A Theoretical Framework for Mitigating Delay in 3D Wireless Data Center Networks

Kai Zhou, Xiaohua Tian
Department of Electronic Engineering
Shanghai Jiao Tong University, Shanghai, China, 200240
Email:{beyondzk,xtian} @sjtu.edu.cn

Yu Cheng
Department of Electrical and Computer Engineering
Illinois Institute of Technology, Chicago, IL, USA ,60616
Email: cheng@iit.edu

Abstract—**Recently, a novel 3D wireless mechanism based on 60 GHz band has been proposed to mitigate the job completion time (JCT) in Data Center Networks (DCNs), where signals bounce off DC ceilings to establish wireless connections. The 3D scheme could alleviate hotspots in DCNs with flexible multi-gigabit wireless links, which bypasses the line-of-sight limitation of 60 GHz wireless links. However, the novel wireless transmission mechanism incurs significant change in the traditional interference model for wireless networks, and the theoretical analysis tool for such hybrid networks is still unavailable. This paper presents a theoretical framework for such 3D DCNs, where the entire network is first transformed from 3D to 2D by remodeling the 3D interference effects. The transformed graph is then processed with a multi-dimensional conflict graph methodology, where the wired and wireless sub-graphs induced by the DCN topology are jointly analyzed. The last but not least, the processed graph is modeled as a minimum job completion time (MJCT) problem, where the optimal traffic engineering, channel allocation and scheduling schemes in the original DCN can be obtained. Simulation results are presented to demonstrate the delay performance of our proposed approach.**

I. Introduction

The increasingly popular cloud services are driving the creation of data center networks (DCNs) that hold massive servers and concurrently support a large number of Internet services. The prevailing DCNs evolving from the LAN networks, however, are suffering from hotspots problem by oversubscription [1] and unbalanced traffic [2], which make the job completion time (JCT) for the supported Internet services unsatisfactory. To tackle with the issue, advanced network topologies are designed to avoid oversubscription, and some design could even achieve non-oversubscription [3] [4]; nevertheless, these designs usually incur significant cabling updating, which could be a major challenge to large-scale DCNs.

Several works propose to augment traditional wired DCNs with wireless links, where 60 GHz technology is utilized due to its reasonably large capacity up to multi-gigabit per second [1] [5]. Fully exploiting the augmented wireless links needs overcoming the inherit drawbacks of 60GHz technology such as link blockage and radio interference [6]. Recently, a novel 3-dimensional (3D) wireless mechanism based on 60GHz band has been proposed to mitigate the JCT in DCNs, where signals bounce off DC ceilings to establish wireless connections [6]. While wireless networking technique provides a new perspective in resolving DCN challenges, a theoretical framework to guide efficient utilization of wireless links in DCNs is still unavailable.

In this paper, we propose a theoretical framework for 3D wireless DCNs. With our framework, the 3D wireless network is first transformed into 2D by remodeling the 3D interference effects. The yielded network could induce a conflict graph by considering interference among wireless links as well as the contention of channels and radio interfaces. As wireless DCNs normally utilize multiple radio and multiple channel configuration, we leverage our previous work multiple dimensional conflict graph (MDCG) [7] [8] [13] to derive maximum independent sets (MISs), each of which consists all wireless links without interference with each other. Based on the MISs, we can perform scheduling and channel assignment to let non-interfered wireless links work simultaneously. For each MIS, we design optimal traffic engineering scheme in order to minimize the job completion time over the MIS.

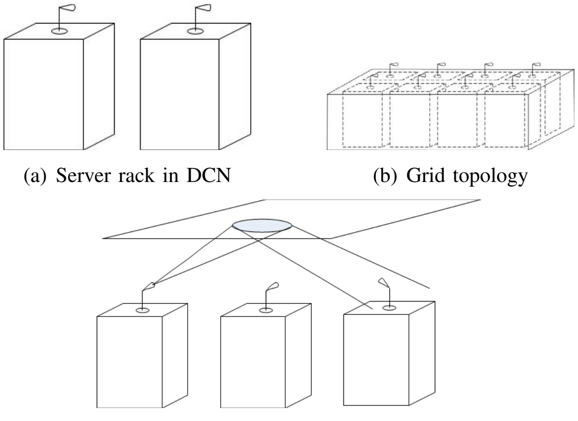
The rest of the paper is organized as follows. Section II describes the system model. Section III presents the 3D interference protocol model, and the MDCG construction. Section VI shows how to achieve the optimal traffic engineering, scheduling and channel assignment. Simulation results are presented in Section V, and conclusion remarks are given in Section VI.

II. System model

The 3D wireless DCNs under study consist of hundreds of racks and each rack contain 20-80 servers [6], as shown in Fig. 1. All the racks are organized in a grid topology, and there is a top-of-rack (TOR) switch connecting severs within each rack and transmitting data among racks. We assume that all this racks are connected by cables to a central controller which are responsible for information exchange and traffic engineering. Wireless devices with 60GHz beamforming antennas are placed on top of each rack, and each rack are equipped with multiple radios due to the multiple source traffic demand from the servers. These beam directions can be adjusted in both azimuth and elevation by placing the horn antennas on rotators with high accuracy. Thus devices can adjust directions

This paper is supported by NSFC 61202373, China Postdoctoral Science Foundation Grant 2011M500774 and 2012T50417, and STCSM Grant 12JC1405200.

978-1-4673-3122-7/13/$31.00 ©2013 IEEE


(a) Server rack in DCN

(b) Grid topology

(c) 3D beamforming

Fig. 1. (a) Racks containing massive servers. (b) All racks are organized in a grid topology. (c) 3D beamforming utilizes the ceiling to bounce off signal to establish connections.

of antennas to meet the demand of wireless transmissions. A metal reflector is placed on the ceiling and a source radio can reflect the signal on the ceiling and adjust the direction to reach the desired destination.

In the DCN environment, due to the special transmission direction, the interference zone has been greatly reduced and constrained in a relatively small area around the receiver. These advantages greatly benefit the performance of DCNs. Compared with its 2D counterpart, this 3D scheme has a larger wireless coverage and higher potential capacity due to its lower interference, which is a promising technique for designing DCNs.

III. Interference Model of 3D Wireless DCNs

A. 3D Interference Protocol Model

In the 3D wireless DCNs, the interference behavior among different wireless links are quite different from that in 2D wireless networks [11] [12]. We here are to describe a systematical approach to transform 3D interference effect [9] to 2D interference model, based on which the multiple dimensional conflict graph (MDCG) approach [7], [8] will be applied for following theoretical analysis.

Given two wireless links: two transmitters X_i and X_j with two receivers $X_{R(i)}$ and $X_{R(j)}$ respectively, we can derive two conditions for link i is not interfered by link j:

1. (1) Transmitter X_j is separated far away enough so that for receiver $X_{R(i)}$ the ratio of the power received from X_i to that of X_j exceeds an limitation.
2. (2) Receiver $X_{R(i)}$ is out of the signal coverage of transmitter X_j.

An illustration of these two conditions is shown in Fig. 2. For condition (1), a free-space propagation model [6] can be used to describe the signal propagation characteristics in 3D wireless data centers:

$$P_r = \frac{P_t G_t G_r \lambda^2}{(4\pi)^2 (L^2 + 4h^2)}, \tag{1}$$

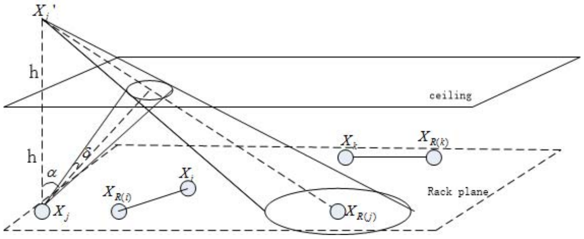


Fig. 2. Illustration for two conditions that link i is not interfered by other links: (1)receiver $X_{R(i)}$ is separated far away enough from transmitter X_k. (2) receiver $X_{R(i)}$ is out of the signal coverage of transmitter X_j.

where P_r and P_t are the receive and transmit power, G_r and G_t are the receive and transmit antenna gain, L is the distance between the transmitter and the receiver, h is the distance from the ceiling to the antenna and λ is the radio wavelength. Let N be the node set and $L = (X_i, X_{R(i)}) : i \in N$ be the active link set in data centers, for the transmission from node X_i to node $X_{R(i)}$ to be successful, the following must be satisfied:

$$(P_r)_{ii} \geq (1 + \Delta)(P_r)_{ij}, \tag{2}$$

where $(P_r)_{ii}$ and $(P_r)_{ij}$ are the power received by $X_{R(i)}$ from X_i and X_j respectively and $\Delta > 0$ is a parameter modeling the tolerance of interference. Combined (1) and (2) we can easily get

$$(L_{ij}^2 + 4h^2) \geq (1 + \Delta)(L_{ii}^2 + 4h^2), \tag{3}$$

where L_{ij} and L_{ii} are the distance from X_j and X_i to $X_{R(i)}$ respectively. And this is a sufficient condition for link i not being interfered by link j.

For condition (2), as shown in figure 2, it's easy to see that the signal coverage of transmitter X_j is an eclipse on the rack plane and as a limitation of space, we simply give the result of the location of this eclipse. Suppose the coordinate of X_j and $X_{R(j)}$ are (x_1, y_1) and (x_2, y_2) on the rack plane, respectively. The angle of the beam spread is 2δ. And the angle between the beam center line and the perpendicular line is

$$\alpha = arctan \frac{\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}}{2h}. \tag{4}$$

From this, we can obtain that the major axis of the eclipse lies on the extension cord of the line from X_j to $X_{R(j)}$ with a length of

$$2h[tan(\alpha + \delta) - tan(\alpha - \delta)], \tag{5}$$

and the center at $(x_1 + dcos\theta, y_1 + dsin\theta)$,where

$$d = h[tan(\alpha + \delta) + tan(\alpha - \delta)] \tag{6}$$

$$\theta = arctan \frac{y_2 - y_1}{x_2 - x_1}. \tag{7}$$

The length of the minor axis is $2\sqrt{(2h)^2 + d^2}tan\delta$. With (4)-(7), we can determine the interference zone of a wireless transmission, thus transforming the 3D interference to 2D. Thus, another sufficient condition for link i is not interfered by link j is that receiver is out of the coverage of the eclipse. In summary, based on these two conditions, we can determine whether two links are interfered with each other.

B. MDCG under 3D Interference Protocol Model

We utilize the MDCG to model the overall interference relationship. In a MDCG, the node is a tuple $((u,v),x_u,x_v,c)$, where (u,v) is the link in the transmission graph, x_u and x_v are the radios utilized by the transmission nodes respectively and c is the channel occupied by this link. An edge between two tuples indicates that they are interfered with each other thus in conflict.

With the observations mentioned above, we define following three events:

E_1 According to the protocol interference model, the nodes in two different tuples are with in each other's interference range.

E_2 The links in two different tuples are occupying the same channel.

E_3 Two different tuples share common radio interfaces at one or two nodes.

Based on these events,we can define that two tuples in the MDCG are in conflict as follows:

***Proposition** 1:* Two tuples in the MDCG are in conflict if the condition $E_1E_2E_3 \cup E_3$ is true.

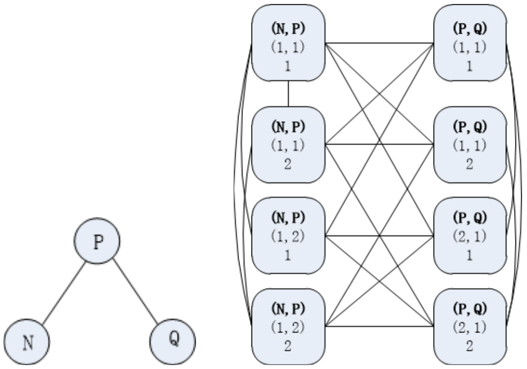
Specifically, the condition $E_1E_2E_3$ means that co-channel transmissions within the interference range of each other will cause interference. Condition E_3 means that nodes involved in different links (even if occupying different channels) utilizing the same radio interface will cause interference.

To illustrate the definition of MDCG, we use an example shown in Fig. 3. The left side of figure is a transmission graph composed of three nodes and two corresponding links (N,P) and (P,Q). We assume that node P has two radio interfaces and node N,Q has one.And there are a total of two orthogonal channels. For instance, tuple $((P,Q),(1,2),1)$ indicates that link (P,Q) occupies channel 1 with node P using its radio interface 1 and node Q using its radio interface 2. Thus, there are total $2 \times 1 \times 2$ tuples and according to the conflict conditions, we can draw an edge between two tuples if they are in conflict. In this way, we can map the transmission graph into the MDCG shown in the right side of Fig. 3.

IV. Minimizing Job Completion Time

A. Job Completion Time

In this section, we model the data transmissions in WDCNs as a minimum job completion time (MJCT) problem. On one hand, the interference nature makes it hard to determine the scheduled capacity of the wireless links, which make it even harder to determine the flow amount on wireless links. On the other hand, the total throughput is not a suitable merit for data centers any more. It is because in data centers the aim is to alleviate the hot spots problem, that is to assign heavy traffic to the lighter links. Although the maximum total throughput can be achieved, it cannot guarantee that all the data flows can finishing transmission at the same time. Thus, we propose job completion time as a new merit for data transmission in data centers.



(a) Transmission graph

(b) Conflict graph

Fig. 3. Map transmission graph to conflict graph

We define the transmission graph as follows:

***Definition** 1:* A transmission graph is a undirected graph $G = (V,E)$,where V denotes the set of transmission unit or node and E is the set of edges denoting the transmission links between those nodes.

In data centers, a typical application or job can be composed of many data transmissions or flows from source nodes to destination nodes. We denote a flow from source node s to destination node t as f_{st} with $s \in N_s, t \in N_t$ where N_s and N_t are the sets of source and destination nodes, respectively. The flow demand $F = \{f_{s,t}|s,t \in V\}$ is gathered from the TOR switches on each rack and this information is transmitted to the control center through wired links. It is easy to see that the completion time of a job depends on the completion of the last flow. This is the main difference compared with MCF problems in multiple-radio multiple-hop wireless networks, where the central concern is to maximize the throughput or capacity. While in DCNs, we aim at finishing the jobs as quickly as possible. Thus, in order to achieve this goal (i.e.,to minimize the job completion time), we should distribute all flow amount evenly among all the links to make all the flows complete roughly at the same time.

In a transmission graph $G = (V,E)$, each link $l_{uv} = (u,v)$ has a capacity c_{uv}. Given a flow f_{st}, $f_{st}(u,v)$ of it is assigned to link l_{uv} and the time it takes to complete the transmission on this link is defined as

$$t_{uv} = \sum_{s \in N_s, t \in N_t} \frac{f_{st}(u,v)}{c_{uv}}. \tag{8}$$

In order to minimize the job completion time, we should assign $f_{st}(u,v)$ so as to make t_{uv} distributed evenly. Based on this, we define the minimum job completion time problem as follows:

***Definition** 2:* Given a flow vector $F = \{f_{s,t}|s,t \in V\}$ and a transmission graph $G = (V,E)$, a minimum job completion time problem in WDCNs is to assign flow amount to each link so as to minimize job completion time.

Flow gathering
Algorithm decision
Flow transmission

Fig. 4. Joint TCS scheme design

B. Joint Design for MJCT

In this section, we aim at designing a joint traffic engineering, channel allocation and scheduling (TCS) scheme to minimize the job completion time. As shown in Fig. 4, the timeline of the system is divided into time slots and each slot is divided into three stages: at first stage, each rack detects flow demands from the severs within the rack and this demand is transmitted to a central controller in DCNs along extra wired links. Then, in the central controller, the flow transmission decision including traffic engineering results, channel allocation and scheduling are calculated and this decision is transmitted to each rack. At the last stage, each flow is transmitted according to the decision.

In order to calculate the flow distribution over each link, the network topology and the capacity of each link must be known. However, due to interference, the wireless links in data centers cannot be active simultaneously and work at full capacity. Thus, when all the wireless links are considered, it is hard to determine the capacity of each link. To solve this problem, we augment the MJCT problem with conflict graph constraints to take interference among different wireless links into account. Specifically, given a conflict graph, we define an independent set I as a set of tuples where no edge exists between two tuples indicating that the links involved in the tuples can be active simultaneously. A maximum independent set (MIS) is an independent set where adding one more tuple to this independent set will make it nonindependent. Suppose that there are s MISs in a conflict graph denoted as $I_1, I_2, ..., I_s$ and for MIS I_k, a fraction of α_k of transmission time is allocated to it in the transmission stage. Suppose that a wireless link l_{uv} has a capacity of w_{uv}^c where c is the channel this link occupied. After scheduling all MISs, the capacity of link l_{uv} becomes

$$C_{uv} = \sum_{k:l_{uv} \in I_k} \alpha_k w_{uv}^{c(l_{uv}, I_k)}, \forall l_{uv} \in L, \tag{9}$$

where $c(l_{uv}, I_k)$ indicates that link l_{uv} occupies channel c in MIS I_k, and α_k should satisfy the following constrains:

$$\sum_{k=1}^{s} \alpha_k \leq 1. \tag{10}$$

The capacity of each wireless link can be determined by (9) based on which we can solve the MJCT problem using the nonlinear programming approach. To minimize the total flow completion time, we should let each sub-flow completes over each link simultaneously, that is to make t_{uv} on each link close to each other. Thus, we construct the objective function as:

$$min \frac{1}{|L|} \sum_{l_{uv} \in L} (t_{uv} - T)^2 \tag{11}$$

Subject to:

$$T = \frac{1}{|L|} \sum_{l_{uv} \in L} t_{uv} \tag{12}$$

$$\sum_{l_{uv} \in L} f_{st}(u, v) = \sum_{l_v u} f_{st}(v, u), \forall f_{st} \in F \tag{13}$$

$$\text{and } v \in N / \{N_s, N_t\}$$

$$\sum_{l_u s \in L} f_{st}(u, s) = 0, \forall f_{st} \in F \tag{14}$$

$$\sum_{l_{tv} \in L} f_{st}(t, v) = 0, \forall f_{st} \in F \tag{15}$$

$$f_{st}(u, v) \geq 0, \forall l_{uv} \in L \tag{16}$$

$$\text{and } f_{st} \in F$$

The meanings of the constrains are stated as follows:

- (14): except the source and the destination node, the amount of incoming flow equals the amount of outgoing flow at each node;
- (15): for each data flow, the incoming flow to the source node is 0;
- (16): for each data flow, the outgoing flow from the destination node is 0;
- (17): the amount of flow distribution on each link should be non-negative.

Along with the constrains (9)(10), a nonlinear programming approach is formulated.

The solution of this minimum job completion time problem consist of flow distribution $f_{st}(u, v)$, scheduling time fraction α_k and the corresponding MIS I_k. For instance, assuming a large total scheduling time T and a set of s MISs. The scheduling result is all the MISs take turns to access channels for time length $\alpha_k T$. The advantages of this scheme is that 1) flow distribution $f_{st}(u, v)$ actually reflects the optimum engineering to achieve minimum flow completion time. 2) time fraction α_k indicates the scheduling result of those wireless links and 3) the tuples in each MIS indicates the channel allocation to each link. Thus, we have designed a joint traffic engineering, channel allocation and scheduling scheme for 3D wireless data centers to achieve the minimum job completion time.

V. Numeric Results

In this section, we use simulations by MATLAB to give a performance evaluation of the proposed approach. We investigate the influence of the number of flow demands on flow completion time. The architecture of 3D DCN is shown in Fig. 1, where the topology of the wired network is tree based, as shown in Fig. 5. The topology is divided to three layers. The first layer is a router and the second layer consist of many aggregation switches which connects racks that constituting the third layer. Wireless links can be established between racks

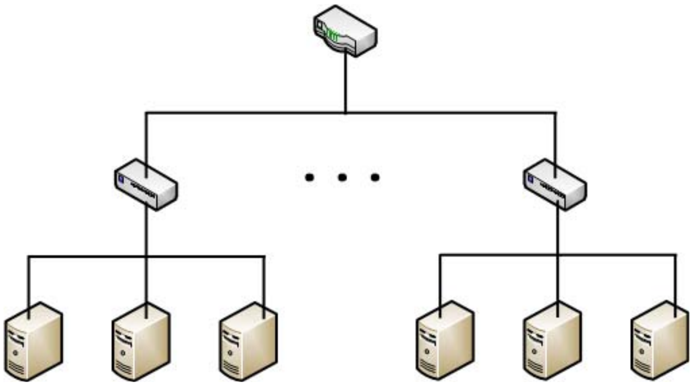


Fig. 5. DCN topology

in the third layer dynamically. For simulation simplicity, we assume that each aggregation switch connects to 10 racks, and the capacity of wired links in each layer is fixed.

By injecting to the networks with different number of flows, we compare the performance of our proposed scheme with that of wired data centers. We assume that the wired links in the first layer shown in fig. 5 have a fixed capacity of 20 $units/s$ and links in the second layer have 10 $units/s$. The wireless links have fixed while different capacities before scheduling. And suppose that each flow has 100 $units$ to transmit and the flow number vary from 1 to 10. And there is a total of 50 racks and each rack is equipped with four radio interfaces. The number of channels is three. Fig. 6 shows the job completion time and flow number under our proposed scheme and in traditional wired data centers without wireless links. We can see that as more flows are injected into the network, the flow completion time have increased a lot since more congestion occurs. While in our scheme, the data flow can be transferred to the right wireless links at the right time. As a consequence, the flow completion time increases in a much slower speed. One thing we should notice is that, in our simulation, we only consider 50 racks for simplicity while in real data centers the number of racks could be thousands where our proposed scheme should have a much bigger improvement.

VI. Conclusion

In this paper, we have proposed a theoretical framework for 3D wireless DCNs. With our framework, the 3D wireless network is first transformed into 2D by remodeling the 3D interference effects. The yielded graph could induce a conflict graph by considering interference among wireless links. As wireless DCNs normally utilize multiple radio and multiple channel configuration, we have leveraged our previous work multiple dimensional conflict graph (MDCG) to derive maximum independent sets (MISs), each of which consists all wireless links without interference with each other. Based on the MISs, we have performed scheduling and channel assignment to let non-interfered wireless links work simultaneously. For each MIS, we design optimal traffic engineering scheme in order to minimize the job completion time over the MIS. Simulation results are presented to demonstrate the delay performance of our proposed approach.

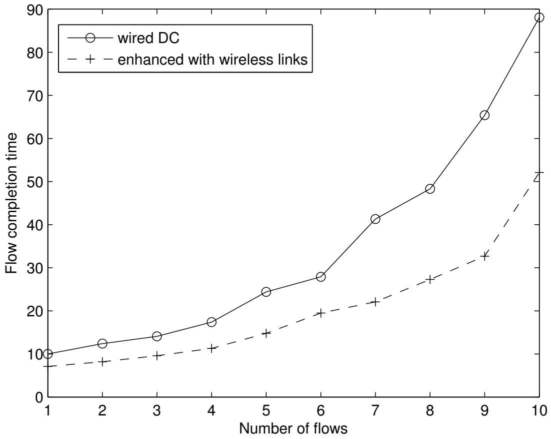


Fig. 6. Flow completion time with different number of flows

References

[1] S. Kandula, J. Padhye, and V. Bahl, “Flyways to de-congest data center networks,” in *8th ACM Workshop on Hot Topics in Networks*. New York City, NY, Otc. 2009, pp. 45–50.

[2] B. Theophilus, A. Ashok, A. Aditya, and Z. Ming, “Understanding data center traffic characteristics,” *SIGCOMM Comput. Commun. Rev.*, vol. 40, no. 1, pp. 92–99, Jan. 2010.

[3] A. Greenberg, N. Jain, S. Kandula, C. Kim, P. Lahiri, D. Maltz, P. Patel, and S. Sengupta, “Vl2: A scalable and flexible data center network,” in *Proceedings of the ACM SIGCOMM 2009 conference*. Barcelona, Spain, Aug. 2009, pp. 51–62.

[4] A. Mohammad, L. Alexander, and V. Amin, “A scalable, commodity data center network architecture,” in *Proceedings of the ACM SIGCOMM 2008 conference on Data communication*. Seattle, WA, Aug. 2008, pp. 63–74.

[5] H. Daniel, K. Srikanth, A. Jitendra, B. Paramvir, and W. David, “Augmenting data center networks with multi-gigabit wireless links,” in *Proceedings of the ACM SIGCOMM 2011 conference*. Toronto, Canada, Aug. 2011, pp. 38–49.

[6] Z. Xia, Z. Zengbin, Z. Yibo, L. Yubo, K. Saipriya, V. Amin, Z. B. Y., and Z. Haitao, “Mirror mirror on the ceiling: Flexible wireless links for data centers,” in *Proceedings of the ACM SIGCOMM 2012 conference on Applications, technologies, architectures, and protocols for computer communication*. Helsinki, Finland, Aug. 2012, pp. 443–454.

[7] Y. Cheng, H. Li, and P. Wan, “A theoretical framework for optimal cooperative networking in multiradio multichannel wireless networks,” *Wireless Communications, IEEE*, vol. 19, no. 2, pp. 66–73, Apr. 2012.

[8] H. Li, Y. Cheng, C. Zhou, and P. Wan, “Multi-dimensional conflict graph based computing for optimal capacity in mr-mc wireless networks,” in *Proceedings of the 2010 IEEE 30th International Conference on Distributed Computing Systems*. Genova, Italy, Jun. 2010, pp. 774–783.

[9] G. P. and K. P.R, “The capacity of wireless networks,” *Information Theory, IEEE Transactions on*, vol. 46, no. 2, pp. 388–404, Mar. 2000.

[10] P. Gupta and P. R. Kumar, “Internets in the sky: The capacity of three dimensional wireless networks,” *Communications in Information and Systems*, vol. 1, no. 2, pp. 33–50, Mar. 2001.

[11] X. Wang, W. Huang, S. Wang, J. Zhang and C. Hu, “Delay and Capacity Tradeoff Analysis for MotionCast,” in *IEEE/ACM Transactions on Networking*, Vol. 19, no. 5, pp. 1354–1367, Oct. 2011.

[12] W. Huang and X. Wang, “Capacity Scaling of General Cognitive Networks,” in *IEEE/ACM Transactions on Networking*, vol 20, no. 5, pp. 1501-1513, 2012.

[13] Y. Cheng, X. Ling, and W. Zhuang, “A protocol-independent approach for analyzing the optimal operation point of CSMA/CA protocols”, in *IEEE INFOCOM 2009*, Rio de Janeiro, Brazil, Apr. 19-25, 2009.