Augmenting Data Center Networks with Multi-Gigabit Wireless Links

Daniel Halperin*†, Srikanth Kandula†, Jitendra Padhye†, Paramvir Bahl†, and David Wetherall*
Microsoft Research† and University of Washington*

Abstract – The 60 GHz wireless technology that is now emerging has the potential to provide dense and extremely fast connectivity at low cost. In this paper, we explore its use to relieve hotspots in oversubscribed data center (DC) networks. By experimenting with prototype equipment, we show that the DC environment is well suited to a deployment of 60 GHz links contrary to concerns about interference and link reliability. Using directional antennas, many wireless links can run concurrently at multi-Gbps rates on top-of-rack (ToR) switches. The wired DC network can be used to sidestep several common wireless problems. By analyzing production traces of DC traffic for four real applications, we show that adding a small amount of network capacity in the form of *wireless flyways* to the wired DC network can improve performance. However, to be of significant value, we find that one hop indirect routing is needed. Informed by our 60 GHz experiments and DC traffic analysis, we present a design that uses DC traffic levels to select and adds flyways to the wired DC network. Trace-driven evaluations show that network-limited DC applications with predictable traffic workloads running on a 1:2 oversubscribed network can be sped up by 45% in 95% of the cases, with just one wireless device per ToR switch. With two devices, in 40% of the cases, the performance is identical to that of a non-oversubscribed network.

Categories and Subject Descriptors

C.2.1 [**Computer-Communication Networks**]: Network Architecture and Design—Wireless Communication

General Terms

Design, Experimentation, Measurement, Performance

1. INTRODUCTION

Millimeter wavelength wireless technology is rapidly being developed. Spectrum between 57–64 GHz, colloquially known as the 60 GHz band, is available world-wide for unlicensed use. The band contains over 80 times the bandwidth available for 802.11b/g at 2.4 GHz, and supports devices with multi-Gbps data rates. Furthermore, 60 GHz devices with directional antennas can be deployed densely, because the signal attenuates rapidly due to the high frequency. The VLSI technology has now matured to the point where 60 GHz radio hardware can be built using CMOS technology, and companies like SiBeam [26] promise to deliver 60 GHz devices at less than $10 per unit at OEM quantities. In summary, 60 GHz technology can lead to dense, high-bandwidth wireless connectivity at low cost.

SIGCOMM'11, August 15–19, 2011, Toronto, Ontario, Canada.

To date, 60 GHz technology has been explored for isolated point-to-point links. A common scenario is home entertainment, e.g., a Blu-Ray player that communicates wirelessly with a nearby television instead of using bulky HDMI cables.

In this paper, we consider the novel possibility of using 60 GHz links in a data center (DC), to augment the wired network. This is a promising approach to explore for several reasons. First, we note that the machines in a DC are densely packed, so wireless devices that provide high bandwidth over short ranges are a natural fit. Second, the radio environment is largely static since people and equipment move around infrequently, minimizing fluctuations in wireless link quality. Third, line-of-sight communication is achievable by mounting 60 GHz radios on top of racks. Finally, the wired DC network is available as a reliable channel for coordinating wireless devices, thereby simplifying many traditional wireless problems such as aligning directional senders and receivers, and interference avoidance.

Traditional, wired DC networks are tree-structured and oversubscribed to keep costs down [15]. For example, a typical DC rack comprises 40 machines connected to a top-of-the-rack (ToR) switch with 1 Gbps links. The ToR is connected to an aggregation switch (to network with other racks) with 10 Gbps links. Thus, the link from the ToR to the aggregation switch can be oversubscribed with a ratio of 1:4. However, each oversubscribed link is a potential hotspot that hinders some DC application. Recent research tackles this problem by combining many more links and switches with variants of multipath routing so that the core of the network is no longer oversubscribed [1, 8, 9]. Of course, this benefit comes with large material cost and implementation complexity [15]. Some designs require so many wires that cabling becomes a challenge [1], and most require "fork lift" [5] upgrades to the entire infrastructure.

In prior work [15], we argued instead for a more modest addition of links to relieve hotspots and boost application performance. The links, called *flyways*, add extra capacity to the base network to alleviate hotspots. When the traffic matrix is sparse (i.e. only a few ToR switches are hot), a small number of flyways can significantly improve performance, without the cost of building a fully non-oversubscribed network.

The basic design of a DC network with 60 GHz flyways is as follows. The *base wired network* is provisioned for the average case and can be oversubscribed. Each top-of-rack (ToR) switch is equipped with one or more 60 GHz wireless devices, with electronically steerable directional antennas. A central controller monitors DC traffic patterns, and switches the beams of the wireless devices to set up flyways between ToR switches that provide added bandwidth as needed.

Other researchers have explored use of fiber optic cables and MEMS switches [7, 30] for creating flyways. We believe that 60 GHz flyways are an attractive choice because wireless devices simplify DC upgrades, as no wiring changes are needed. Furthermore, 60 GHz technology is likely to become inexpensive as it is commoditized by consumer applications, while optical switches are not. Wireless devices can introduce additional issues as well—for example, with dynamic topology, the network management may become more

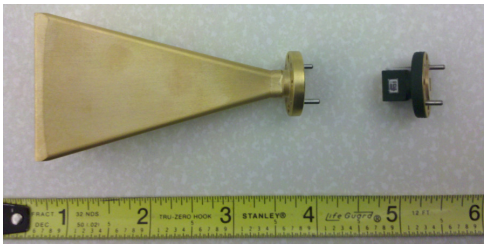


Figure 1: Narrow-beam (left) and wide-beam (right) horn antennas for 60 GHz. Note the small size.

complicated. However, before cost or management considerations come into play, we need to first understand whether 60 GHz wireless will perform well in the DC environment despite the challenges of interference and reliability. Answering this question is the primary focus of this paper.

We make three contributions. First, we report experiments with prototype 60 GHz devices and measurement-based simulations that show the feasibility of 60 GHz networks in the DC. To our knowledge, we are the first to report such results. Second, we show by analyzing four DC traffic traces that real workloads have few hotspots even when they lack predictable elephant flows. This implies that flyways can provide substantial benefits to real applications at low cost. Prior work [7, 30] has used synthetic workloads. Third, we present the design of a 60 GHz wireless flyway system motivated by our measurements. It differs from previous work on flyways [7, 30] in its use of indirect routing to obtain good gains from flyways. A trace-driven evaluation shows that in a 73-rack cluster with a 1:2 oversubscribed network, and just one wireless device per TOR, our system improves performance of a network-limited DC application by 45% in 95% of the cases. With two devices per ToR, the performance is identical to that of a non-oversubscribed network in 40% of the cases.

The rest of the paper proceeds as follows. We give background on 60 GHz in §2. Then, using experiments and simulations, we show that of 60 GHz data center deployments are feasible (§3). We then analyze DC traffic traces for different applications to understand what flyway characteristics are needed (§4). We present the design of our system (§5) followed by evaluation results (§6). We wrap up with a discussion (§7), related work (§8) and our conclusions (§9).

2. 60 GHz TECHNOLOGY

Recent advances in CMOS technology have reduced the cost of 60 GHz devices significantly, leading to commercial interest in indoor applications. This differs from initial, limited uses of 60 GHz for outdoor, point-to-point infrastructure [16, 29]. This section gives a brief primer on the 60 GHz physical layer and ongoing research and standardization efforts.

The nature of 60 GHz radio waves leads to significant challenges for operating high rate links. All other factors being equal, a 60 GHz link is roughly 55 dB (a factor of 300,000× worse) than a 2.4 GHz link in terms of the signal-to-noise ratio (SNR) that determines packet delivery. This is due to three factors. First, the free-space path loss is higher due to the small 5 mm wavelength (Friis' law [23]). Second, the channels are 100 times wider and thus 20 dB noisier [10, 32] to enable high multi-Gbps bit-rates. Third, most commercial equipment uses only 10 mW transmit power (compared to 802.11's typical 50 mW) in order to meet regulations and energy budgets. To compensate for these losses, indoor 60 GHz technologies such as 802.11ad [10] target a *short range* of 10 meters, and 60 GHz links use *highly directional antennas*.

Directionality represents the primary novel aspect of 60 GHz technology and is the key to enable a dense indoor deployment of 60 GHz links. With directional antennas, 60 GHz links can support multi-Gbps rates over distances of several meters [31, 32]. Directional antenna effectiveness is inversely proportional to the square of the radio wavelength, and so the short wavelength of 60 GHz leads to compact antennas. For fixed links, such as long range outdoor deployments, physically directional antennas (e.g., pyramidal gain horns such as in Figure 1) are used. For dynamic links, such as indoor Wireless HDTV [32], phased array antennas are used. Phased array antennas can rapidly change their directional radiation pattern electronically, i.e. with no moving parts.

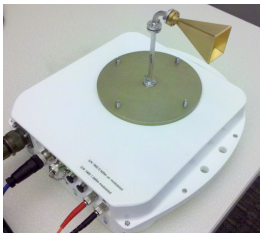


Figure 2: HXI device, paired with a horn antenna.

WirelessHD [32] and IEEE 802.11ad/WiGig [10, 31] are the two main ongoing efforts to standardize the PHY and MAC of the 60 GHz band. WirelessHD standardizes streaming media traffic in home entertainment systems. It was explicitly not intended for general data communication [32] and is a poor fit for our goals. IEEE 802.11ad enhances 802.11 for 60 GHz. It operates like standard Wi-Fi, with changes to the 802.11 PHY and MAC that support higher data rates that range from 385 Mbps to 6.76 Gbps. We use 802.11ad as the starting point for our flyways system.

3. 60 GHz LINKS IN THE DATA CENTER

The goal of this section is to find out whether 60 GHz links perform well in a DC environment. This primarily means that the links must offer steady, high throughput, even when deployed in a dense manner. We first measure 60 GHz propagation, link stability, and spatial reuse using prototype 60 GHz hardware. We then use these measurements to simulate dense collections of 60 GHz links in the DC. The results from this section guide our system design (§5).

3.1 Hardware

Our results are based on the device shown in Figure 2, built by HXI. It provides a full duplex, 60 GHz Gigabit Ethernet data link. It has a 1000BASE-SX fiber interface, and directly modulates the 1.25 Gbps line rate Ethernet protocol onto a 60 GHz carrier wave using On-Off-Keying (OOK). Rather than use any MAC protocol, this hardware employs frequency division to support the full duplex links: a paired set of devices operates on center frequencies that are 3.7 GHz apart sharing a single antenna per node. An SNMP management interface to the device provides continuous estimates of signal quality in the form of RSSI with 0.1 dB resolution.

This device interfaces with a removable antenna using the standard 60 GHz WR-15 waveguide. We use two physical directional antennas: a wide-beam horn antenna, marketed as a 10 dBi (60°) gain antenna, and a narrow-beam horn antenna, marketed as a 23 dBi (15°) gain antenna. Figure 1 shows how small these antennas are. We measured their radiation patterns in a large, free-space environment; unsurprisingly, the actual gain values (Figure 3) differ slightly from manufacturer claims. We will refer to these two antennas as wide-beam (WB) and narrow-beam (NB), respectively.

The measurement results in this paper use these fixed-beam directional antennas. However, we envision the use of electronically steerable phased array antennas that can change their radiation pat-

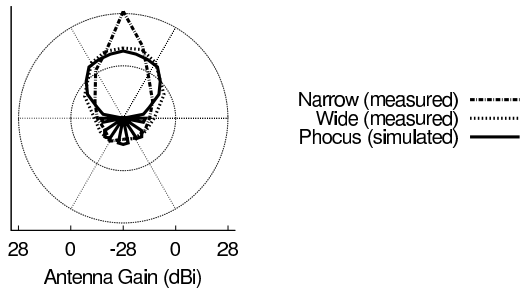


Figure 3: Radiation patterns for our antennas.

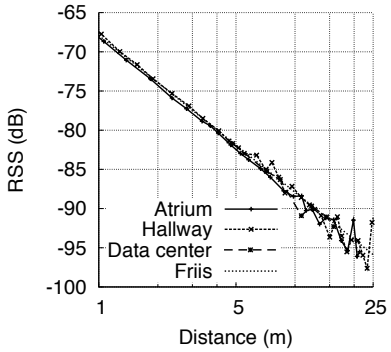


Figure 4: RSS vs Distance. RSS is relative to the transmitter power level, and fits the Friis model with exponent 2. The signal degrades by around 88 dB at 10 meters.

tern rapidly and with no moving parts. Phased arrays can be significantly more powerful than fixed-beam antennas, as they can generate patterns of variable beam width, control the amount and angle of side lobes, and can be used in more advanced ways to, e.g., null-form away from specific interferers [20]. As 60 GHz arrays are not yet available, we instead simulated the radiation pattern from the Geo-fencing project [25] that uses the commercially available Phocus phased array which operates at 2.4 GHz. Lacking real phased arrays for 60 GHz, we do not speculate on antenna properties, such as steering time. However, prior work [17, 19] shows that such antennas can be steered in hundreds of microseconds. Further details about the simulated Phocus array pattern and our assumptions about the 60 GHz phased arrays are in Appendix A.

Note that the Phocus pattern actually has smaller back and side lobes than our measured directional antennas (Figure 3). This is because, in our measurements with the NB and WB antennas, we conservatively assumed that any angle at which we measured no signal strength (e.g., sender facing directly away from the receiver) is in fact a received signal with strength just below the noise floor.

3.2 Signal propagation

We studied 60 GHz propagation in multiple environments. First, the atrium of our building, which resembles a free-space environment with no walls closer than 40 m from either end of the link. Second, a 1.5 m wide interior hallway, where multiple paths and physical obstructions exist. We chose line-of-sight environments as they come closest to the space on top of racks in our data centers. Finally, we measured propagation across the tops of rows of racks in a production data center, similar to the way in which wireless flyways could be deployed. As real data centers do, this production environment has a low ceiling, rows of racks (Figure 9), pipes for cabling as well as to and from the cooling systems and metal cages. In each scenario, we set up one sender and one receiver and varied the distance between the two, measuring the signal strength at the receiver at each step.

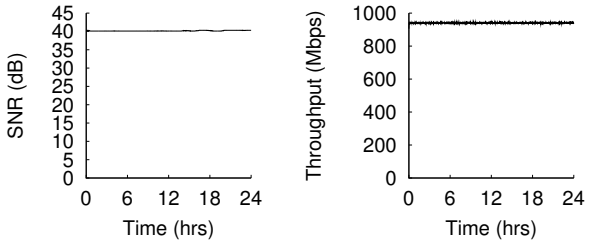


Figure 5: SNR and TCP are stable for 24 h in a data center.

The results are shown in Figure 4. We see that signal strength degrades rapidly with distance. The path exponent is 2, reflecting near-perfect Friis free-space propagation. Prior studies show that line-of-sight links in multi-path environments (waves in the 900–2400 MHz frequency with omni-directional antennas) have path exponents between 1.6–1.8 [23]. Thus, we believe that our directional antennas effectively mitigate the impact of multi-path. In fact, even at distances of 25 m, the signal variation (likely due to multi-path) is no more than 3 dB in the atrium and 5 dB in the hallway. This conclusion is supported by prior 60 GHz measurements [18] that showed that directionality at just one side of the link greatly reduced indoor multi-path effects.

These results show that the Friis model is appropriate for indoor line-of-sight 60 GHz links when the endpoints use narrow directional antennas.

3.3 Link stability

The adjective "flaky" is often associated with performance of wireless links, and is a potential concern for using wireless links in the DC. However, the performance variability seen in typical WLAN/Wi-Fi deployments comes from device mobility, environmental movement (people, doors opening and closing), temperature changes, and interference. The data center offers a stable, temperature-controlled environment, with infrequent movement of equipment, people, or doors. With devices mounted on top of racks and using directional antennas, the impact of these movements is even less. There is also no external interference in the 60 GHz band due to high attenuation by atmospheric oxygen and by walls. Thus, we expect individual links to be extremely stable.

To verify link stability, we set up a 60 GHz link in our data center using HXI devices with NB antennas. We deployed the devices atop two racks, facing each other across an aisle. We ran a long-lived TCP flow (using `iperf`) for 24 hours across two normal workdays, measuring throughput and SNR information every second. During the last five minutes of the measurement, one of the authors repeatedly walked under the link.

Figure 5 shows the link SNR and TCP throughput over the 24 hour period. TCP throughput achieves the full 1 Gbps rate, with almost no variation. In fact, none of the 1 s RSSI samples was off the average by more than 0.1 dB. The throughput curve shows that all the end-to-end components, not just the wireless link, are stable as perceived by the application. Even in the last five minutes, there is no variation in the throughput.

To provide a counterpoint, we set up a link with the same hardware, but at 3 feet above the ground. We then walked across it. Figure 6 shows the resulting variation due to line-of-sight obstruction.

These results show that in a typical DC, line-of-sight 60 GHz links set up at rack height provide stable performance.

3.4 Interference (Spatial reuse)

So far, we have studied wireless link properties in isolation. However, our system will require multiple flyways to be active simulta-

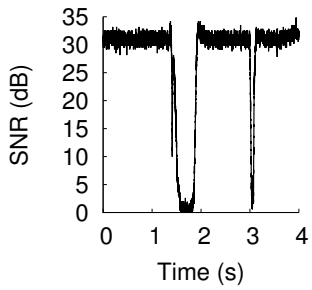


Figure 6: SNR fluctuates wildly when people walk (left) or wave hands (right) across the line-of-sight path.

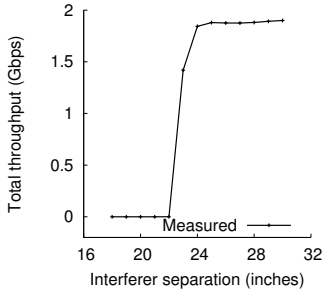


Figure 7: Interference experiment results

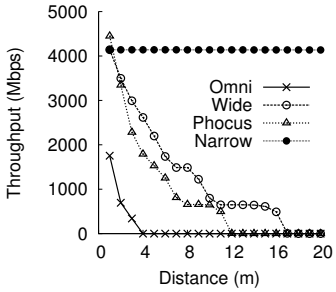


Figure 8: TCP throughput at various distances when sender and receiver both use directional antennas of various gains.

neously. Interference between flyways must be mitigated for good performance. This can be accomplished in a number of ways: by using multiple channels, by using directional antennas at both the sender and the receiver, and by carefully controlling which flyways are activated. We use all these techniques in our system design, but the bulk of interference mitigation happens due to directional antennas.

We now devise an experiment to see how directionality impacts spatial reuse. We configured two parallel links using HXI devices equipped with NB antennas. Recall that these links use frequency division to support bidirectional communication; we configured the links so that nodes facing in the same direction used the same frequency to maximize interference. We separated source and destination by a fixed 85 inches to mimic the width of an aisle, and varied the separation between the links in small increments. At each position, each source sends a greedy TCP flow to its destination. The cumulative throughput, shown in Figure 7, indicates whether the two links interfere with each other. Note that this prototype hardware has no MAC and uses no physical- or link-layer backoff, so the links interfere completely or not at all. We see that parallel links closer than 24 inches interfere, but directional antennas enable them to coexist perfectly with slightly more separation. Note that 24 inches is about 1 rack wide, and with 3 available 802.11ad channels, a large number of flyways can operate simultaneously.

These results show that directional antennas can isolate links and enable spatial reuse.

Later in this section, we will study the impact of interference with more links, and at various data rates, using simulations.

3.5 TCP throughput

In §3.3, we saw that a 60 GHz link set up over an aisle can provide stable 1 Gbps throughput. That throughput, however, was limited by the capabilities of HXI equipment. To get a better idea of what TCP throughput a full-fledged 802.11ad link can support, we rely on packet-level simulations. The simulations are done using the `ns-3` simulator [21], which we have extensively modified to model 60 GHz propagation, 802.11ad MAC, directional antennas and data center layouts. For more details on the changes we made to the simulator, see Appendix B.

We simulate the TCP throughput obtained over a 60 GHz link at various distances for the four antenna models. Note that these simulations account for overheads such as headers and various MAC overheads. The results are shown in Figure 8 and underscore the need for directional antennas. Omni-directional antennas provide no throughput at under 4 m, but modestly directional WB antennas can provide nearly 1 Gbps of throughput between nodes that are 15 m apart. With NB antennas, the TCP performance barely degrades with distance because the RSSI is sufficient to use the highest encoding rate of 6.76 Gbps even at 20 m. The performance

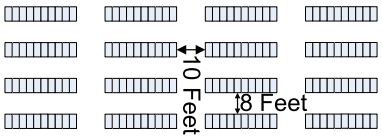


Figure 9: Partial top view of data center of a large search provider. Each row has ten 24x48 inch racks. The aisles are 10 and 8 feet wide, as shown. Overall area is roughly 14 m x 14 m.

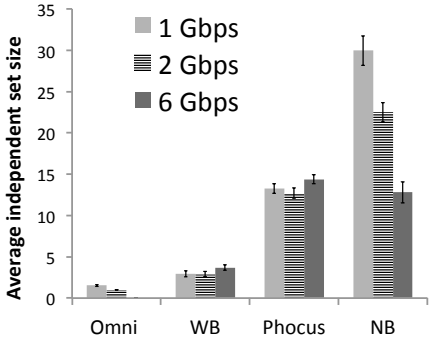


Figure 10: Number of flyways that can operate concurrently on one channel. Mean and standard deviation over 100 runs.

of the Phocus array is comparable to that of the WB antenna. Arrays with more elements (e.g., 30 as opposed to the 8 used here which we derived from [25]) should perform comparable to the NB antennas.

These results show that directional antennas are necessary to achieve high throughputs over links more than 1 m long.

Note that there is a gap between the maximum TCP throughput achieved ($\approx$ 4 Gbps) and the highest link transmission rate (6.76 Gbps). This gap is due to various wireless MAC and TCP overheads. In Appendix C, we describe some ideas on how to reduce these overheads by exploiting the unique hybrid nature of a wired data center network enhanced with wireless flyways.

3.6 Dense deployment of links

In §3.4, we showed that two high-rate 60 GHz links can coexist in close proximity. Using simulations, we now investigate the number of 60 GHz links that can operate simultaneously in a typical data center while still offering reasonable performance. We simulate the data center layout shown in Figure 9. This layout is based on an operational data center of a large search provider. We consider the case of a number of racks (160) connected to a single aggregation switch. We assume that each top-of-rack (ToR) switch is equipped with a single 60 GHz device, connected to a steerable antenna with a specified gain. All devices operate on the same channel and so may interfere.

Name	# Servers	Description
Cosmos	O(1K)	Map-Reduce
IndexSrv	O(10K)	Index lookup
Neon	O(100)	Car Simulation: HPC
3Cars	O(100)	Car Simulation: HPC

Table 1: Datasets

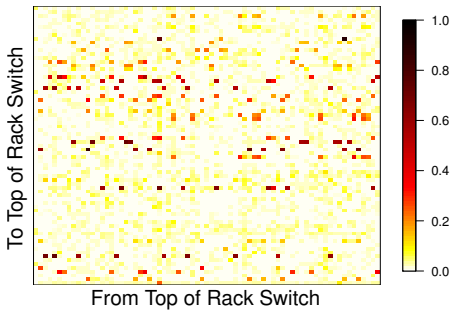


Figure 11: Traffic Demands (normalized) between ToR Switches.

We use the Monte-Carlo method to find maximal independent sets of flyway links [12]. Given n devices, note that $n \cdot (n-1)$ links are potentially feasible. A set of links is deemed independent if every link in the set provides some minimal throughput, even when all links in the set are active concurrently. The set is maximal if no other links can be added to it without violating the independence property. To test for independence, we simulate running long-lived TCP flows across the links. To limit the complexity of the simulation, we allow each device to participate in only one link at a time.

For a given DC layout, the average size of the maximal independent set tells us how many flyways may be set up at the same time. It depends on the antenna used, and the minimum throughput required from each link. Figure 10 shows the average size and one standard deviation over 100 randomly generated maximal independent sets for various antenna gains and minimum throughputs. Since each ToR can participate in only one flyway, and we have 160 ToRs, the set size cannot exceed 80.

With a Phocus antenna array or NB antennas, the number of flyways that can operate together increases dramatically. If the ToRs are equipped with NB antennas, the average size of the independent set is more than 30 for 1 Gbps links. Note that this is with just one channel; the set size increases linearly with more channels. We shall see later in the paper that for our workloads, these numbers suffice to provide significant performance gains.

In summary, these results give us confidence that in a typical data center, a large number of 60 GHz links can operate while delivering desired performance.

4. ANALYZING DATA CENTER TRAFFIC

We now examine traffic from four real applications in the data center to understand how much value flyways can add.

4.1 Datasets

Table 1 summarizes the analyzed datasets. Together, these logs represent over 76 hours worth of traces, and over 114 terabytes of traffic. The Cosmos dataset was measured on a pre-production cluster with O(1K) servers running Dryad. It supports a data mining workload for a large web search engine. Jobs on this cluster are a mix of repetitive production scripts (e.g., hourly summaries) and jobs submitted by users. The IndexSrv dataset is from a production cluster with O(10K) servers. The cluster stores the web search index and assembles search results for queries. This workload is latency sensitive. Unlike the Cosmos cluster, links here rarely see high utilizations. In both clusters, we instrumented every server to log network send and read system calls and the amount of data involved. The next two datasets are from an HPC platform with O(100) servers spread across 5 racks, running car simulation software. In most of the datasets, the servers were in racks underneath a single core switch pair. However, servers in the IndexSrv dataset spanned multiple core switches. In all clusters, ToR switches have enough backplane bandwidth such that intra-rack communication is only limited by the server NICs. However, the links connecting the ToR switches to the core are oversubscribed.

4.2 Estimating demand matrices

We want to understand the demands of data center applications without being impacted by the topology and capacity of the observed networks. To do so, we aggregate the traffic exchanged at time scales that are pertinent to the application. For example, most Dryad tasks finish within a few minutes, so the total traffic exchanged between racks in the Cosmos cluster every few minutes is a good indicator of application requirements. Unless otherwise noted, the datasets in this paper average traffic over 300 s periods to compute demands.

Consider an example demand matrix from the Cosmos dataset; Figure 11 depicts a heat map of the demands between pairs of the ToR switches. The color palette is on a logarithmic scale, i.e., black corresponds to the largest demand entry D, deep red (0.5 on the scale) corresponds to $\sqrt{D}$ and white indicates zero demand.

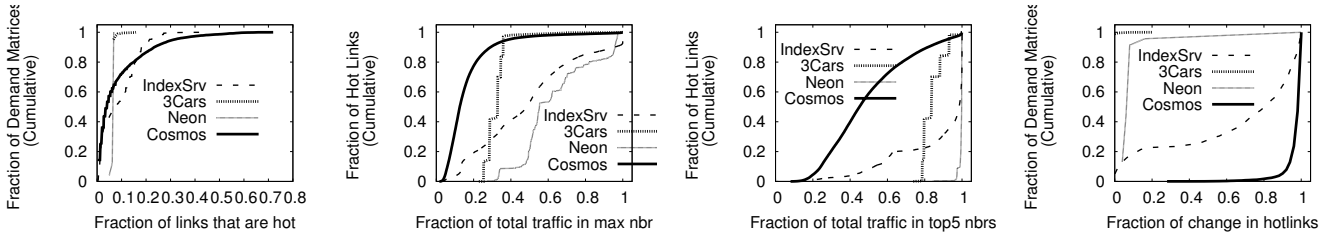
A few trends are apparent. First, only a few ToR pairs are hot, i.e., send or receive a large volume of traffic (darker dots). The bulk of the ToR pairs are yellow, i.e., less than $D^{\frac{1}{10}}$. Second, hot ToRs exchange much of their data with a few, but not all, of the other ToRs (horizontal and vertical streaks). It follows that providing additional bandwidth at hotspots would dramatically reduce the maximum temperature of the matrix. But, does this hold across all demand matrices? What form should the additional bandwidth take? How do the hotspots change over time? We look at these questions next.

4.3 Prevalence of hotspots

Figure 12(a) plots the fraction of hot links—links that are at least half as loaded as the most loaded link—in each of our datasets. In every dataset, over 60% of the matrices have fewer than 10% of their links hot at any time. In fact, every matrix in the Neon dataset has less than 7% hot links. *This means that for measured traffic patterns in the DC, avoiding oversubscription over the entire network may not be needed.* Instead, performance may be improved by adding capacity to a small set of links. We see in the evaluation of our system (§6) that, indeed, a few flyways have a large effect.

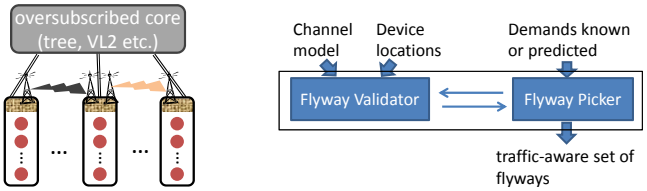
4.4 Traffic contributors to hotspots

To be useful, additional capacity provided to a hotspot should be able to offload a substantial fraction of the load. Prior proposals [7, 30] establish one additional flyway, in the form of an optical circuit, per congested link. Figure 12(b) estimates the maximum potential value of doing so, and suggests there will be little benefit in real data centers. Across hot links the traffic share of the largest ToR neighbor is quite small; on the Cosmos dataset, it is less than 20% for 80% of the matrices. In fact, Figure 12(c) shows that in some cases, even the top five ToR pairs can cumulatively add up to a small fraction of load on the hotlink. In other words, we find that hot links are associated with a high fan-in (or fan-out). This observation was a surprise; it means that *at hotspots, the existing proposals that offload traffic going to just the best neighbor would*



(a) Sparsity of Hotspots (b) Share of HotLink's traffic to Top Neighbor (c) Cumulative Share of HotLink's Traffic To Top Five Neighbors (d) Predictability of flyways

Figure 12: Nature of hotspots in measured DC traffic



(a) A flyway augmented network (b) Flyway controller

Figure 13: Proposed Architecture

be of limited value for real data center workloads. Even the optimistic wavelength multiplexing-based optical extensions proposed in Helios would not suffice in these cases. We propose and evaluate a one hop indirection technique (§5) that overcomes this weakness.

4.5 Predictability of hotspots

Figure 12(d) compares the change in the pairs of hot-links across consecutive matrices. We observe a dichotomy—some matrices are highly predictable, others are very unpredictable. In both HPC datasets, we see less than a 10% change in hot links whereas in the Cosmos dataset fewer than 10% of hot links repeat. We tried a few more complicated predictors, and find that the results are qualitatively similar. Likely, this is due to the nature of workload. While Cosmos churns work at the granularity of map and reduce tasks which typically last about a few minutes, work in HPC clusters manifests in more long lived groups. We also verify that flow sizes and arrival rates in the DC [6, 14] indicate that traffic in the DC lies in a fast-changing collection of medium-sized flows. This property of real DC workloads renders predictors that rely on identifying elephant flows [2] to be of less use.

Take-aways: In a broad study of many types of DC workload, we find that hotspots are sparse. The potential benefit of selectively providing additional bandwidth at these hotspots, as opposed to building for the worst case with non-oversubscribed networks, appears significant. We also see that real data center traffic matrices are more complex than synthetic workloads evaluated by prior proposals [7, 30], and flyway placement algorithms developed by these proposals are likely be of marginal value. The key issue is that hotspots are often correlated with a high fan-in (or fan-out) implying that to be useful traffic from (or to) many destinations needs to be offloaded. Our system design (§5) includes a novel one hop indirection method designed to resolve this problem.

5. FLYWAYS SYSTEM DESIGN

In this section, we propose a design for a DC network with flyways. The basic architecture is shown in Figure 13; we consider the set of racks in a cluster, each equipped with one or more wireless devices that can be used to construct flyways as needed. Our design is independent of the specific topology used in the oversubscribed core, which could be the typical tree structure, or recent proposals for non-oversubscribed networks [1, 8, 9] with proportionally fewer switches and links.

Our goal is to configure the flyway links and the routing to improve the time to satisfy traffic demands. The metric of interest is the **completion time of the demands** (CTD), defined as the time it takes for the last flow to complete.

The system has three tasks: (i) measure and estimate traffic demands, (ii) decide which flyways to instantiate, and (iii) make appropriate routing changes to route traffic over flyways. Inputs to the system include the measured 60 GHz channel model, antenna characteristics, device locations and traffic demands if available. We focus on flyways instantiation, and discuss traffic estimation and routing only briefly (§5.3).

Computing an optimal choice of flyways is challenging since wireless constraints such as range and interference are hard to incorporate into a max-flow formulation. Hence, our design decomposes the problem into two sub-parts. A 'flyway picker' (§5.1) proposes flyways that will improve the completion time of demands. A measurement and channel-model driven 'flyway validator' (§5.2) confirms or rejects this proposal. The validator ensures that the system only adds feasible, non-interfering flyways. In addition, the validator also predicts how much capacity the flyways will have. This allows the picker to add the "approved" flyway and propose flyways for subsequent hotspots. The process repeats until no more flyways can be added. This decomposition is not optimal and there is room to improve. However, it finishes quickly, scales well and provides impressive gains, as we will show.

5.1 Choosing flyways

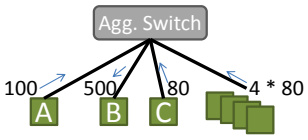
In this section, we will assume that traffic demands are known. We begin with an example. Consider the network in Figure 14(b). Six ToR switches A and C–G have traffic to send to ToR B. A has 100 units to send, whereas the rest each send 80 units. Each ToR has one wireless device connected to it. Wired link capacity in and out of the ToRs is 10 units/sec and for simplicity assume that these are the only potential bottlenecks. The downlink into B is the bottleneck here. It carries 500 units of traffic in total and takes 50 s to do so. Hence, completion time (CTD) is 50 s.

Suppose we add a flyway (capacity 3) from A to B to *improve performance of the straggler*, i.e., the ToR pair that sends the most amount of traffic on the bottleneck link and completes last, by bypassing the bottleneck. As Figure 14(c) shows, traffic on the bottleneck drops to 400 units, and time to complete drops to 40 s. However, as our traffic analysis shows, the straggler often contributes only a small proportion of the total demand on that link (in this case 100/500). Alleviating the straggler provides only 20% gains, reducing CTD to 40 s.

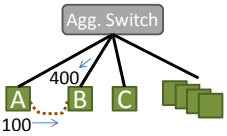
Note that there is room to spare on the flyway; the demand from A to B completes after 33.3 s, 6.7 s before traffic from C–G. Our datasets indicate that this is quite common; very few of the ToR pairs on hot links require substantial capacity. Hence, we also *allow*

Demand	A→B	100 units
	C,D,E,F,G→B	80 units
Capacity	wired links	10 units/s
	flyway: A→B	3 units/s
	flyway: C→B	6 units/s

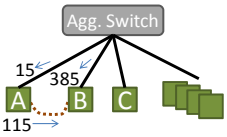
(a) Setup



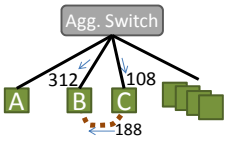
(b) Baseline: CTD=50s



(c) Direct flyway for the straggler: CTD=40s



(d) Allowing transit traffic: CTD=38.5s



(e) Greedy choice of flyways: CTD=31.2s

Figure 14: A motivating example: Greedy choice of flyways to add and allowing transit traffic through flyways are crucial.

indirect transit traffic to use the flyway, i.e., as Figure 14(d) shows, traffic from other sources to B bypasses the bottleneck by flowing via node A and the flyway. This improves CTD to $\frac{115}{3} = \frac{385}{10} = 38.5$ s.

Often the flyway to the straggler is infeasible or an inferior choice, the devices at either ends might be used up in earlier flyways or the link may interfere with an existing flyway or the ToR pairs might be too far apart. Allowing transit traffic ensures that *any flyway that can offload traffic on the bottleneck will be of use*, even if it is not between the straggler pair. In this case, it is more effective to enable the flyway from C to B, with twice the capacity of the flyway from A. This decision allows more traffic to be offloaded results in a CTD of $\frac{312}{10} = \frac{188}{6} = 31.2$ s.

Proposed algorithm: Our approach formalizes these two insights. By allowing transit traffic on a flyway, via indirection, we skirt the problem of high fan-in (or fan-out) that we saw to be correlated with congestion. Further, doing so opens up the space of potentially useful flyways, greedy choice among this set adds substantial value. In particular, at each step, we choose the flyway that diverts the most traffic away from the bottleneck link. For a congested downlink to ToR p, the best flyway will be from the ToR that has a high capacity flyway and sufficient available bandwidth on its downlink to allow transit traffic through, i.e.,

$$\arg\max_{\text{ToR } i} \min\left(C_{i\to p},\ \ D_{i\to p} + down_i\right).$$

The first term $C_{i\to p}$ denotes the capacity of the flyway. The amount of transit traffic is capped by $down_i$ the available bandwidth on the downlink to i and $D_{i\to p}$ is i's demand to p. Together, the second term indicates the maximum possible traffic that i can send to p. The corresponding expression of the best flyway for a congested uplink to ToR is similar,

$$\arg\max_{\text{ToR } i} \min\left(C_{p\to i},\ \ D_{p\to i} + up_i\right).$$

5.2 Validating flyway choice

The flyways validator determines whether a specified set of flyways can operate together — it computes the effects of interference and what capacity each link is likely to provide. It operates using the same principle as DIRC's conflict graph [17]: If we know how much signal is delivered between all pairs of nodes in all transmit and receive antenna orientations, we can combine these measurements with knowledge of which links are active, and how the antennas are oriented, to compute the SINR for all nodes. We can then use the simple SINR-based auto-rate algorithm (§B) to select rates.

Our SINR model (§B) is very conservative: we compute interference assuming *all nodes from all other flyways* send concurrently, and add an additional 3 dB. Hence, we disable carrier sense on our flyways links, managing contention between sender and receiver with other types of coordination (§C). Recall that both the SINR model and rate selection are appropriate for our data center environment because of the high directionality (§3 and [17]).

Obtaining the conflict graph: If there are N racks and K antenna orientations, the input to the validator is an $(NK)^2$-size table of received signal strengths. How can we generate this very large table? In the simulator, we compute delivered signal power using the models of antennas and signal propagation developed in §3. In a real DC deployment, we can measure it—a data center provides a stable, line-of-sight environment (§3.3) and a fixed set of nodes with known geographic coordinates. Hence, unlike DIRC's dynamic, non-line-of-sight environment with unknown client locations, we can afford to measure this table only once when the data center is configured, and measurements will remain valid over time. We can refresh entries in the table opportunistically, without disrupting ongoing wireless traffic, by having idle nodes measure signal strength from active senders at various receive antenna orientations and sharing these measurements, along with transmitter antenna orientation, over the wired network.

The table can be used not just to compute interference, but also to determine the best antenna orientation for two ToRs to communicate with each other, and the complex antenna orientation mechanisms prescribed in 802.11ad are no longer needed. In this paper, we evaluate antennas that use purely directional radiation patterns and point directly at their intended receivers. Advanced, more powerful antenna methods such as null-steering to avoid interference [20] could increase flyway concurrency, but we defer these to future work. Our results (§6) will show that even this simple antenna model is effective at improving data center performance.

5.3 Traffic estimation and routing

Traffic estimation and routing are not the main focus of this paper, and our system design in these areas is largely similar to prior work [30, 7, 8]. We describe it briefly for the sake of completeness.

Estimating traffic demands: Traffic demand can be estimated in one of two ways. First, for clusters that are orchestrated by cluster-wide schedulers (e.g., map-reduce schedulers such as Quincy [11]), logically co-locating our system with such a scheduler makes traffic demands visible. In this mode, our system can pick flyways appropriate for these demands. C-Through [30] takes a somewhat similar approach: it assumes that applications hint at their traffic demands.

Second, in clusters that have predictable traffic patterns, such as the HPC datasets we analyzed, we can use instrumentation to estimate current traffic demands and pick flyways appropriate for demands predicted based on these estimates. Such distributed, end-host based, traffic measurement instrumentation is already used, for e.g., at EC2 and Windows Azure, for billing and accounting, and can provide up-to-date inputs for our system as well.

We have designed a simple traffic estimation scheme that uses a shim layer (an NDIS filter driver) on servers to collect traffic statistics, in a manner similar to prior work [6, 14]. We use a simple moving average of estimates from the recent past [13]. This estimator works as well with our traces as the more complex alternatives that we tried. Micro-benchmarks show this estimator to be feasible at line rate with negligible increase in server load.

In future work, we will address traffic that is neither predictable nor orchestrated. See (§7) for some ideas.

Routing: We present a simple mechanism that routes traffic across the potentially multiple paths that are made feasible with flyways.

Our approach is straightforward and similar to prior work [8, 7]. We treat flyways as point-to-point links. Note that every path on the flyway transits through exactly one flyway link, so all that the routing does is to encapsulate packets to the appropriate interface address. For example, to send traffic via $A \rightarrow Core \rightarrow C \rightarrow B$, the servers underneath A encapsulate packets with the address of C's flyway interface to B. The flyway picker computes the fraction of traffic to flow on each path and relays these decisions to the servers. We have built this functionality into the aforementioned NDIS filter driver. Our micro-benchmarks tests on standard DC servers equipped with 1 Gbps NICs indicate that these operations can be performed at line speed with negligible additional load.

When changing the flyway setup, we simply disable encapsulation, and remove the added routes. The default routes on the ToR and Agg switches are never changed, and direct traffic on the wired network. Thus, as soon as we remove the flyway route, the traffic flows over wired links. Thus, during flyway changes (and flyway failures, if any), packets are simply sent over wired network.

6. EVALUATING FLYWAYS

In this section we combine our measurement- and standard-driven wireless models with our traces from real data centers to evaluate the practical benefits to data center workloads in oversubscribed networks that come from our wireless flyways system.

6.1 Methodology

Demands: We replay the traffic described in §4, which is measured from four different clusters and includes workloads from latency- and throughput-sensitive applications and highly tuned HPC applications.

Wireless models: We use the wireless physical and MAC layers and channel models described in §B. Here, we recall a few salient specifics: We use the three channels defined in 802.11ad to increase the number of concurrent links. Devices use a uniform 10 mW transmit power. The system uses the interference model and rate selection algorithms described in §B and the flyways validator described in §5.2. We use the 802.11ad OFDM rates, which peak at 6.76 Gbps, only about 85% of which is usable for traffic (§C).

Geography: We mimic the geographical layout of racks as per measurements from an open floor-plan data center (see Figure 9). We assume that each ToR is equipped with K wireless devices, often 1, which are mounted atop the rack. ToR switches in the observed data centers have a few unused ports for occasional network management tasks.

We compare these variants of our system:

Straggler is the simplest alternative, in which the picker proposes a flyway between the pair of ToRs taking the longest time to complete. If the validator accepts this proposal as safe then the flyway is added, and if not then the process terminates — the CTD cannot be further improved.

Transit augments Straggler by allowing for transit traffic on the added flyways. As we saw in §5, doing so improves performance by offloading more traffic from the bottleneck link, and potentially changes which link will next be the bottleneck.

Greedy augments Transit by preferentially picking, in each iteration, the flyway that offloads the most traffic from the bottleneck link. In practice, this results in using flyways between close-by nodes that have high capacity. As a side-effect, this process tends to add shorter links and thus results in more feasible flyways than Straggler.

Metric: Our primary metric of goodness is the completion time of demands (CTD) as defined in §5 and shown in the example (Figure 14) of §5.1. To facilitate comparison, we report the *normalized CTD*: CTD/$\text{CTD}_{\text{ideal}}$, where $\text{CTD}_{\text{ideal}}$ is the CTD with an ideal, non-oversubscribed network. In a 1:N oversubscribed network, the baseline network has a CTD of N, and obtaining a CTD of 1 implies that with flyways, the network has performed as well as the ideal, non-oversubscribed network. We will also report statistics on the numbers of flyways used, the capacities of those flyways and their utilization.

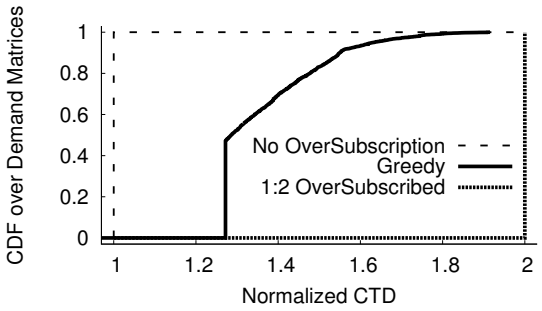


Figure 15: With just one device/ToR with NB antennas, the greedy traffic-aware choice of flyways provides significant improvements for demands observed in data centers.

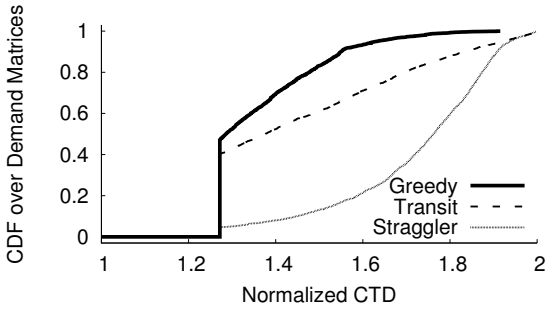


Figure 16: Improvements from the traffic-oblivious scheme as well as from each of the constituent ideas.

6.2 Benefits from flyways

Figure 15 plots a CDF of the normalized CTD over all the demands in the dataset on a 1:2 oversubscribed network. For reference, the normalized CTD of the ideal non-oversubscribed network and the baseline are 2 and 1 as shown in the figure. With just one device per ToR (with NB antennas), Greedy provides significant improvements. About 50% of the demand matrices have a normalized CTD of 1.27, i.e., 27% off the optimal. More than 90% of the demand matrices experience a speed-up of at least 45% (normalized CTD < 1.55). This configuration trades roughly half the number of switches, links and ports (by running at 1:2 oversubscription) for one wireless device per ToR.

At first blush, it is surprising that a large number of demand matrices reach CTD=1.27, but none go lower. The reason is that CTD improvement is limited by the additional capacity in or out of each ToR. Given a baseline network oversubscribed N times and K flyways per ToR of capacity F, the best possible CTD is $N/\left(1+\frac{KF}{C}\right)$, where C is the uplink capacity at each ToR. Then with the flyway capacity 85% of the ideal 6.756 Gbps wireless bitrate and a ToR uplink of 10 Gbps, it follows that for the default configuration of one device per ToR, the best possible normalized CTD value is about 1.27. Thus, half of the demand matrices obtain almost the best possible savings.

Figure 16 compares Greedy with other schemes. We see that Straggler performs quite poorly. Since high fan-in (and fan-out)

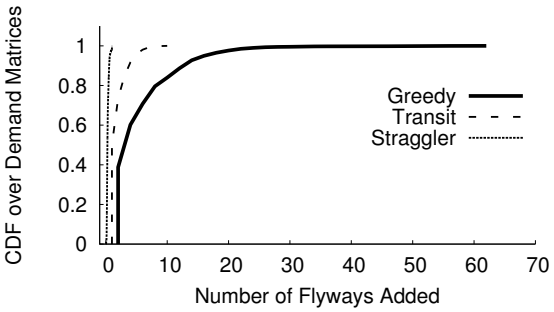


Figure 17: Average numbers of flyways used

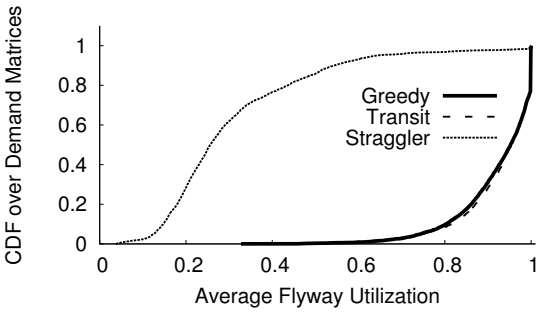


Figure 18: Average utilization of the flyways that are added

correlate with congestion, Straggler runs out of flyways that it can add. As expected, the supporting result in Figure 17 shows that Straggler adds many fewer flyways than all of the other schemes. Offloading the demands to just the largest neighbor does not impact the hotlinks by much. Instead, by allowing indirect traffic across flyways, Transit improves the performance for every demand matrix. Greedy performs even better. Building on the ability to indirect, Greedy searches among many more flyway possibilities and adds those that allow the most traffic to be offloaded. Figure 18 shows that for both Transit and Greedy, almost all the flyways are fully utilized. In addition, Greedy primarily picks short flyways that achieve the full possible rate. This indicates that were more capacity achievable on the flyway link, Greedy's performance would improve. These results reaffirm the value of allowing transit traffic across flyways and greedily picking the best over the resulting many possibilities.

6.3 Evaluating alternate configurations

To understand the solution space better, we evaluate alternatives with more wireless devices available at each ToR, different antennas and different degrees of oversubscription on the core.

More wireless devices/ToR: Figure 19 plots the benefits due to flyways when more than one wireless device is available at each ToR. We see that with just one additional device ($K = 2$), the improvements in completion time are significant. In fact, over 40% of the matrices finish as fast as they would have in a non-oversubscribed network. There are two reasons for this. First, as we saw in Figure 15, with just one device available per ToR, some of the demand matrices are constrained by the maximum capacity that a flyway adds. Additional wireless devices provide immediate benefit to these matrices. Second, even matrices that are unconstrained by flyway capacity experience benefit because with more flyways many more indirect routes are now feasible. Ever more traffic gets diverted away from congested parts of the wired network via flyways to other wired links that have spare capacity.

Different antenna configurations: All the results so far were with a narrow beam, 23 dBi gain, antenna. Figure 20 compares the benefits when different directional antennas are used. We compare Greedy with Transit, its next best alternative. We find that Greedy works best with the Phocus antennas, even though they are less directional than NB antennas, due to two reasons. First, Greedy biases the algorithm to use shorter higher capacity flyways and to then route traffic indirectly via these links. Most of these short links will continue to exist even with lower gain antennas. Second, the Phocus array has smaller back and side lobes, resulting in lower interference, and hence more simultaneously usable links. We find that unlike Greedy, Transit is sensitive to antenna directionality. With the wider beam antennas, Transit performs considerably worse and is on par with Straggler+NB. That is, the benefits from allowing transit traffic are lost with wider antennas. The reason is that with the wider antennas, there is a greater variation of capacities across flyways (as predicted by Figure 8) and quicker decay with distance. The inability to pick flyways other than those between the straggling ToR pair causes Transit to lose its gains. On the other hand, Greedy's selectivity allows it to retain most of its gains even with the wider antennas.

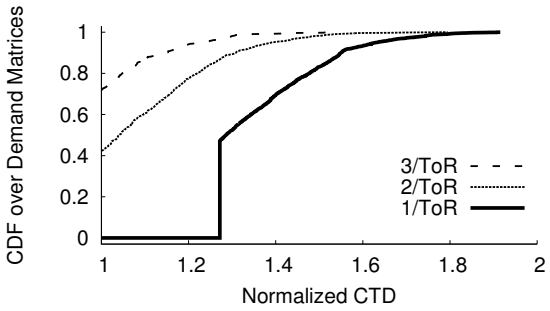


Figure 19: Increasing number of wireless devices/ToR: One more device per ToR provides significant additional benefit.

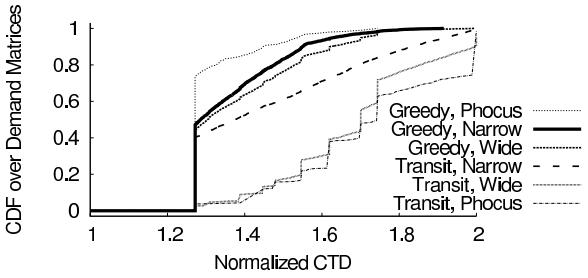


Figure 20: Different antenna configurations: The greedy approach is more robust with wide-beam antennas where flyways can have very different capacities.

Different oversubscription factors: With a greater oversubscription factor, e.g., slower links between the ToR and the core or fewer core switches in a VL2-like architecture, the network core would be relatively less expensive. Figure 21 plots the median normalized CTD across demand matrices for different oversubscription factors. We see a reasonable trade-off: one can increase the oversubscription factor on the wired network and instead spend a small fraction of that amount to deploy additional wireless devices at each ToR. The marginal improvement from each additional device decreases, but the savings are considerable. On a 1:4 oversubscribed network, flyways with 3 devices per ToR provide a median CTD of 1.78, i.e., performance better than a 1:2 oversubscribed network.

7. DISCUSSION

Flyways limitations: For some workloads, such as all-pairs-shuffle, non-oversubscribed networks are indeed more appropriate. However, these workloads are not reflected in our many traces, and we

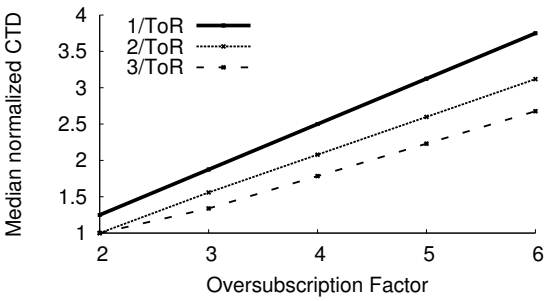


Figure 21: Flyway benefit vs. Oversubscription Ratio

believe that such workloads are rare in practice. Our current "flyway picker" algorithm requires knowledge of traffic patterns. In some cases (e.g., multi-tenant data centers) traffic patterns may not be predictable, and there may be no cluster-wide scheduler. In such cases, we believe that an online traffic engineering approach such as that described in [5], combined with the ability to rapidly steer antennas (every few seconds) may be the right solution. Design of flyway validator and flyway picker becomes more interesting when rapid beam steering is possible; since now a single flyway device can divide time across multiple neighbors. We are currently investigating the practical issues (e.g., routing) involved in this approach.

Scaling with faster wired networks: The maximum rate specified in 802.11ad is 6.76 Gbps. We have shown (§4) that flyway bandwidth needs to be only a fraction of the hot link's capacity. Still, as the speed of wired links in the data center continues to grow, we may need faster flyways. Our results in Figure 8 show that many links have ample SNR headroom and thus have plenty of room to grow with higher modulations. In addition, the flyway architecture is not specific to 60 GHz technology. Other frequencies in the 50–75 GHz band have similar properties and as 60 GHz devices become a mature technology, it may be possible to convince the FCC to open up more spectrum around the 60 GHz band for indoor data center use. Given the large lot sizes of data centers and the short propagation distance of 60 GHz links, it may be possible to use a wider band while ensuring that no detectable signal leaves the data center premises.

Scaling the data center size: Network architectures such as VL2 [8] and FatTree [1] allow the data center to scale easily in addition to providing full bisection bandwidth. In these designs it is easy to build a bigger network by simply adding additional switches, instead of investing in larger aggregation switches, or adding more layers to hierarchy. We can use the flyway architecture in conjunction with oversubscribed VL2-like networks. The VL2 architecture can be used for easy scaling, while the flyways are used to address congestion in a dynamic manner.

Containerized data center networks: While many of today's large data centers use a large, open floor plan (and new ones continue to be built), some of the new data centers are being built with containerized architecture. In a container environment, we can either deploy flyways inside a container, or between containers. Deploying flyways inside a container, instead of building a full bisection bandwidth network may allow for cheaper containers, as much less hardware will be required. On the other hand, inside a container, flyways may suffer from multipath effects, as radiation bounces off metal walls of the container. This issue can be addressed in numerous ways — by lining the inside of a container with adsorbent materials, or by employing very narrow beam antennas. For inter-container traffic, flyways are an ideal choice, since a number of devices can be mounted atop a container. At the same time, inter-container links will need higher bandwidth. We plan to study this scenario further.

Comparison with Helios and c-Through: A direct comparison between Helios [7] and c-Through [30] is difficult to perform at this point. There are several reasons for this. First, the performance of any flyway scheme will depend on the speed at which flyways can be switched between nodes. We believe (Appendix A) that wireless flyways can be switched extremely fast compared to optical MEMs switches. However, to do a meaningful comparison, we would need access to electronically steerable antennas, which we do not have. Second, the Helios system is meant for inter-container traffic, while we focus on inter-rack traffic. Third, the workload used in Helios is artificial, while our evaluation is based on real traces. Finally, our modest attempt at comparison was not successful. We built a test-bed using MEMs optical switches from the same vendor as Helios. However, we found switching times to be above 100 ms and reducing it to the values in the Helios paper (10 ms) would have required significant resources and time (software switches and NIC modifications).

Signal leakage: A concern with using wireless in a data center environment is that the signal may leak outside the data center and be picked up by an attacker. Our measurements show that the concern is unfounded. We found that common construction materials such as wood, glass and metal significantly attenuate the signal by a large margin. Coupled with normal free-space attenuation, this margin makes it very unlikely that the signal can be decoded outside the data center, even with a highly directional antenna. We omit detailed results due to lack of space.

Power consumption: Our experimental 60 GHz HXI devices consume 25 Watts of power. Several startups report devices that consume at most a few Watts [24, 27]. As a typical server rack draws thousands of Watts of power, a few additional wireless devices per rack increase consumption by a negligible fraction.

8. RELATED WORK

60 GHz wireless: Millimeter wavelength wireless communication is a very active research area, especially at the hardware/PHY level, with several dedicated conferences and workshops. Much of this work focuses on characterizing signal propagation, proposing new modulation schemes and devising antenna hardware, and much has been synthesized into WiGig and WirelessHD standards. Our work benefits from advances in this area. However, our use of 60 GHz links for data center communications is significantly different from the other applications that the field has explored.

We are aware of only one other[1] paper [22] that has discussed using 60 GHz links in data centers. In it, the authors give a high-level vision for an all-wireless data center network. In contrast, we present a hybrid architecture, backed by an experimental study of 60 GHz wireless propagation in the data center environment and evaluate its merit upon several types of traffic measured in the data center.

Data center networks: A number of papers have addressed the problem of congestion in data center networks. We discuss a representative sample. Researchers have proposed [1, 8, 9] building full bisection bandwidth networks to eliminate hot-spots. Deploying such networks is expensive, and the cabling complexity of some of them is quite daunting [15].

Hedera [2] and MicroTE [5] advocate fine-grained traffic engineering over a fixed topology to alleviate congestion. However, this approach has limitations. For example, in a tree-structured net-

[1]Apart from the preliminary version [15] of this paper.

work, if the downlink from an aggregator switch to a ToR switch is congested, only extra bandwidth can relieve the congestion. At the other extreme, Proteus [28] explored the idea of a completely reconfigurable, all-optical network topology.

Instead, we explore the idea of adding additional bandwidth to data center networks on demand. The closest work to our own is Helios [7] and c-Through [30], both of which propose flyways using optical switches. Wireless flyways have the benefits of potentially lower costs and limited cabling complexity but also face challenges unique to wireless, such as interference, and we show how we address these.

There are other differences in the designs as well. C-Through delays TCP connections to ensure that optical links are used optimally; we do not require any such delay. Helios is targeted towards inter-container traffic; we target sub-agg-switch traffic. Most importantly, both Helios and c-Through use synthetic workloads, while we use extensive traffic traces from a variety of real data centers to motivate our system and assess the value of flyways.

9. CONCLUSION

We have presented the design and evaluation of a 60 GHz wireless flyway system. It adds wireless links to the wired data center network to relieve hotspots and thus improve performance. Working with prototype 60 GHz devices, we measured and simulated performance to show that wireless flyways can provide a dense deployment of stable, multi-Gbps paths in the DC environment. By analyzing traces of DC traffic for four real applications, we find that only a relatively small number of top-of-rack switches and links are congested at any time. This implies that a set of flyways with relatively small capacity can relieve hotspots and boost application performance. Informed by our exploration of 60 GHz and DC traffic, we designed a wireless flyway system that sets up the most beneficial flyways and routes over them both directly and indirectly to reduce congestion on hot links. Our trace-driven simulation shows this design speeds up network-limited DC applications with predictable traffic workloads by 45% in 95% of the cases at a fraction of the cost of avoiding oversubscription.

10. REFERENCES

[1] M. Al-Fares et al. A scalable, commodity data center network architecture. In *SIGCOMM*, 2008.
[2] M. Al-Fares et al. Hedera: Dynamic flow scheduling for data center networks. In *NSDI*, 2009.
[3] S. Alalusi. CMOS multi-antenna systems at 60 GHz. In *Communications Design Conference*, July 2004.
[4] M. Alizadeh et al. DCTCP: Efficient packet transport for the commoditized data center. In *SIGCOMM*, 2010.
[5] T. Benson et al. The case for fine-grained traffic engineering in data-centers. In *WREN*, 2010.
[6] T. Benson et al. Network traffic characteristics of data centers in the wild. In *IMC*, 2010.
[7] N. Farrington et al. Helios: A hybrid electrical/optical switch architecture for modular data centers. In *SIGCOMM*, 2010.
[8] A. Greenberg et al. VL2: A scalable and flexible data center network. In *SIGCOMM*, 2009.
[9] C. Guo et al. BCube: High performance, server-centric network architecture for data centers. In *SIGCOMM*, 2009.
[10] IEEE P802.11ad/D0.1: Enhancements for very high throughput in the 60 GHz band. Draft 0.1, June 2010.
[11] M. Isard et al. Quincy: fair scheduling for distributed ocmputing clusters. In *SOSP*, 2009.
[12] K. Jain et al. Impact of interference on multi-hop wireless network performance. In *MOBICOMM*, 2003.
[13] S. Kandula et al. Walking the tightrope: responsive yet stable traffic engineering. In *SIGCOMM*, 2005.
[14] S. Kandula et al. The nature of datacenter traffic: Measurements and analysis. In *IMC*, 2009.
[15] S. Kandula, J. Padhye, and P. Bahl. Flyways to de-congest data center networks. In *HotNets*, Nov. 2009.
[16] LightPointe. http://www.lightpointe.com/.
[17] X. Liu et al. DIRC: Increasing indoor wireless capacity using directional antennas. In *SIGCOMM*, 2009.
[18] T. Manabe et al. Effects of antenna directivity on indoor multipath propagation characteristics at 60 GHz. In *PIMRC*, 1995.
[19] V. Navda et al. MobiSteer: Using steerable beam directional antenna for vehicular network access. In *MobiSys*, 2007.
[20] G. Nikolaidis et al. (Poster) Cone of silence: Adaptively nulling interferers in wireless networks. In *SIGCOMM*, 2010.
[21] The ns-3 network simulator. http://www.nsnam.org/.
[22] K. Ramachandran et al. 60 GHz data-center networking: Wireless ⇒ worry less? Technical report, NEC, 2008.
[23] T. S. Rappaport. *Wireless Communications: Principles and Practice*. Prentice-Hall, 2002.
[24] Sayana Networks.
[25] A. Sheth et al. Geo-fencing: Confining Wi-Fi coverage to physical boundaries. In *Pervasive*, 2009.
[26] SiBeam. http://sibeam.com/whitepapers/.
[27] SiBeam. http://www.earthtimes.org/articles/press/generation-solutions-ces-2011,1604984.html.
[28] A. Singla et al. Proteus: A topology malleable data center network. In *HotNets*, 2010.
[29] Terabeam Wireless. http://www.terabeam.com/.
[30] G. Wang et al. c-Through: Part-time optics in data centers. In *SIGCOMM*, 2010.
[31] Wireless Gigabit Alliance. http://wirelessgigabitalliance.org/.
[32] WirelessHD. http://wirelesshd.org/.
[33] Y. Yu et al. A 60 GHz digitally controlled phase shifter in CMOS. In *ESSCIRC*, 2008.

APPENDIX

A. PHASED ARRAYS

Our assumptions about phased array design are based on three sources of information: (1) research literature on commercially available 2.4 GHz Phocus arrays from the wireless research communities, (2) existing 60 GHz silicon such as SiBeam's WirelessHD [26] products, and (3) research on 60 GHz phased arrays technology.

Phased array technology: A phased array comprises multiple antenna elements each transmitting (or receiving) an attenuated, phase-shifted copy of the same RF signal. By varying the amount of attenuation or phase shift, a device can control the radiation pattern of the antenna, including the direction of maximum gain or the size and location of lobes. The flexibility of a phased array increases rapidly with the number of antenna elements. The 2.4 GHz Phocus array is commercially available today and uses 8 elements; in contrast, 60 GHz phased array design is an area of ongoing research [3, 33], but SiBeam's initial WirelessHD products include 32 elements and fit into 1 in^2.

Antenna pattern: In this work, we simply replicate the radiation pattern used in the Geo-fencing project [25]. This pattern minimizes the size and extent of back and side lobes and can be produced by the 8-element Phocus arrays. Though 60 GHz and 2.4 GHz phased array technologies will likely differ, we expect that the increased number of elements at 60 GHz will enable patterns of similar or better flexibility. For simplicity, in this paper we assume that the antenna pattern can be steered to an arbitrary angle. Extrapolating from the Phocus array, a 30-element array might in practice have a 6° granularity.

Switching times: The phased array technology used in the Phocus arrays can be switched within 250 μs [19]. In personal communication, the authors of [33], informed us that their 60 GHz phase shifting technology can be switched in picoseconds. However since

we cannot yet implement and evaluate phase-switched flyways, we ignore switching overhead in this paper and instead focus on individual traffic matrices in isolation.

Discovering the steering coefficients: Steering coefficients allow directional antennas of the sender and transmitter to point at each other. Optimizing the search process to discover steer coefficients is currently an active area of research. However, the flyways scenario greatly simplifies this problem. We use the wired network to coordinate the steering. Rather than needing to optimize the combination of multiple reflections off walls and in-home objects, in the DC, we can use physically meaningful steering patterns that point in a particular direction. A movement of a few mm does not dramatically reduce gains. The stable DC environment allows us to use history and retrain infrequently. Any nodes not involved in flyways (there will be many of these) can opportunistically measure their directional gains with respect to ongoing transmissions, as well as update measurements for patterns between idle nodes that will not interfere with ongoing traffic.

B. 60 GHz SIMULATOR

We implemented an 802.11ad simulator in `ns-3`. To have confidence that our simulations are a good reflection of reality, we base wireless effects directly on the physical layer measurements we took in §3 and the WiGig/802.11ad PHY and MAC design [31]. Here we describe wireless aspects of our `ns-3` model. We extended `ns-3` with other support too, such as automatic generation of DC layouts and routing, but these components are straightforward and we omit them due to lack of space.

Directional antennas: We built table-driven models from the measured radiation patterns of the antennas in our lab (Figure 3). We interpolate between measurements when needed. As well as using measured patterns, rather than the manufacturer antenna specifications, we take care to simulate the full 360° radiation pattern, not just the primary lobe. We also added a simple isotropic antenna model, and the radiation pattern used for Geo-fencing [25].

IEEE 802.11ad PHY and MAC: We implemented in `ns-3` the physical and MAC layers defined in the draft 802.11ad standard. We limit ourselves to the faster OFDM PHY. We fix transmit power to 10 mW to match commercial devices.

Signal propagation: We model signal propagation using Friis' law. Our measurements (§3.2) show that this is a good fit for line-of-sight environments. Still, we conservatively subtract an additional 3 dB from the signal power (but not from interference) to represent potential destructive multi-path interference received via side lobes.

Interference (SINR): To calculate the SINR needed for bit error rate estimation, `ns-3` uses the standard SINR modeling technique. It adds together the power from multiple interferers, combines it with noise, and compares it with signal strength. `ns-3` does not model symbol-level fading, i.e., it assumes that the received power (RSS) from each transmitter is consistent throughout its transmission. It does, however, compute different SINR levels for different parts of packets when interference stops or starts during reception.

Our measurements of the stability of real links (§3.3) show that we can use this SINR model and ignore fading at the sub-packet level. Prior work (DIRC [17]) has also found this simple SINR model to be appropriate with directional antennas, even when using the 802.11g OFDM rates in non-line-of-sight environments and omni-directional antennas at receivers. The model is much more fitting in our 60 GHz domain: both transmitter and receiver use directional antennas so that secondary rays (multi-path) have little impact (§3.2); and the channel is very stable due to little environmental mobility (§3.3).

Ideal Rate	Wireless TCP	Offload ACKs to Wired	No DCF
693 Mbps	656 Mbps	672 Mbps	676 Mbps
6.76 Gbps	4.58 Gbps	5.36 Gbps	5.62 Gbps

Table 2: Impact of sending TCP ACKs over wire

Bit error rate (BER): Estimating the bit error rate, and hence whether a transmitted packet is received correctly, forms a key function of any wireless model. The input to this calculation is the SINR and the 802.11ad wireless rate. To estimate BER, we use the 802.11ad standard as our guide. It defines the sensitivity for each rate and coding as the (SINR) power level down to which a device much successfully receive more than 99% of 4096-byte packets sent using that rate. This reception rate corresponds to a BER less than 3.07×10^{-7}, and thus we calibrate our error model for each rate by assuming its BER is 3×10^{-7} when its SINR is the sensitivity threshold. The sensitivities defined in the standard implicitly include the (≈-81 dBm) thermal noise for a 2.16 GHz channel, and a 15 dB combined implementation loss. We compute BERs at other SINR values using textbook formulas [23] for BER as a function of SNR in Gaussian noise. To receive a packet, all bits must be correct.

Auto-rate algorithm: The 802.11ad standard does not mandate use of a specific auto-rate algorithm. We select rates based on received SINR. This is reasonable for our stable data center environment.

C. IMPROVING WIRELESS PERF.

Data center performance will improve the most when the flyways deliver the largest possible throughputs. A unique aspect of our flyways scenario is the hybrid wired and wireless nature of the network, and in this section we describe and evaluate two wireless optimizations that leverage the wired backbone in the DC to increase flyway TCP throughput by 25%

Wired offload of MAC-inefficient packets: TCP ACKs are far smaller than data packets, and make inefficient use of wireless links because payload transmission time is dwarfed by overheads such as preamble and SIFS. The hybrid wired-wireless design of our network lets us improve efficiency by sending ACK packets over the wire instead. To measure the improvement, we simulated a single TCP flow on a 20 m link and configured `ns-3` to send the TCP ACKs over the wired network. Table 2 shows the resulting TCP throughput. For fast links enabled by the narrow-beam antenna, the performance improves 17%. Note that the TCP ACK traffic will use some wired bandwidth, but this will be trivial compared to the increase in throughput.

Removing DCF: For the common case of one-way TCP flows in the data center [4], if we divert TCP ACKs over the wire as above then all traffic over a given wireless link will flow in only one direction. Furthermore, our system design (§5) is based on independent flyways that do not interfere with one another. Thus, there are no collisions in our wireless network, and we can eliminate the DCF backoff mechanism. This change improves the TCP throughput by an additional 5%, as seen in the third column of Table 2.

Occasionally there may be bidirectional data flows over the flyway. Even in this case, we can remove the cost of DCF. Since only the two communicating endpoints can interfere with each other, we can easily schedule transmissions on the link by passing a token between the endpoints. This naturally fits into the 802.11 link layer protocol because after transmitting a packet batch, the sender waits for a link layer Block-ACK. We can exploit this scheduled hand-off to let the receiver take the token and send its own batch of traffic.