On the Feasibility of Completely Wirelesss Datacenters

Ji-Yong Shin, Emin Gün Sirer, Hakim Weatherspoon, and Darko Kirovski

***Abstract*—Conventional datacenters, based on wired networks, entail high wiring costs, suffer from performance bottlenecks, and have low resilience to network failures. In this paper, we investigate a radically new methodology for building wire-free datacenters based on emerging 60-GHz radio frequency (RF) technology. We propose a novel rack design and a resulting network topology inspired by Cayley graphs that provide a dense interconnect. Our exploration of the resulting design space shows that wireless datacenters built with this methodology can potentially attain higher aggregate bandwidth, lower latency, and substantially higher fault tolerance than a conventional wired datacenter while improving ease of construction and maintenance.**

***Index Terms*—Communication technology, computer networks, millimeter-wave communication, wireless networks.**

I. Introduction

PERFORMANCE, reliability, cost of the switching fabric, power consumption, and maintenance are some of the issues that plague conventional wired datacenters [1]–[3]. Current trends in cloud computing and high-performance datacenter applications indicate that these issues are likely to be exacerbated in the future [4], [5].

In this paper, we explore a radical change to the construction of datacenters that involves the removal of all but power-supply wires. The workhorses of communication in this new design are the newly emerging directional, beamformed 60-GHz radio frequency (RF) communication channels characterized by high bandwidth (4–15 Gb/s) and short range ($\leq$ 10 m). New 60-GHz transceivers [6], [7] based on standard 90-nm CMOS technology make it possible to realize such channels with low cost and high power efficiency ($<$ 1 W). Directional ($25°$–$60°$ wide) short-range beams employed by these radios enable a large number of transmitters to simultaneously communicate with multiple receivers in tight confined spaces.

Manuscript received April 01, 2013; accepted June 21, 2013; approved by IEEE/ACM TRANSACTIONS ON NETWORKING Editor L. E. Li. Date of publication August 15, 2013; date of current version October 11, 2013. This work was supported in part by the National Science Foundation under Grants No. 0424422, No. 1040689, No. 1053757, No. 1111698, and No. SA4897–10808PG, the DARPA under Grants No. D11AP00266 and No. FA8750–11-2–0256, and an Intel Early Career Faculty Honor. A conference version of this paper appeared in the Proceedings of the 8th ACM/IEEE Symposium on Architectures for Networking and Communications Systems (ANCS), Austin, TX, USA, October 29-30, 2012.

J.-Y. Shin, E. G. Sirer, and H. Weatherspoon are with the Department of Computer Science, Cornell University, Ithaca, NY 14853 USA (e-mail: jyshin@cs.cornell.edu; egs@cs.cornell.edu; hweather@cs.cornell.edu).

D. Kirovski is with Jump Trading, Chicago, IL 60654 USA (e-mail: darko@kirovski.org).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNET.2013.2274480

The unique characteristics of 60-GHz RF modems pose new challenges and tradeoffs. The most critical questions are those of feasibility and structure: Can a large number of transceivers operate without signal interference in a densely populated datacenter? How should the transceivers be placed and how should the racks be oriented to build practical, robust, and maintainable networks? How should the network be architected to achieve high aggregate bandwidth, low cost, and high fault tolerance? Can such networks compete with conventional wired networks?

To answer these questions, we propose a novel datacenter design. Because its network connectivity subgraphs belong to a class of Cayley graphs [8], we call our design a Cayley datacenter. The key insight behind our approach is to arrange servers into a densely connected, low-stretch, failure-resilient topology. Specifically, we arrange servers in cylindrical racks such that inter- and intra-rack communication channels can be established and form a densely connected mesh. To achieve this, we replace the network interface card (NIC) of a modern server with a Y-switch that connects a server's system bus with two transceivers positioned at opposite ends of the server box. This topology leads to full disappearance of the classic network switching fabric (e.g., no top-of-rack switches, aggregation switches, access routers, copper and optical interconnects) and has far-reaching ramifications on performance.

Overall, this paper makes three contributions. First, we present the first constructive proposal for a fully wireless datacenter. We show that it is possible for 60-GHz technology to serve as the sole and central means of communication in the demanding datacenter setting. Second, we propose a novel system-level architecture that incorporates a practical and efficient rack-level hardware topology and a corresponding geographic routing protocol. Finally, we examine the performance and system characteristics of Cayley datacenters. Using a set of 60-GHz transceivers, we demonstrate that signals in Cayley datacenters do not interfere with each other. We also show that, compared to a fat-tree [9], [10] and a conventional datacenter, our proposal exhibits higher bandwidth, substantially improved latency due to the switching fabric being integrated into server nodes, lower power consumption, and easier maintenance as a result of the plug-and-play simplicity of connecting servers. Cayley datacenters exhibit strong fault tolerance due to a routing scheme that can fully explore the mesh: Cayley datacenters can maintain connectivity to over 99% of live nodes until up to 31% of total racks or 55% of total nodes fail.

II. 60-GHz Wireless Technology

In this section, we briefly introduce the communication characteristics of the newly emerging 60-GHz wireless technology, which is the foundation of our datacenter.

TABLE I
60-GHz WIRELESS TRANSCEIVER CHARACTERISTICS [6]

Category	Characteristic
Technology	Standard 90nm CMOS
Packaging	Single chip Tx/Rx in QFN
Compliance	ECMA TC48
Power	0.2W (at output power of 3dBm)
Range	$\leq$ 10m
Bandwidth	4-15Gbps

Propagation of RF signals in the 57–64-GHz subband is severely attenuated because of the resonance of oxygen molecules, which limits the use of this subband to relatively short distances [11]. Consequently, 57–64 GHz is unlicensed under FCC rules and open to short-range point-to-point applications. To date, 60 GHz as a technology has been mostly pursued as a wireless replacement for high-definition multimedia interface (HDMI) connections [12]. Several efforts are aiming to standardize the technology, with most of them tailored to home entertainment: two IEEE initiatives, IEEE 802.15.3c and 802.11.ad [13], [14], WiGig 7 Gb/s standard with beamforming [15], and ECMA-387/ISO DS13156 6.4 Gb/s spec [16] based upon the Georgia Institute of Technology's (Georgia Tech's) design [6].

In this paper, we focus on a recent integrated implementation from Georgia Tech whose characteristics are summarized in Table I.

More details about 60-GHz transceiver characteristics can be explained using a link margin, which models communication between a transmitter (Tx) and a receiver (Rx). The link margin M is the difference between the received power at which the receiver stops working and the actual received power and can be expressed as follows:

$$M = P_{\mathrm{TX}} + G_{\mathrm{TX+RX}} - L_{\mathrm{Fade}} - L_{\mathrm{Implementation}} - FSL - NF - SNR \quad (1)$$

where P_{TX} and $G_{\mathrm{TX+RX}}$ represent transmitted power and overall joint transmitter and receiver gain that is dependent upon the geometric alignment of the $\mathrm{Tx} \leftrightarrow \mathrm{Rx}$ antennas [17]. Free-space loss equals $FSL = 20\log_{10}(4\pi D/\lambda)$, where D is the line-of-sight $\mathrm{Tx} \leftrightarrow \mathrm{Rx}$ distance and λ wavelength. The noise floor $NF \sim 10\log_{10}(R)$ is dependent upon R, the occupied bandwidth. SNR is the signal-to-noise ratio in decibels, which links a dependency to the bit error rate as $BER = \frac{1}{2}\mathrm{erfc}(\sqrt{SNR})$ for binary phase-shift keying (BPSK) modulation for example. Loss to fading and implementation are constants given a specific system. From (1), one can compute the effects of constraining different communication parameters.

Fig. 1 illustrates a planar slice of the geometric communication model we consider in this paper. A transmitter antenna radiates RF signals within a lobe—the surface of the lobe is a level-set whose signal power is equal to one half of the maximum signal power within the lobe. Because the attenuation is very sharp in the 60-GHz frequency range, a receiver antenna should be within the bound of a transmitter's beam for communication. The beam is modeled as a cone with an angle θ and length L. Using a spherical coordinate system centered at transmitter's antenna, one can define the position of the receiver antenna with its radius, δ, elevation α, and azimuth β. The plane of the receiver antenna can then be misaligned from the plane

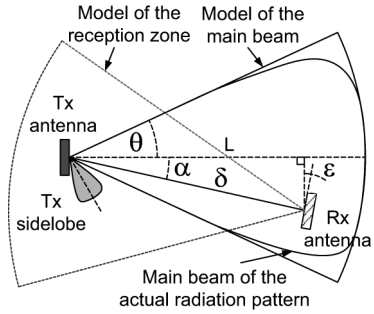

Fig. 1. Geometric communication model.

of the transmitter antenna by an angle ε along the elevation and γ along the azimuth. We use a modeling tool developed at Georgia Tech to convert $\{\alpha, \beta, \gamma, \varepsilon, \delta, L, \theta\}$ into $G_{\mathrm{TX+RX}}$. Through personal communication with Georgia Tech's design team, we reduced our space of interest to $25^\circ \leq \theta \leq 45^\circ$ as a constraint to suppress sidelobes. Based on design parameters from the antenna prototypes developed by the same team, we model a reception zone of the receiver that is in identical shape to the main transmitter beam. We limit ε and γ to be smaller than θ such that the transmitter is located within the reception zone and assume a BER of 10^{-9} at 10 Gb/s bandwidth within $L < 3$ m range. We do not utilize beam-steering[1] and assume that the bandwidth can be multiplexed using both time- (TDD) and frequency-division duplexing (FDD).

The design parameters of the transceiver are optimized for our datacenter design and lead to a higher bandwidth and less noisy transceiver design compared to off-the-shelf 60-GHz transceivers for HDMI [12]. We validate the assumptions behind these parameters in Section IV with a conservative 60-GHz hardware prototype built by Terabeam/HXI [18]. More research in 60-GHz RF design with a focus on Cayley datacenters can further improve performance.

III. CAYLEY DATACENTER DESIGN

This section introduces Cayley datacenter architecture, the positioning of the 60-GHz transceivers in a wireless datacenter, and the resulting network topology. We introduce a geographical routing protocol for this unique topology and discuss how to overcome failures. We also adopt a MAC—layer protocol to address the hidden terminal and the masked node problems.

A. Component Design

In order to maximize opportunities for resource multiplexing in a wireless datacenter, it is important to use open spaces efficiently because the maximum number of live connections in the network is proportional to the volume of the datacenter divided by that of a single antenna beam. We focus on the network topology that would optimize key performance characteristics, namely latency and bandwidth.

To separate the wireless signals for communications within a rack and among different racks, we propose cylindrical racks [Fig. 2(a)] that store servers in prism-shaped containers [Fig. 2(c)]. This choice is appealing because it partitions the datacenter volume into two regions: intra- and inter-rack free

[1]Typically, reconnection after beam-steering involves training of communication codebooks involving delays on the order of microseconds. This may be tolerated in home networking scenarios, but not in the datacenter.

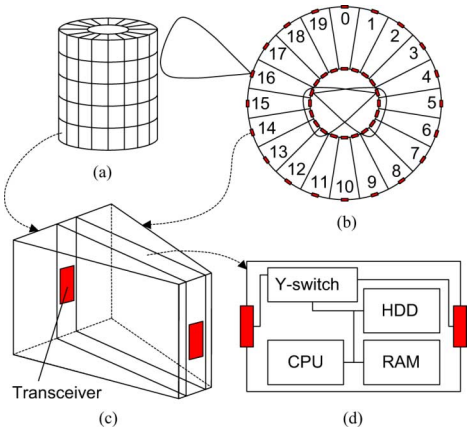

Fig. 2. Rack and server design. (a) Rack (3-D view). (b) Rack (2-D view from the top). (c) Container. (d) Server.

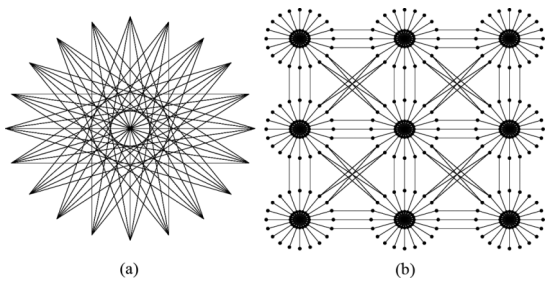

Fig. 3. Cayley datacenter topology when $\theta = 25°$. (a) Intra-rack. (b) Inter-rack.

space. A single server can be positioned so that one of its transceivers connects to its rack's inner-space and another to the inter-rack space as the rack illustrated in Fig. 2(b). A rack consists of S stories, and each story holds C containers; we constrain $S = 5$ and $C = 20$ for brevity of analysis and label servers in the same story from 0 to 19 starting from the 12 o'clock position in a clockwise order.

The prism containers can hold commodity half-height blade servers. A custom-built Y-switch connects the transceivers located on opposite sides of the server [Fig. 2(d)]. The Y-switch, whose design is discussed at the end of this section, multiplexes incoming packets to one of the outputs.

B. Topology

The cylindrical racks we propose utilize space and spectrum efficiently and generalize to a topology that can be modeled as a mesh of Cayley graphs. A Cayley graph [8] is a graph generated from a group of elements G and a generator set $S \subseteq G$. Set S excludes the identity element $e = \mathrm{g} \cdot \mathrm{g}^{-1}$, where $\mathrm{g} \in G$, and $h \in S$ *iff* $h^{-1} \in S$. Each vertex $v \in V$ of a Cayley graph (V, E) corresponds to each element $\mathrm{g} \in G$ and edge $(v_1, v_2) \in E$ *iff* $\mathrm{g}_1 \cdot \mathrm{g}_2^{-1} \in S$.

This graph is vertex-transitive, which facilitates the design of a simple distributed routing protocol and is generally densely connected, which adds fault tolerance to the network [19].

When viewed from the top, connections within a story of the rack form a 20-node, degree-k Cayley graph, where k depends on the signal's radiation angle [Fig. 3(a)]. This densely connected graph provides numerous redundant paths from one server to multiple servers in the same rack and ensures strong connectivity.

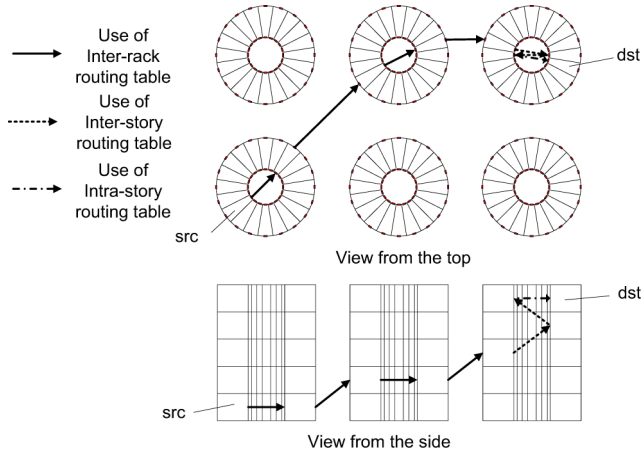

Fig. 4. Diagonal XYZ routing example.

The transceivers on the exterior of the rack stitch together Cayley subgraphs in different racks. There is great flexibility in how a datacenter can be constructed out of these racks, but we pick the simplest topology by placing the racks in rows and columns for ease of maintenance. Fig. 3(b) illustrates an example of the two-dimensional connectivity of 9 racks in 3×3 grids: Small black dots represent the transceivers, and the lines indicate the connectivity. A Cayley graph sits in the center of each rack, and the transceivers on the exterior of the racks connect the subgraphs together. Relatively long lines connecting the transceivers on the exterior of the racks show the wireless inter-rack connections. Furthermore, since the wireless signal spreads in a cone shape, a transceiver is able to reach servers in different stories, both within and across racks.

C. Routing Protocol

A routing protocol for datacenters should enable quick routing decisions, utilize a small amount of memory, and find efficient routes involving few network hops. A geographic routing technique for our topology can fulfill these conditions.

1) Diagonal XYZ Routing: The uniform structure of Cayley datacenters lends itself to a geographical routing protocol. The routing protocol that we investigate in this paper is called diagonal XYZ routing.

Similar to XY routing [20], diagonal XYZ routing finds an efficient route to the destination at a low computational and storage cost using geographical information. We define the geographical identity g_k of a server k as (rx, ry, s, i), where rx and ry are the x- and y-coordinates of the rack, s corresponds to the ordinal number for the story, and i is the index of the server within a story. Cayley datacenters use this identity to address the servers. Once a datacenter administrator manually configures the identity of several servers, the rest of the servers can identify their identities by querying the neighbors and propagating the information.

The geographical identity facilitates finding a path in the Cayley datacenter network. The routing protocol determines the next hop by comparing the destination of a packet to the identity of the sever holding the packet. Based on rx and ry values, the protocol finds an adjacent rack of the server that is closest to the destination. The s value is then used to reach the story height of the destination that the packet should arrive. Finally, the i value is used to forward the packet using the shortest path to the destination server within the same story.

Algorithm 1: Diagonal XYZ Routing

Require: g_{curr}: geographical identity of the server, where the packet is currently at
g_{dst}: geographical identity of the packet's final destination
r_{curr}: rack of g_{curr}
r_{dst}: rack of $g_{\mathrm{dst}} R_{\mathrm{adj}}$
R_{adj}: set of adjacent racks of r_{curr}
$T_{\mathrm{InterRack}}$: inter-rack routing table of curr
$T_{\mathrm{InterStory}}$: inter-story routing table of curr
$T_{\mathrm{IntraStory}}$: intra-story routing table of curr
Ensure: g_{next}: geographical identity of next destination
if IsInDifferentRack $(g_{\mathrm{curr}}, g_{\mathrm{dst}})$ **then**
$r_{\mathrm{next}} \leftarrow r_{\mathrm{dst}}$.GetMinDistanceRack (R_{adj})
$dir \leftarrow r_{\mathrm{curr}}$.GetHorizontalDirection (r_{next})
$G \leftarrow T_{\mathrm{InterRack}}$.LookupGeoIDs $(dir, g_{\mathrm{dst}}.s)$
else if IsInDifferentStory $(g_{\mathrm{curr}}, g_{\mathrm{dst}})$ **then**
$dir \leftarrow g_{\mathrm{curr}}$.GetHorizontalDirection (g_{dst})
$G \leftarrow T_{\mathrm{InterStory}}$.LookupGeoIDs $(dir, g_{\mathrm{dst}}.s)$
else if IsDifferentServer $(g_{\mathrm{curr}}, g_{\mathrm{dst}})$ **then**
$G \leftarrow T_{\mathrm{IntraStory}}$.LookupGeoIDs $(g_{\mathrm{dst}}.i)$
else
$G \leftarrow g_{\mathrm{dst}}$
end if
$g_{\mathrm{next}} \leftarrow$ RandomSelect (G)

Algorithm 1 describes the details about the routing algorithm, and Fig. 4 illustrates an example of using this algorithm.

Because the topology has a constant fanout, diagonal XYZ routing requires very little state to be maintained on each host. Every host keeps and consults only three tables to determine the next destination for a packet.

- *Inter-rack routing table*: Maps eight horizontal directions toward adjacent racks to directly reachable servers on the shortest path to the racks.
- *Inter-story routing table*: Maps two vertical directions to directly reachable servers in the same rack of the table owner leading to the desired story.
- *Intra-story routing table*: Maps 20 server index i's to directly reachable servers in the same story in the same rack of the table owner. The servers in the table are on the precomputed shortest path leading to server i.

Inter-rack and inter-story routing tables maintain story s as the secondary index for lookup. Using this index, LookupGeoIDs $(dir, g_{\mathrm{dst}}.s)$ returns the identities with the closest s value to $g_{\mathrm{dst}}.s$ among the ones leading to dir.

For all three tables, LookupGeoIDs returns multiple values because a transceiver can communicate with multiple others. The servers found from the table lookup all lead to the same number of hops to the final destination. Thus, the routing protocol pseudo-randomly selects one of the choices to evenly distribute the traffic and to allow a TCP flow to follow the same path. We use a pseudo-random hashing of the packet header like the Toepliz Hash function [21].

The directionality of the radio beam, the presence of multiple transceivers per node, and the low latency of the Y-switch make it possible for Cayley datacenters to deploy cut-through switching [22], which starts routing a packet immediately after receiving and reading the packet header. While this is generally not usable in wireless communication based on omnidirectional antennas, unless special methodologies such as signal cancellation are employed [23], [24], Cayley datacenter servers employ this optimization.

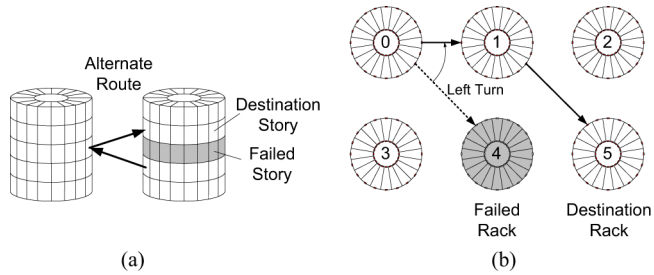

Fig. 5. Routing under failures. (a) Routing when stories fail. (b) Routing when racks fail.

2) Adaptive Routing in Case of Failure: Compared to a conventional datacenter, a Cayley datacenter has a distinct failure profile. While a group of servers within a rack shares much of the switching gear (switches and wires) in the former, thus causing dependencies on the failure probability model, in the Cayley datacenter, these probabilities are both lower (fewer components can fail) and less dependent (servers within a rack share only the same power supply and similar operating temperature) [2]. When they happen, failures of nodes in a Cayley datacenter can lead to a failure on a network path. In this section, we show that our proposal is capable of dealing with substantial server failures. Employing a simple timeout scheme in the MAC-layer protocol enables easy detection of such failures.

Isolated node failures typically do not lead to disconnection in Cayley datacenters because the servers around the failed server are likely to provide comparable routing functionality as the failed node. Similarly, racks that may be missing individual servers appear like a failed server, so other servers around a "hole" can be used. Therefore, we focus our discussion on two cases with massive, correlated failures: failure of an entire rack or a story of servers within a rack.

Failure of all servers in a single story can affect different paths (i.e., a set of next-hop nodes G in Algorithm 1) among stories within the same rack.

In this case, we can use adjacent racks to deliver the packet to different stories as shown in Fig. 5(a).

Rack failures are potentially more catastrophic as they can affect larger number of packets in the inter-rack routing level. We adopt a geographic routing technique based on face routing [25] to deal with these cases. Once the MAC layer detects that a failed rack is blocking the path for a packet, our routing protocol sets up a temporary destination for the packet using the left turn (a.k.a. right hand) rule. If the next rack to visit has failed [e.g., the southeast rack in Fig. 5(b)], our routing protocol will temporarily route the packet to the adjacent rack to the left [e.g., the east in the example of Fig. 5(b)]. Once the packet arrives at its temporary destination, the protocol routes the packet to its original destination.

However, using only the left turn rule can lead to infinite loops. Assume a packet's source is rack 2 and the destination is rack 3 in Fig. 5(b). Routing using only left turns will endlessly route the packet between racks 2 and 5. To prevent this, the failure routing switches to the right turn (a.k.a. left hand)

rule, when the following conditions are satisfied: The packet is at the rack on the edge of the grid of racks [e.g., rack 5 in Fig. 5(b)], and there is no rack on the very left of the failed rack [e.g., from the viewpoint of rack 5, there is no rack on the very left of rack 4 in Fig. 5(b)]. In this example, the full path of racks that the packet goes through becomes racks 2 (left turn applies), 5 (right turn applies), 1, and 3, in that order.

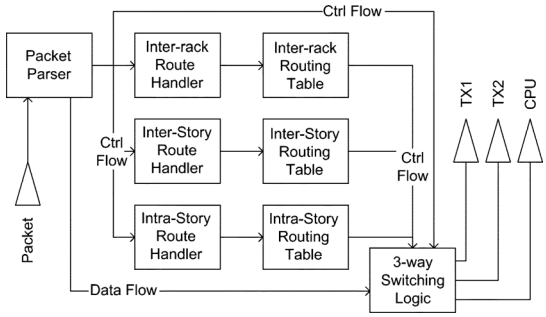

Fig. 6. Y-switch schematic.

Even if we use both left and right turn rules, certain rare failures, such as racks failing in a "⊔" pattern, can lead to live locks. These cases require several racks, which consist of several hundreds of servers, to fail at the same time. However, field data indicates that failures usually involve less than 20 components in datacenters with 100 K nodes, so such failures are very unlikely [2]. Modification of our failure routing can overcome such failures, but given the rarity of the scenario, we do not discuss them here.

D. MAC-Layer Arbitration

A transceiver in a Cayley datacenter can communicate with approximately seven to over 30 transceivers depending on its configuration. As a result, communication needs to be coordinated. However, due to the directionality of the signal, all transceivers that can communicate with the same transceiver act as hidden terminals for each other. Such multiple hidden terminals can lead to a masked node problem [26] that causes collisions if a regular ready-to-send/clear-to-send (RTS/CTS)-based MAC protocol [27] is used.

Therefore, we adopt a dual busy-tone multiple access (DBTMA) [28], [29] channel arbitration/reservation scheme. DBTMA is based on an RTS/CTS protocol, but it employs an additional out-of-band tone to indicate whether the transceivers are transmitting or receiving data. This tone resolves the masked node problem by enabling nodes both at the sending and receiving end to know whether other nodes are already using the wireless channel.

We use a fraction of the dedicated frequency channel for this tone and control messages using FDD so that they do not interfere with the data channel.

E. Y-Switch Implementation

The Y-switch is a simple customized piece of hardware that plays an important role in a Cayley datacenter. High-level schematic of this switch is shown in Fig. 6. When the Y-switch receives a packet, it parses the packet header and forwards the packet to the local machine or one of the transceivers.[2] The decisions are made by searching through one of the three routing tables described in Section III-C.1. To analyze the feasibility of the proposed Y-switch design, we implemented the Y-switch design for Xilinx FPGA in Simulink [30] and verified that, for an FPGA running at 270 MHz, its switching delay is less than 4 ns.

[2]Note that the Y-switches could also share the main memory resident on the server to buffer packets if necessary.

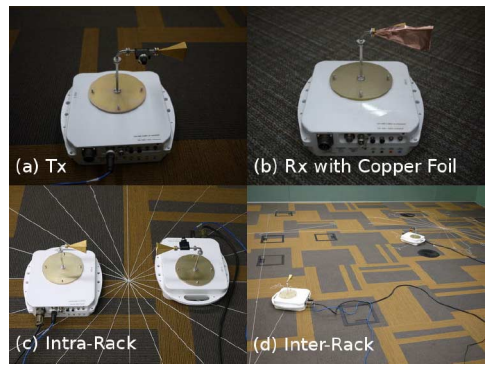

Fig. 7. 60-GHz Tx, Rx, and measurements on a Cayley datacenter floor plan. (a) Tx. (b) Rx with copper foil. (c) Intra-rack. (d) Inter-rack.

IV. Physical Validation

Before evaluating the performance of Cayley datacenters, we validate the assumptions behind the Cayley design with physical 60-GHz hardware. Specifically, we quantify communication characteristics and investigate the possibility of interference problems that may interfere with realizing the Cayley datacenter.

We conduct our experiments using Terabeam/HXI 60-GHz transceivers [18] [Fig. 7(a)]. While the Terabeam/HXI transceivers are older and therefore not identical to the Georgia Tech's transceiver described in Section II, they provide a good baseline for characterizing 60-GHz RF signals. This is a conservative platform, previously used in [31], over which modern hardware would provide further improvements. For instance, the Terabeam antennas are large and emit relatively broad sidelobes, and the signal-guiding horns catch some unwanted signals. In contrast, recently proposed CMOS-based designs can be smaller than a dime, effectively suppress sidelobes, and do not use signal-guiding horns at all [6], [32]. To compensate for the noise stemming from the older horn design, we augment one side of the receiver's horn with a copper foil [Fig. 7(b)]. The devices are statically configured to emit signals in a $\theta = 15^\circ$ arc, which is narrower than the Georgia Tech's transceiver.

We validate our model with physical hardware by first measuring how the received signal strength (RSS) varies as a function of the angle between the transmitter and receiver. We then build a real-size floor plan of a Cayley datacenter with a 2×3 grid of racks based on Table II, place transmitter–receiver pairs, and examine whether the signal strength is sufficient for communication [Fig. 7(c) and (d)]. Finally, we quantify the amount of interference for all possible receiver and transmitter pairs in intra-rack space, in inter-rack space both between adjacent and nonadjacent racks, and in different rack stories (Fig. 8). Due to the symmetric circular structure of racks on a regular grid, evaluating a subset of transceiver pairs on the 2×3 grid is sufficient to cover all cases. In the following experiments, we primarily examine RSS as a measure of signal quality in relationship to

TABLE II
CAYLEY DATACENTER CONFIGURATIONS

Cayley datacenter parameter	Value
Inner radius	0.25 (meter)
Outer radius	0.89 (meter)
Distance between racks	1 (meter)
Height of each story	0.2 (meter)
# of servers per story	20
# of stories per rack	5
# of servers per rack	100
Bandwidth per wireless data link	10 Gbps
Bandwidth per wireless control link	2.5 Gbps
Switching delay in Y-switch	4 ns

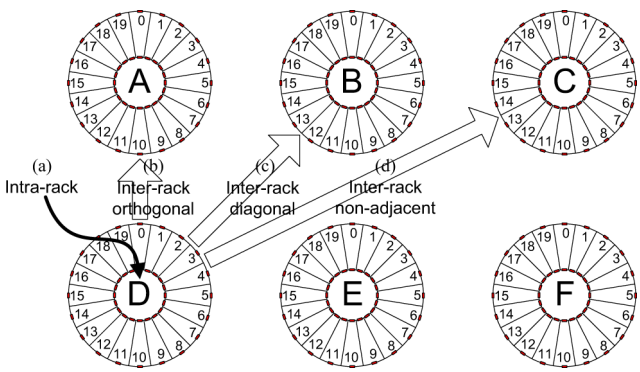

Fig. 8. Interference measurement summary. (a) Intra-rack. (b) Inter-rack orthogonal. (c) Inter-rack diagonal. (d) Inter-rack nonadjacent.

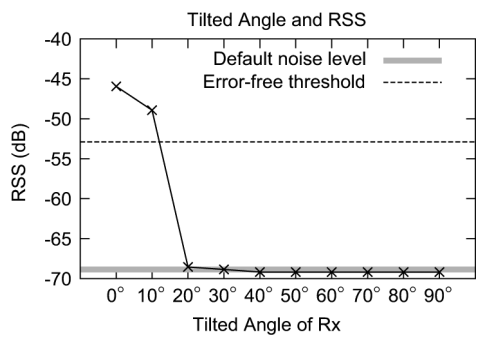

Fig. 9. Facing direction of Rx and RSS.

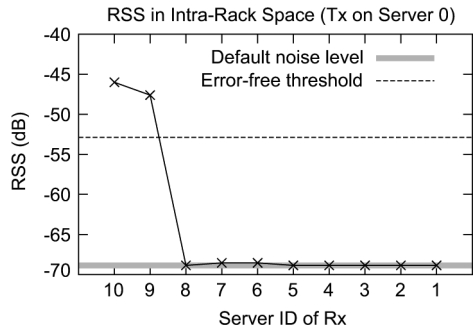

Fig. 10. RSS in intra-rack space.

a vendor-defined base.[3] We configure the transmission power of the Terabeam transmitter for all experiments such that a receiver directly facing the transmitter receives signal at -46 dB. This is a conservative level, as the minimum error-free RSS for this hardware is -53 dB in a noisy environment [33], and the typical default noise level we measured in a datacenter-like environment was approximately -69 dB.

A. Received Signal Strength and Facing Directions

The most basic assumption that the Cayley datacenter design makes of the underlying hardware is that a transmitter and a receiver pair can communicate when they are within each other's signal zone. To validate this assumption, we examine the signal strength of a transmitter-receiver pair, placed 1 m apart, as a function of the facing angle ε (i.e., $\alpha, \beta = 0°$ and $\delta = 1$ m in Fig. 1). In an ideal scenario with no interference, a receiver would not read any signals when ε exceeds θ.

Fig. 9 shows that the received signal strength is significantly above the error-free threshold when $\varepsilon \leq \theta = 15°$ and is negligible when $\varepsilon > 15°$. This confirms that the pair can communicate when oriented in the prescribed manner, and more importantly, that there is negligible interference from a transmitter on an unintended receiver whose reception zone does not cover the transmitter.

B. Intra-Rack Space

The cylindrical rack structure we propose divides free space into intra- and inter-rack spaces in order to achieve high free-space utilization. Such cylindrical racks would not be feasible if there was high interference within the dense intra-rack space [Fig. 8(a)]. To evaluate if this is the case, we measure the interference within a rack by measuring the signal strength at all receivers during a transmission.

Fig. 10 demonstrates that only the receivers within the 15° main signal lobe of the transmitter (receivers at positions 9 and 10 for transmitter 0) receive a signal at a reliable level. The rest of the servers do not receive any signal interference. In part, this is not surprising given the previous experiment. However, it confirms that any potential sidelobes and other leaked signals from the transmitter do not affect the adjacent receivers.

[3]The raw RSS value we get from the interface is the received signal strength indicator (RSSI). Due to some missing form factors, which the vendor did not provide, we translate RSSI to decibels using polynomial regression based on known RSSI to dB mappings. The coefficient of determination R^2 we get is 0.999993.

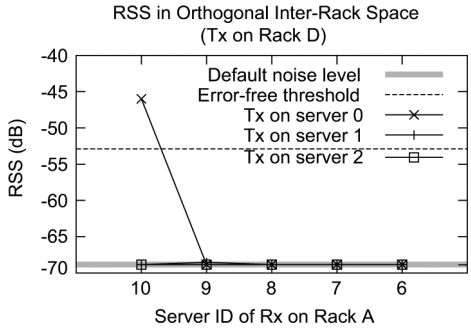

Fig. 11. RSS in inter-rack space between racks in orthogonal positions.

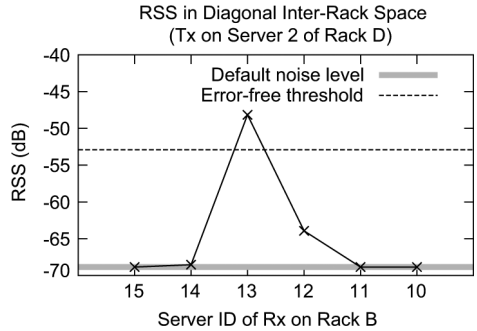

Fig. 12. RSS in inter-rack space between racks in diagonal positions.

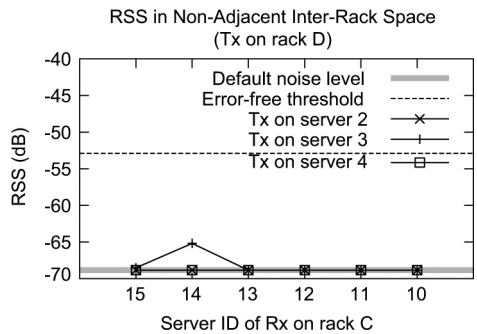

Fig. 13. RSS in inter-rack space between nonadjacent racks.

C. Orthogonal Inter-Rack Space

Eliminating all wires from a datacenter requires the use of wireless communication between racks. Such communication requires that the signals from nodes on a given rack can successfully traverse the free space between racks. We first examine the simple case of communication between racks placed at 90° to each other [Fig. 8(b)].

Fig. 11 shows that a transmitter–receiver pair can communicate between racks only when their signal zones are correctly aligned. For clarity, the graph omits symmetrically equivalent servers and plots the received signal strength of servers 6–10 on rack A. Other servers on rack A at positions less than 6 or greater than 14 show no signal from servers 0 to 2 on rack D. The graph shows that server 0 on rack D can transmit effectively to server 10 on rack A without any interference to any other servers, as expected.

D. Diagonal Inter-Rack Space

Cayley datacenters take advantage of diagonal links between racks in order to provide link diversity and increase bandwidth. We next validate whether the transceivers in our cylindrical racks can effectively utilize such diagonal paths [Fig. 8(c)].

Fig. 12 shows the received signal strength between diagonally oriented racks and demonstrates that the intended transmitter-receiver pairs can communicate successfully. Once again, the figure omits the symmetrical cases (e.g., transmitter on server 3 of rack D), and no signal from faraway servers (e.g., 0, 1, 4, 5 of rack D) reaches rack B at all. The signal strength in this experiment is as high as the orthogonal case despite the increased distance due to transmit power adjustment. The case of receiver on server 12 represents an edge case in our model: The signal strength is slightly above the background level because the node is located right at the boundary of the transmission cone. This signal level, while not sufficient to enable reliable communication, can potentially pose an interference problem. To avoid this problem, one can slightly increase the transmitter's signal's angle so that it sends a stronger signal. Alternatively, one can narrow the transmitter's signal angle to eliminate the signal spillover.

E. Nonadjacent Racks

While Cayley datacenters utilize only the wireless links between adjacent racks, it is possible for signals from nonadjacent racks to interfere with each other [Fig. 8(d)]. We examine the attenuation of the signal between nonadjacent racks and quantify the impact of such interference.

Fig. 13 shows the impact of three transmitters on rack D and the receivers on nonadjacent rack C. The transmitters are calibrated to communicate with their adjacent racks B and E. The measurements show that receivers on rack C receive no signal or weak signal not strong enough for communication, but when multiple nonadjacent transmitters send the weak signal (i.e., transmitter on server 3 and receiver on server 14), the noise rate could potentially become too great. For this reason, we propose placing nonreflective curtains, made of conductors such as aluminum or copper foil, that block the unwanted signal. Such curtains can be placed in the empty triangles in Fig. 3(b) without impeding access.

F. Inter-Story Space

Finally, we examine the feasibility of routing packets along the z-axis, between the different stories on racks. To do so, we orient a transmitter–receiver pair exactly as they would be oriented when mounted on prism-shaped servers placed on different stories of a rack and examine signal strength as the receivers are displaced from 0° to 30° following the z-axis.

Fig. 14 shows that the signal is the strongest at the center of the main lobe and drops quickly toward the edge of the signal zone. When the receiver reaches the borderline (15°) of the signal, it only picks up a very weak signal. Once the receiver moves beyond the 15° point, it receives no signal. Overall, the signal strength drops very sharply towards the edge of the signal, and except for the 15° borderline case, transceivers on different stories can reliably communicate.

G. Summary

In summary, we have evaluated transceiver pairs in a Cayley datacenter and demonstrated that the signal between pairs that

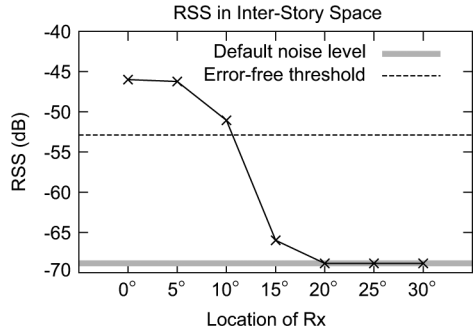

Fig. 14. RSS in inter-story space.

should communicate is strong and reliable, with little interference to unintended receivers. Calibrating the antenna or using conductor curtains can address the few borderline cases when the signal is weaker than expected or where there is potential interference. Although not described in detail, we also tested for potential constructive interference. We verified with two transmitters that even when multiple nodes transmit simultaneously, the signals do not interfere with the unintended receivers, namely the receivers in positions that received negligible or no signal in Figs. 9–Figs. 14. Overall, these physical experiments demonstrate that extant 60-GHz transceivers achieve the sharp attenuation and well-formed beam that can enable the directed communication topology of a Cayley datacenter while controlling interference.

V. Performance and Cost Analysis

In this section, we quantify the performance, failure resilience, and cost of Cayley datacenters in comparison to a fat-tree and a conventional wired datacenter (CDC).

A. Objectives

We seek to answer the following questions about the feasibility of wireless datacenters.

- *Performance*: How well does a Cayley datacenter perform and scale?
  By measuring the maximum aggregate bandwidth and packet delivery latency using a fine-grain packet level simulation model with different benchmarks, we compare the performance with fat-trees and CDCs.
- *Failure resilience*: How well can a Cayley datacenter handle failures?
  Unlike wired datacenters, server failures can affect routing reliability in Cayley datacenters because each server functions as a router. Thus, we measure the number of node pairs that can connect to each other, maximum aggregate bandwidth, and packet delivery latency under an increasing number of server failures.
- *Cost*: How cost-effective is a Cayley datacenter compared to wired datacenters?
  The wireless transceivers and Y-switches are not yet available in the market. We estimate and parameterize costs based on the technologies that wireless transceivers use and compare the price of a Cayley datacenter with a CDC based on the expected price range of 60-GHz transceivers. In addition, we compare the amount of power consumption in both datacenters.

TABLE III
Conventional Datacenter Configurations

Conventional datacenter parameter	Value
# of servers per rack	40
# of 1 GigE ports per TOR	40
# of 10 GigE port per TOR	2 to 4
# of 10 GigE port per AS	24
# of 10 GigE port per CS sub-unit	32
Buffer per port	16MB
Switching delay in TOR	6 μs
Switching delay in AS	3.2 μs
Switching delay in CS	5 μs

B. Test Environments

Because datacenters involve tens of thousands of servers and 60-GHz transceivers in Section II are not yet massively produced, it is impossible to build a full Cayley datacenter at the moment. Therefore, we built a fine-grained packet level simulation to evaluate the performance of different datacenters.

We model, simulate, and evaluate the MAC-layer protocol including busy tones, routing protocol, and relevant delays in the switches and communication links both for Cayley datacenters and CDCs. From the simulation, we can measure packet delivery latency, packet hops, number of packet collisions, number of packet drops from buffer overflow or timeout, and so on. The simulator can construct the three-dimensional wireless topology depending on the parameters such as the transceiver configurations, the distance between racks, and the size of servers. We also model, simulate, and evaluate the hierarchical topology of a fat-tree and a CDC given the number of ports and oversubscription rate of switches.

C. Base Configurations

Throughout this section, we evaluate Cayley datacenters along with fat-trees and CDCs with 10 K server nodes. Racks are positioned in a 10×10 grid for Cayley datacenters. We use the smallest configurable signal angle of 25° to maximize the number of concurrent wireless links in the Cayley datacenter and distance of one meter between racks for ergonomic reasons.

For CDCs and fat-trees, we simulate a conservative topology consisting of three levels of switches—top-of-rack switches (TOR), aggregation switches (AS), and core switches (CS)—in a commonly encountered oversubscribed hierarchical tree [34]. Oversubscription rate x indicates that among the total bandwidth, the rate of the bandwidth connecting the lower hierarchy to that connecting the upper hierarchy is $x : 1$. The oversubscription rates in a real datacenter are often larger than 10 and can increase to over several hundred [2], [35]. To be conservative, we configure CDCs to have oversubscription rates between 1 and 10, where the rate 1 represents the fat-tree.

The basic configurations for Cayley datacenters and CDCs are described in Tables II and III, respectively. The number of switches used for CDC varies depending on the oversubscription rate in each switch. The configuration and delays for the switches are based on the data sheets of Cisco products [36]–[38]. All switches use cut-through switching.

We focus exclusively on traffic within the datacenter, which account for more than 80% of the traffic even in client-facing web clouds [2]. Traffic in and out of the Cayley datacenter can

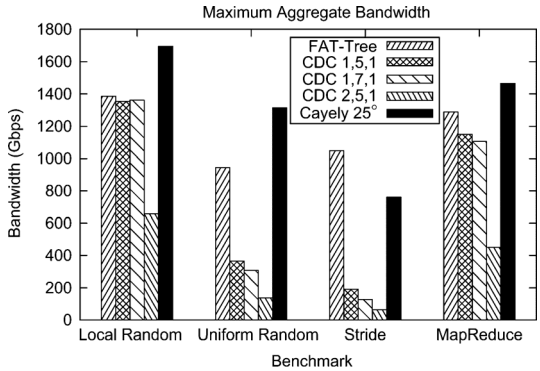

Fig. 15. Maximum aggregate bandwidth.

be accommodated without hot spots through transceivers on the walls and ceiling as well as wired injection points.

D. Performance

In this section, we measure the key performance characteristics, maximum aggregate bandwidth, and average and maximum packet delivery latency of Cayley datacenters, fat-trees, and CDCs using a detailed packet level simulator. The evaluation involves four benchmarks varying the packet injection rates and packet sizes.

- *Local random*: A source node sends packets to a random destination node within the same pod. The pod of a CDC is set to be the servers and switches connected under the same AS. The pod of a Cayley datacenter is set to be the servers in a 3×3 grid of racks.
- *Uniform random*: Source and destination nodes for a packet are randomly selected among all nodes with uniform probability.
- *Stride*: Source node with a global ID x sends packets to the destination node with ID $\mathrm{mod}(x + (\mathrm{total}\ \#\ \mathrm{of\ servers})/2, \mathrm{total}\ \#\ \mathrm{of\ servers})$.
- *MapReduce*: 1) A source node sends messages to the nodes in the same row of its rack. 2) The nodes that receive the messages send messages to the nodes in the same columns of their racks. 3) All the nodes that receive the messages exchange data with the servers in the same pod and outside the pod with 50% probability each. This benchmark resembles the MapReduce application used in Octant [39], where server nodes compute and exchange information with other nodes during the reduce phase.

We use different oversubscription rates in each level of switch in the CDC and use three numbers to indicate them: Each number represents the rate in TOR, AS, and CS in order. For example, (2, 5, 1) means the oversubscription rate of TOR is 2, that of AS is 5, and that of CS is 1, and a fat-tree is equivalent to (1, 1, 1).

1) Bandwidth: We measure the maximum aggregate bandwidth while every node pair is sending a burst of 500 packets.[4] The results are shown in Fig. 15, and the standard deviation of repeated experiments was within 2% of the plotted values.

For all cases, the Cayley datacenter shows higher maximum aggregate bandwidth than any CDC. A Cayley datacenter takes advantage of high bandwidth, oversubscription-free wireless channels, whereas a CDC uses 1-Gb/s links in each server and oversubscribed switches. The figure clearly shows the disadvantage of having oversubscribed switches in CDCs: When the majority of packets travel outside of a rack or above an AS, as in uniform random and stride, the bandwidth falls below 50% of Cayley datacenter's bandwidth.

Fat-trees perform noticeably better than all CDCs except for local random, where no packet travels above AS's. However, Cayley datacenters outperform fat-trees for all cases except the stride benchmark. Packets from the stride benchmark travel through the largest amount of hop counts, thus it penalizes the performance of the Cayley datacenter.

Although Cayley datacenters generally show higher maximum aggregate bandwidth for most cases, they have long bandwidth tails. Comparing the execution time, Cayley datacenters generally have shorter execution time than CDCs but slightly longer execution time than fat-trees (Fig. 16). The issue is that the MAC-layer contention allows only one transceiver to send at a time among seven to eight others that share the overlapping signal space. Still, for the realistic MapReduce benchmark, the Cayley datacenter performs the best.

2) Packet Delivery Latency: We measure packet delivery latencies by varying the packet injection rate and packet size. Figs. 17 and 18 show the average and maximum latencies, respectively.

The columns separate the type of benchmarks and the rows divide the packet sizes that we use for the experiments. Packets-per-server-per-second injection rates ranged from 100 to 500.

Local random is the most favorable and stride is the least favorable traffic for all datacenters from a latency point of view: Packets travel a longer distance in order of local random, MapReduce, uniform random, and stride.

Overall, the average packet delivery latencies of Cayley datacenters are an order of magnitude smaller (17–23 times) than those of fat-trees and all CDCs when the traffic load is small. This is because datacenter switches have relatively larger switching delay than the custom-designed Y-switch and Cayley datacenters use wider communication channels. For local random and MapReduce benchmarks that generate packets with relatively small network hops [Fig. 17(a) and (d)], Cayley datacenters outperform fat-trees and CDCs for almost all cases.

For all other benchmarks, CDC (2, 5, 1) performs noticeably worse than all others, especially when traffic load is large, because the TOR is oversubscribed. The latency of CDC (2, 5, 1) skyrockets once uniform random and stride traffic overloads the oversubscribed switches and packets start to drop due to buffer overflow [Fig. 17(b) and (c)]. Besides CDC (2, 5, 1), fat-tree and other CDCs maintain relatively stable average latencies except for during the peak load. The amount of traffic increases up to 8 MB/s per server during peak load: This is approximately the same amount of traffic generated as the peak traffic measured in an existing datacenter [40].

Cayley datacenters generally maintain lower latency than fat-trees and CDCs. The only case when the Cayley datacenters' latency is worse is near the peak load. When running uniform random and stride benchmarks under the peak load, Cayley datacenters deliver packets slower than fat-tree, CDC (1, 5, 1), and CDC (1, 7, 1) [the last row of Fig. 17(b) and (c)]. The numbers of average network hops for a Cayley datacenter are 11.5

[4]We configure the MapReduce benchmark to generate equivalent amount of packets.

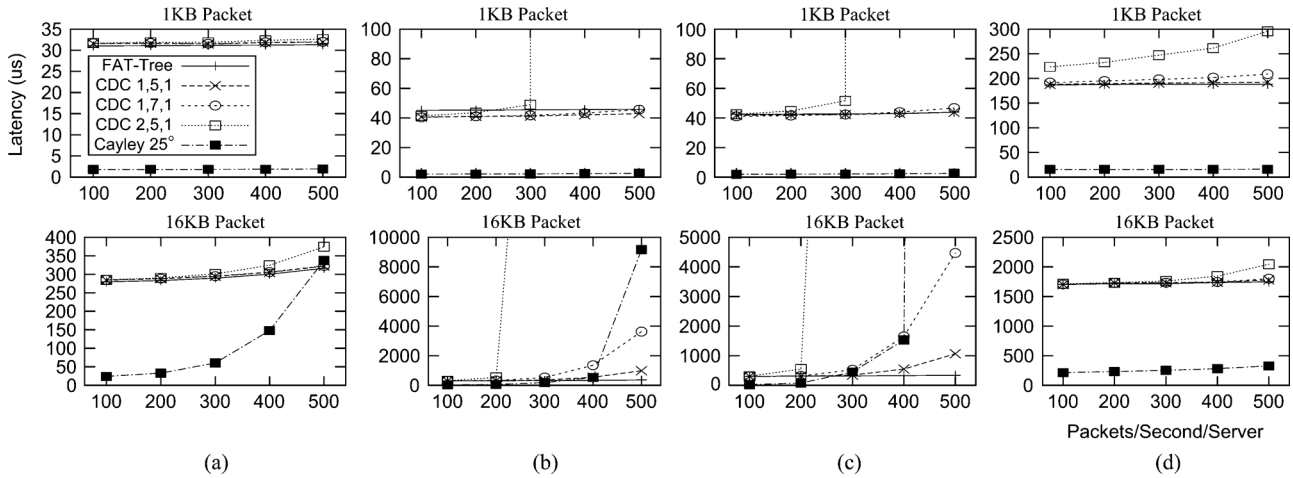

Fig. 17. Average packet delivery latency. (a) Local random. (b) Uniform random. (c) Stride. (d) MapReduce.

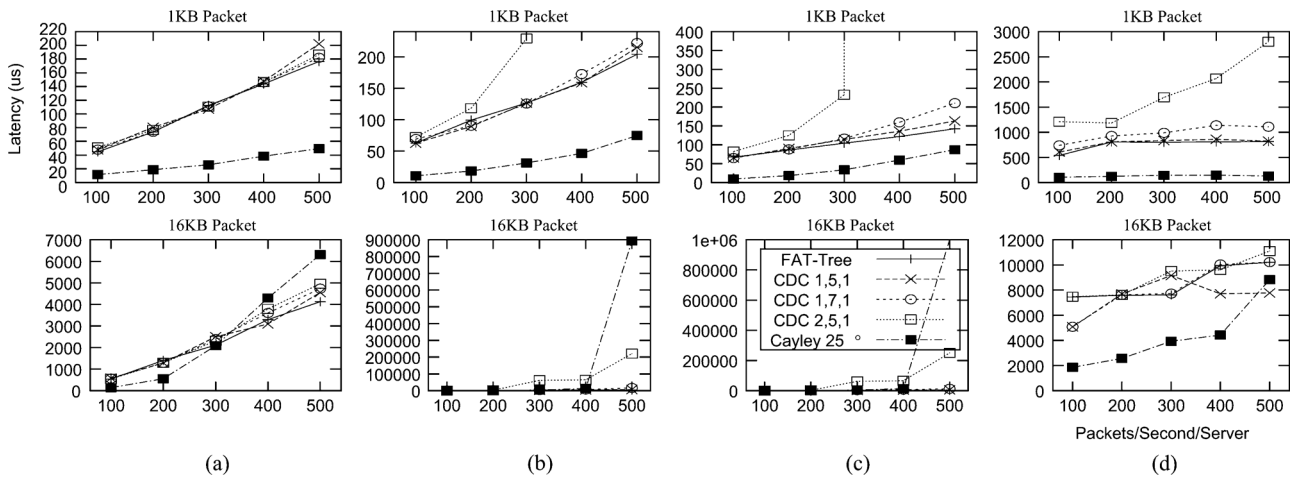

Fig. 18. Maximum packet delivery latency. (a) Local random. (b) Uniform random. (c) Stride. (d) MapReduce.

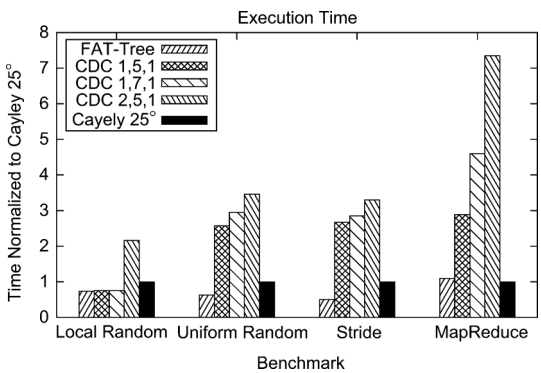

Fig. 16. Execution time.

and 12.4, whereas those of the tree-based datacenters are 5.9 and 6 for uniform random and stride benchmarks. Competing for a data channel at each hop with relatively large packets significantly degrades the performance of Cayley datacenters compared to fat-trees and CDC (1, 5, 1) and (1, 7, 1).

The maximum packet delivery latency shows the potential challenge in a Cayley datacenter (Fig. 18). Although the average latencies are better than CDCs, Cayley datacenters show a relatively steep increase in maximum latency as traffic load increases. Therefore, the gap between average and maximum latency for packet delivery becomes larger depending on the amount of traffic. However, except for under the peak traffic load, the maximum latency of the Cayley datacenter is less than 3.04 times as large as the latency of a fat-tree, and is smaller than CDCs for most cases. Therefore, Cayley datacenters are expected to show significantly better latency on average than fat-tree and CDCs, except under peak load for applications similar to stride.

In summary, except for handling the peak traffic for uniform random and stride benchmark, the Cayley datacenter performance is better than or comparable to fat-tree and CDC. As the average number of hops per packet increases, the performance of Cayley datacenters quickly decreases. This shows that Cayley datacenters may not also be as scalable as CDC, which has stable wired links with smaller number of network hops. Cayley datacenters may not be suitable to handle applications requiring large number of network hops per packet, but this type of application also penalizes the CDC performance as we observed for CDC (2, 5, 1). In reality, datacenter applications such as MapReduce usually resemble the local random benchmark, which does not saturate oversubscribed (aggregate) switches [35], [40]. Furthermore, the experimental results demonstrate that Cayley datacenters perform the best

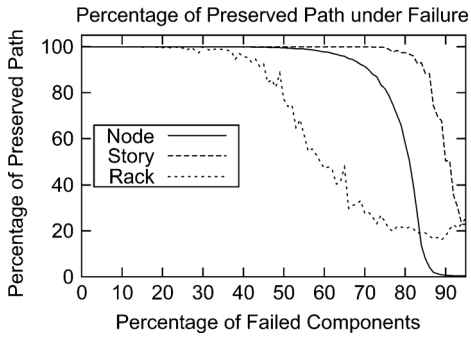

Fig. 19. Percentage of preserved path under failure.

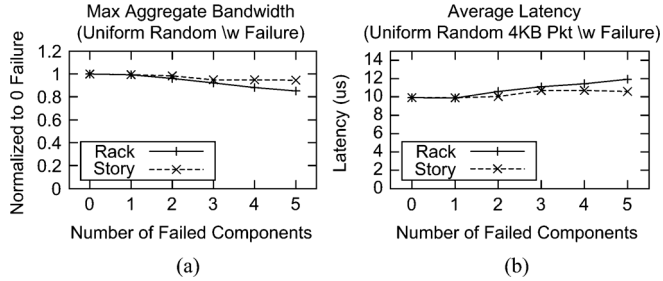

Fig. 20. Performance under failure. (a) Bandwidth. (b) Latency.

for MapReduce. Consequently, Cayley datacenters may be able to speed up a great portion of datacenter applications. Even for larger-scale datacenters, engineering the application's traffic pattern as in [41] will enable applications to run in Cayley datacenters more efficiently than in fat-trees and CDCs.

E. Failure Resilience

We evaluate how tolerant Cayley datacenters are to failures by investigating the impact of server failures on network connectivity and performance.

1) Connectivity: To measure how well Cayley datacenters preserve paths between two nodes, we select the failing nodes randomly in units of individual node, story, and rack. Fig. 19 shows the preserved connections between live nodes. We run 20 tests for each configuration and average the results. The average of standard deviation for the 20 run is less than 6.5%.

Server nodes start to disconnect when 20%, 59%, and 14% of the nodes, stories, and racks fail, respectively. However, over 99% of the network connections are preserved until more than 55% of individual nodes or stories fail. Over 90% of the connections are preserved until 45% of racks fail. Assuming failure rates of server nodes are the same in wireless datacenters as fat-tree-based datacenters and CDCs, then a Cayley datacenter can be more resilient to network failures. This is mainly because wireless datacenters do not have conventional switches that can be critical points of failure, and the failures catastrophic enough to partition a Cayley datacenter are very rare [2].

2) Performance Under Failure: While Section V-E.1 showed the connectivity of a Cayley datacenter, we explore the maximum aggregate bandwidth and packet delivery latency in the presence of failed components in this section. Using the uniform random workload, we generated a burst of 500 packets per node for the bandwidth measurement and injected 4-kB packets at a rate of 300 packets per second per server for the latency measurement. Fig. 20 shows the performance variation as the number of randomly selected racks and stories with failure increases from 0 to 5. We exclude cases where individual server nodes fail without any correlation. In our 10-K-node setting, failing one rack means failing 100 nodes, or 1% of total nodes. Typically, only less than 20 devices fail at the same time, and they are mostly fixed within a day in datacenters with 100 K server nodes [2]. With commodity servers, the rate of simultaneously failing nodes increases, but it is still within 5% range [43].

The average number of hops increases by a maximum of 0.04 for all cases compared to the base case without failure. The hop count shows that alternate routes, involving almost the same number of hops as the original route, can be found for most cases using our routing method.

The maximum aggregate bandwidth degrades by maximum 15.0% and 5.5% [Fig. 20(a)], and the latency increases by maximum 20.3% and 7.9% [Fig. 20(b)] compared to the base case as the number of failures for racks and stories increases, respectively. However, Cayley datacenters still perform better than CDCs and fat-trees without failure in terms of both bandwidth and latency. The bandwidth degradation shows a linear trend: The maximum aggregate bandwidth within Cayley datacenters is proportional to the number of live nodes. The increased latency mainly originates from detecting the failed nodes using timeout, but once the failure is detected, packets can be directly forwarded to alternate routes. The standard deviation of measurements was within 3% range.

F. Cost Comparison

It is complicated to compare two technologies when one is commercially mature and the other is yet to be commercialized. We can easily measure the cost of a fat-tree and a CDC, but the cost of a Cayley datacenter is not accurately measurable. However, we parameterize the costs of Cayley datacenters and compare the cost for different values of 60-GHz transceiver cost.

1) Hardware Cost: We compare the cost of the wireless and the wired datacenters based on the network configurations that we used so far. The price comparison can start from the NIC—typically priced at several tens of dollars [44]—and the Y-switch. In our system, we replace the NIC with the proposed simple Y-switch and at least two transceivers. Y-switches consist of simple core logic, host interface such as a PCI express bus, and interface controllers. Thus, we expect the price of a Y-switch to be comparable to an NIC.

The price differences between wireless and wired datacenters stem from the wireless transceivers and the switches. The prices of TOR, AS, and CS [42] and the cost required for CDC and fat-tree to connect 10 K servers using datacenter switches are summarized in Tables IV and V. The total price ranges from US $1.8 M to $2.4 M for CDCs and US $6.3 M for a fat-tree. Since the cost of a fat-tree can be very high, it should be able to use commodity switches as in [9], and the cost can vary much depending on the switch configuration. Thus, we mainly focus on the comparison between CDCs and Cayley datacenters.

The 60-GHz transceivers are expected to be inexpensive due to their level of integration, usage of mature silicon technologies (90 nm CMOS), and low power consumption that implies low-cost packaging. We cannot exactly predict the market price, but the total cost of network infrastructure excluding the Y-switch in Cayley datacenters can be expressed as a function

TABLE IV
CDC SWITCHES [42]

Component	Price ($)	Min Unit	Min unit price ($)
TOR	8000	1	4,800
AS	9,000	1	10,000
CS			358,500
CS subunit	42,000	1	42,000
CS chassis	12,000	1	12,000
CS power supply	3500	3	10,500

TABLE V
CDC NETWORKING EQUIPMENT COST FOR 10 K NODES

Config	#TOR	#AS	#CS sub-unit	#CS chassis	Cost ($)
2,5,1	250	26	8	1	1,818,500
1,7,1	250	48	12	2	2,229,000
1,5,1	250	52	16	2	2,437,000
fat-tree	250	88	96	10	6,337,000

$$Cost_{\text{Cayley}}(cost_{\text{t}}, N_{\text{server}}) = 2 \times cost_{\text{t}} \times N_{\text{server}} \quad (2)$$

where $cost_{\text{t}}$ is the price for a transceiver and N_{server} is the number of servers in a datacenter. From this function, we can find out that as long as $cost_{\text{t}}$ is less than US \$90, Cayley datacenters can connect 10 K servers with lower price than a CDC. Similarly, if $cost_{\text{t}}$ becomes US \$10, the cost of transceivers in Cayley datacenters can be 1/9 of CDC switches. Considering the rapidly dropping price of silicon chips [45], we expect the transceiver's price to quickly drop to less than US \$90 even if it starts with a high cost. This comparison excludes the wire price for CDC, so there is an additional margin, where $cost_{\text{t}}$ can grow higher to achieve lower cost than CDC.

2) Power Consumption: The maximum power consumption of a 60-GHz transceiver is less than 0.3 W [6]. If all 20 K transceivers on 10 K servers are operating at their peak power, the collective power consumption becomes 6 kW. TOR, AS, and a subunit of CS typically consume 176, 350, and 611 W, respectively [36], [37], [46]. In total, wired switches typically consumes 58–72 kW depending on the oversubscription rate for datacenter with 10 K servers. Thus, a Cayley datacenter can consume less than 1/12 to 1/10 of power to switch packets compared to a CDC.

Besides the lower price and power, lower maintenance costs stemming from the absence of wires and substantially increased tolerance to failure can be a strong point for wireless datacenters. In summary, we argue that 60 GHz could revolutionize datacenter construction and maintenance.

VI. PUTTING IT ALL TOGETHER

The summary of our findings throughout the evaluation of Cayley datacenters is presented. The merits of completely wireless Cayley datacenters over fat-trees and conventional datacenters are as follows.

- *Ease of maintenance through inherent fault tolerance*: Densely connected wireless datacenters have significantly greater resilience to failures than wired datacenters, in part because they do not have switches that can cause correlated loss of connectivity and in part because the wireless links provide great path diversity. Additionally, installing new or replacing failed components can be easier than in a CDC since only rewiring power cables is necessary.
- *Performance*: Cayley datacenters can perform better than or comparable to fat-trees and CDCs. Cayley datacenters achieve the highest maximum aggregate bandwidth for most benchmarks and deliver packets at a significantly lower latency, especially for MapReduce-like benchmarks, when traffic load is moderate, and even in the presence of failures.
- *Cost*: The price of networking components in a Cayley datacenter is expected to be less than those in CDC depending on the market price of wireless transceivers. Power consumption and expected maintenance costs are significantly lower than CDC.

Characteristics and limitations of Cayley datacenters are the following.

- *Interference*: Orientation of transceivers on the cylindrical racks and characteristics of 60-GHz signals limit the interference and enable reliable communication.
- *MAC-layer contention*: Sharing of the wireless channel followed by MAC-layer contention greatly influence the overall performance: The lower the contention, the greater the performance.
- *Hop count*: Performance depends on the number of network hops because each hop entails MAC-layer arbitration.
- *Scalability*: Due to the multihop nature of the topology, scalability is not as good as CDC. Yet, this limitation can be overcome by tuning applications to exhibit spatial locality when possible.

These points summarize the challenges, open problems, opportunities, benefits, and feasibility for designing a wireless datacenter.

VII. RELATED WORK

Ramachandran *et al.* [47] outlined the benefits and challenges for removing wires and introducing 60-GHz communication within a datacenter, and Vardhan *et al.* [48] explored the potentials of 60-GHz antennas emulating an existing tree-based topology. We share many of their insights and also conclude that 60-GHz wireless networks can improve conventional datacenters. Furthermore, we address some of the problems identified by the authors. We propose a novel rack-level architecture, use real 60-GHz transceivers and realistic parameters, and provide an extensive evaluation of the performance of the proposed wireless datacenters.

Although we focused on Georgia Tech's transmitter design [6], other research groups are also developing CMOS-based 60-GHz transceivers [49], [50]. While the technology was developed initially for home entertainment and mobile devices, other groups are looking at deploying it more broadly [51]. Our work on building completely wireless datacenters extends this line of research and tests the limits of 60-GHz technology.

Flyways [31] and [52] are wireless networks based on 60 GHz or 802.11n organized on top of wired datacenter racks. They provide supplementary networks for relieving congested wired links or for replacing some of the wired switches. In contrast, wireless links are the main communication channels in Cayley datacenters.

Zhang *et al.* [53] proposed using 3-D beamformation and ceiling reflection of 60-GHz signals in datacenters using networks like Flyways to reduce interference. Cayley datacenters use cone-shape 3-D beams, but use a novel cylindrical rack design to isolate signals and avoid interference.

A scalable datacenter network architecture by Al-Fares *et al.* [3] and Portland [9] employs commodity switches in lieu of expensive high-performance switches in datacenters and provides a scalable oversubscription-free network architecture. They achieve high performance at a lower cost, but significantly increase the number of wires.

CamCube consists of a three-dimensional wired torus network and APIs to support application-specific routing [41]. Although the motivation and goal of our paper is different from those of CamCube, combining their approach of application-specific routing is expected to enhance the performance of our Cayley datacenter design.

The MAC-layer protocol that we used [28], [29] is not developed specifically for Cayley datacenters; as a result, there may be inefficiencies that arise. Alternatively, there are other MAC-layer protocols developed specifically for 60-GHz technology and directional antennas [54]–[56], but they require global arbitrators or multiple directional antennas collectively pointing to all directions. These are not suitable for datacenters. Designing a specialized MAC-layer protocol for wireless datacenters is an open problem.

While our design adopted XY routing for Cayley datacenters, other routing protocols for interconnecting networks, such as [20], [57], and [58], can be adapted to our design.

VIII. Conclusion

In this paper, we proposed a radically novel methodology for building datacenters that replaces the existing massive wired switching fabric with wireless transceivers integrated within server nodes.

For brevity and simplicity of presentation, we explore the design space under the assumption that certain parameters such as topology and antenna performance are constant. Even in this reduced search space, we identify the strong potential of Cayley datacenters: While maintaining higher bandwidth, Cayley datacenters substantially outperform conventional datacenters and fat-trees with respect to latency, reliability, power consumption, and ease of maintenance. Issues that need further improvements are extreme scalability and performance under peak traffic regimes.

Cayley datacenters open up many avenues for future work. One could focus on each aspect of systems research related to datacenters and their applications and try to understand the ramifications of the new architecture. We feel that we have hardly scratched the surface of this new paradigm and that numerous improvements are attainable. Some interesting design considerations involve understanding the cost structure of individual nodes and how it scales with applications: Is it beneficial to parallelize the system into a substantially larger number of low-power low-cost less-powerful processors and support hardware? What data replications models yield best reliability versus traffic overhead balance? Could an additional global wireless network help with local congestion and MAC-layer issues such as the hidden terminal problem? What topology of nodes resolves the max-min degree of connectivity across the network? How should software components be placed within the unique topology offered by a Cayley datacenter? How does performance scale as the communication subband shifts higher in frequency? Would some degree of wired connectivity among servers internal to a single rack benefit performance? As the 60-GHz technology matures, we expect many of the issues mentioned here to be resolved and novel wireless networking architectures to be realized.

Acknowledgment

The authors thank the Georgia Tech team, ANCS Chairs, A. W. Moore and T. Wolf, IEEE/ACM Transactions on Networking editors, anonymous reviewers, and the following people for their contributions: H. Wang, S. Kandula, J. Padhye, V. Bahl, D. Harper, D. Maltz, D. Halperin, B. Kleinberg, D. Altinbuken, and T. Marian

References

[1] A. Greenberg, J. Hamilton, D. A. Maltz, and P. Patel, "The cost of a cloud: research problems in data center networks," *Comput. Commun. Rev.*, vol. 39, no. 1, pp. 68–73, Dec. 2008.

[2] A. Greenberg, J. R. Hamilton, N. Jain, S. Kandula, C. Kim, P. Lahiri, D. A. Maltz, P. Patel, and S. Sengupta, "Vl2: a scalable and flexible data center network," in *Proc. ACM SIGCOMM Conf. Appl., Technol., Archit., Protocols Comput. Commun.*, 2009, pp. 51–62.

[3] M. Al-Fares, A. Loukissas, and A. Vahdat, "A scalable, commodity data center network architecture," in *Proc. ACM SIGCOMM Conf. Appl., Technol., Archit., Protocols Comput. Commun.*, 2008, pp. 63–74.

[4] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, and M. Zaharia, "A view of cloud computing," *Commun. ACM*, vol. 53, no. 4, pp. 50–58, Apr. 2010.

[5] R. Buyya, C. S. Yeo, and S. Venugopal, "Market-oriented cloud computing: Vision, hype, and reality for delivering it services as computing utilities," in *Proc. IEEE Int. Conf. High Perform. Comput. Commun.*, 2008, pp. 5–13.

[6] S. Pinel, P. Sen, S. Sarkar, B. Perumana, D. Dawn, D. Yeh, F. Barale, M. Leung, E. Juntunen, P. Vadivelu, K. Chuang, P. Melet, G. Iyer, and J. Laskar, "60 GHz single-chip CMOS digital radios and phased array solutions for gaming and connectivity," *IEEE J. Sel. Areas Commun.*, vol. 27, no. 8, pp. 1347–1357, Oct. 2009.

[7] J. Nsenga, W. V. Thillo, F. Horlin, A. Bourdoux, and R. Lauwereins, "Comparison of OQPSK and CPM for communications at 60 GHz with a nonideal front end," *EURASIP J. Wireless Commun. Netw.*, vol. 2007, no. 1, pp. 51–51, Jan. 2007.

[8] A. Cayley, "On the theory of groups," *Amer. J. Math.*, vol. 11, no. 2, pp. 139–157.

[9] R. N. Mysore, A. Pamboris, N. Farrington, N. Huang, P. Miri, S. Radhakrishnan, V. Subramanya, and A. Vahdat, "Portland: a scalable fault-tolerant layer 2 data center network fabric," in *Proc. ACM SIGCOMM Conf. Appl., Technol., Archit., Protocols Comput. Commun.*, 2009, pp. 39–50.

[10] C. E. Leiserson, "Fat-trees: universal networks for hardware-efficient supercomputing," *IEEE Trans. Comput.*, vol. C-34, no. 10, pp. 892–901, Oct. 1985.

[11] V. Kvicera and M. Grabner, "Rain attenuation at 58 GHz: prediction versus long-term trial results," *EURASIP J. Wireless Commun. Netw.*, vol. 2007, no. 1, pp. 46–46, Jan. 2007.

[12] SiBeam, "Designing for high definition video with multi-gigabit wireless technologies," White Paper, Nov. 2005 [Online]. Available: http://www.sibeam.com/whtpapers/Designing_for_HD_11_05.pdf

[13] "IEEE 802.15 Working Group for WPAN," [Online]. Available: http://www.ieee802.org/15/

[14] "WG802.11—Wireless LAN Working Group," [Online]. Available: http://standards.ieee.org/develop/project/802.11ad.html

[15] Wi-Fi Alliance, Austin, TX, USA, "Wireless Gigabit Alliance," 2010 [Online]. Available: http://wirelessgigabitalliance.org

[16] *High Rate 60 GHz PHY, MAC and HDMI PAL*, Standard ECMA-387, ECMA International, Dec. 2008 [Online]. Available: http://www.ecma-international.org/publications/standards/Ecma-387.htm

[17] S. K. Yong and C.-C. Chong, "An overview of multigigabit wireless through millimeter wave technology: potentials and technical challenges," *EURASIP J. Wireless Commun. Netw.*, vol. 2007, no. 1, pp. 50–50, Jan. 2007.
[18] HXI, Harvard, MA, USA, "HXI," 2012 [Online]. Available: http://www.hxi.com
[19] K. Tang and R. Kamoua, "Cayley pseudo-random (CPR) protocol: a novel MAC protocol for dense wireless sensor networks," in *Proc. IEEE Wireless Commun. Netw. Conf.*, 2007, pp. 361–366.
[20] C. J. Glass, L. M. Ni, and L. M. Ni, "The turn model for adaptive routing," in *Proc. Int. Symp. High Perform. Comput. Archit.*, 1992, pp. 278–287.
[21] H. Krawczyk, "LFSR-based hashing and authentication," in *Proc. Int. Cryptol. Conf. Adv. Cryptol.*, 1994, pp. 129–139.
[22] P. Kermani and L. Kleinrock, "Virtual cut-through: a new computer communication switching technique," *Comput. Netw*, vol. 3, pp. 267–286, 1979.
[23] J. I. Choi, M. Jain, K. Srinivasan, P. Levis, and S. Katti, "Achieving single channel, full duplex wireless communication," in *Proc. Int. Conf. Mobile Comput. Netw.*, 2010, pp. 1–12.
[24] S. Gollakota, S. D. Perli, and D. Katabi, "Interference alignment and cancellation," in *Proc. ACM SIGCOMM Conf. Appl., Technol., Archit., Protocols Comput. Commun.*, 2009, pp. 159–170.
[25] E. Kranakis, H. Singh, and J. Urrutia, "Compass routing on geometric networks," in *Proc. Can. Conf. Comput. Geom.*, 1999, pp. 51–54.
[26] S. Ray, J. Carruthers, and D. Starobinski, "Evaluation of the masked node problem in ad hoc wireless LANs," *IEEE Trans. Mobile Comput.*, vol. 4, no. 5, pp. 430–442, Sep. 2005.
[27] P. Karn, "MACA—A new channel access method for packet radio," in *Proc. ARRL Comput. Networking Conf.*, 1990, pp. 134–140.
[28] Z. Hass and J. Deng, "Dual busy tone multiple access (DBTMA)-a multiple access control scheme for ad hoc networks," *IEEE Trans. Commun.*, vol. 50, no. 6, pp. 975–985, Jun. 2002.
[29] Z. Huang, C.-C. Shen, C. Srisathapornphat, and C. Jaikaeo, "A busy-tone based directional MAC protocol for ad hoc networks," in *Proc. Int. Conf. Military Commun.*, 2002, vol. 2, pp. 1233–1238.
[30] MathWorks, Natick, MA, USA, "Simulink—Simulation and model-based design," [Online]. Available: http://www.mathworks.com/products/simulink/
[31] D. Halperin, S. Kandula, J. Padhye, P. Bahl, and D. Wetherall, "Augmenting data center networks with multi-gigabit wireless links," in *Proc. ACM SIGCOMM Conf. Appl., Technol., Archit., Protocols Comput. Commun.*, 2011, pp. 38–49.
[32] M. M. Khodier and C. G. Christodoulou, "Linear array geometry synthesis with minimum sidelobe level and null control using particle swarm optimization," *IEEE Trans. Antennas Propag.*, vol. 53, no. 8, pp. 2674–2679, Aug. 2005.
[33] *"Terabeam Gigalink Field Installation and Service Manual,"* Terabeam, San Jose, CA, USA, 2003.
[34] Cisco, San Jose, CA, USA, "Cisco data center infrastructure 2.5 design guide," Mar. 2010 [Online]. Available: http://www.cisco.com/en/US/docs/solutions/Enterprise/Data_Center/DC_Infra2_5/DCI_SRND_2_5_book.html
[35] T. Benson, A. Akella, and D. A. Maltz, "Network traffic characteristics of data centers in the wild," in *Proc. Internet Meas. Conf.*, 2010, pp. 267–280.
[36] Cisco, San Jose, CA, USA, "Cisco catalyst 4948 switch," 2010 [Online]. Available: http://www.cisco.com/en/US/prod/collateral/switches/ps5718/ps6021/product_data_sheet0900aecd8017a72e.pdf
[37] Cisco, San Jose, CA, USA, "Cisco Nexus 5000 series architecture: The building blocks of the unified fabric," 2009 [Online]. Available: http://www.cisco.com/en/US/prod/collateral/switches/ps9441/ps9670/white_paper_c11-462176.pdf
[38] Cisco, San Jose, CA, USA, "Cisco Nexus 7000 F1-series 32-port 1 and 10 Gigabit Ethernet module," 2013 [Online]. Available: http://www.cisco.com/en/US/prod/collateral/switches/ps9441/ps9402/data_sheet_c78-605622.pdf
[39] B. Wong, I. Stoyanov, and E. G. Sirer, "Octant: A Comprehensive framework for the geolocalization of Internet hosts," 2007.
[40] S. Kandula, S. Sengupta, A. Greenberg, P. Patel, and R. Chaiken, "The nature of data center traffic: Measurements & analysis," in *Proc. Internet Meas. Conf.*, 2009, pp. 202–208.
[41] H. Abu-Libdeh, P. Costa, A. Rowstron, G. O'Shea, and A. Donnelly, "Symbiotic routing in future data centers," in *Proc. ACM SIGCOMM Conf. Appl., Technol., Archit., Protocols Comput. Commun.*, 2010, pp. 51–62.
[42] PEPPM, "Cisco current price list," Jan. 2012 [Online]. Available: http://www.peppm.org/Products/cisco/price.pdf
[43] D. Ford, F. Labelle, F. I. Popovici, M. Stokely, V.-A. Truong, L. Barroso, C. Grimes, and S. Quinlan, "Availability in globally distributed storage systems," in *Proc. USENIX Conf. Oper. Syst. Design Implement.*, 2010, pp. 1–7.
[44] Newegg.com, Whittier, CA, USA, "Intel PWLA8391GT 10/100/1000 Mbps PCI PRO/1000 GT desktop adapter 1 x RJ45," Jan. 2012 [Online]. Available: http://www.newegg.com/Product/Product.aspx?Item=N82E16833106121
[45] C. Gianpaolo, D. M. Xavier, S. Regine, L. N. Van Wassenhove, and W. Linda, "Inventory-driven costs," *Harvard Business Rev.*, vol. 83, no. 3, pp. 135–141, Oct. 2005.
[46] Cisco, San Jose, CA, USA, "Cisco Nexus 7000 series switches environment data sheet," 2010 [Online]. Available: http://www.cisco.com/en/US/prod/collateral/switches/ps9441/ps9402/ps9512/Data_Sheet_C78-437759.html
[47] K. Ramachandran, R. Kokku, R. Mahindra, and S. Rangarajan, "60 GHz data-center networking: Wireless = worry less?," NEC Tech. Rep., 2008.
[48] H. Vardhan, N. Thomas, S.-R. Ryu, B. Banerjee, and R. Prakash, "Wireless data center with millimeter wave network," in *Proc. Conf. Global Telecommun.*, 2010, pp. 1–6.
[49] B. Floyd, S. Reynolds, U. Pfeiffer, T. Beukema, J. Grzyb, and C. Haymes, "A silicon 60 GHz receiver and transmitter chipset for broadband communications," in *Proc. IEEE Int. Solid-State Circuits Conf.*, 2006, pp. 649–658.
[50] M. Tanomura, Y. Hamada, S. Kishimoto, M. Ito, N. Orihashi, K. Maruhashi, and H. Shimawaki, "Tx and Rx front-ends for 60 GHz band in 90 nm standard bulk CMOS," in *Proc. IEEE Int. Solid-State Circuits Conf.*, 2008, pp. 558–635.
[51] A. M. Niknejad, "Siliconization of 60 GHz," *IEEE Microw. Mag.*, vol. 11, no. 1, pp. 78–85, Feb. 2010.
[52] Y. Katayama, K. Takano, N. Ohba, and D. Nakano, "Wireless data center networking with steered-beam mmwave links," in *Proc. IEEE Wireless Commun. Netw. Conf.*, 2011, pp. 2179–2184.
[53] W. Zhang, X. Zhou, L. Yang, Z. Zhang, B. Y. Zhao, and H. Zheng, "3D beamforming for wireless data centers," in *Proc. Workshop Hot Topics Netw.*, 2011, pp. 4:1–4:6.
[54] X. Chen, J. Lu, and Z. Zhou, "An enhanced high-rate WPAN MAC for mesh networks with dynamic bandwidth management," in *Proc. Conf. Global Telecommun.*, 2005, pp. 3408–3412.
[55] S. Singh, F. Ziliotto, U. Madhow, E. Belding, and M. Rodwell, "Millimeter wave WPAN: cross-layer modeling and multi-hop architecture," in *Proc. IEEE Int. Conf. Comput. Commun.*, 2007, pp. 2336–2340.
[56] T. Korakis, G. Jakllari, and L. Tassiulas, "A MAC protocol for full exploitation of directional antennas in ad-hoc wireless networks," in *Proc. Int. Symp. Mobile Ad Hoc Netw. Comput.*, 2003, pp. 98–107.
[57] J. Kim, D. Park, T. Theocharides, N. Vijaykrishnan, and C. R. Das, "A low latency router supporting adaptivity for on-chip interconnects," in *Proc. Design Autom. Conf.*, 2005, pp. 559–564.
[58] P. Gratz, B. Grot, and S. Keckler, "Regional congestion awareness for load balance in networks-on-chip," in *Proc. Int. Symp. High Perform. Comput. Archit.*, 2008, pp. 203–214.

Ji-Yong Shin is currently pursuing the Ph.D. degree in computer science at Cornell University, Ithaca, NY, USA.

He is interested in designing and implementing novel datacenter network architectures, cloud storage systems, and virtualized cloud infrastructures.

Emin Gün Sirer received the Ph.D. in computer science from the University of Washington, Seattle, WA, USA, in 2002.

He is an Associate Professor of computer science with Cornell University, Ithaca, NY, USA. His interests span distributed systems, networking, and operating systems.

Hakim Weatherspoon received the Ph.D. degree in computer and information science from the University of California, Berkeley, CA, USA, in 2006.

He is an Assistant Professor with the Department of Computer Science, Cornell University, Ithaca, NY, USA. His research interests cover various aspects of distributed systems and cloud computing.

Prof. Weatherspoon is an Alfred P. Sloan Fellow and recipient of an NSF CAREER Award, DARPA Computer Science Study Panel, IBM Faculty Award, the NetApp Faculty Fellowship, and Intel Early Career Faculty Honor.

Darko Kirovski received the Ph.D. degree in computer science from the University of California, Los Angeles, CA, USA, in 2001.

After splitting 12 years as a Researcher with the Crypto and Machine Learning teams at Microsoft Research, Redmond, WA, USA, he joined Jump Trading, Chicago, IL, USA, as a Quantitative Researcher in 2012. His interests include systems, security, and machine learning.