Hekaton: Efficient and Practical Large-Scale MIMO

Xiufeng Xie
University of Wisconsin-Madison
xiufeng@ece.wisc.edu

Eugene Chai
NEC Labs America
eugene@nec-labs.com

Xinyu Zhang
University of Wisconsin-Madison
xyzhang@ece.wisc.edu

Karthikeyan Sundaresan
NEC Labs America
karthiks@nec-labs.com

Amir Khojastepour
NEC Labs America
amir@nec-labs.com

Sampath Rangarajan
NEC Labs America
sampath@nec-labs.com

ABSTRACT

Large-scale multiuser MIMO (MU-MIMO) systems have the potential for multi-fold scaling of network capacity. The research community has recognized this theoretical potential and developed architectures [1,2] with large numbers of RF chains. Unfortunately, building the hardware with a large number of RF chains is challenging in practice. CSI data transport and computational overhead of MU-MIMO beamforming can also become prohibitive under large network scale. Furthermore, it is difficult to physically append extra RF chains on existing communication equipments to support such large-scale MU-MIMO architectures.

In this paper, we present Hekaton, a novel large-scale MU-MIMO framework that combines legacy MU-MIMO beamforming with phased-array antennas. The core of Hekaton is a two-level beamforming architecture. First, the phased-array antennas steer spatial beams toward each downlink user to reduce channel correlation and suppress the cross-talk interference in the RF domain (for beamforming gain), then we adopt legacy digital beamforming to eliminate the interference between downlink data streams (for spatial multiplexing gain). In this way, Hekaton realizes a good fraction of potential large-scale MU-MIMO gains even under the limited RF chain number on existing communication equipments.

We evaluate the performance of Hekaton through over-the-air testbed built over the WARPv3 platform and trace-driven emulation. In the evaluations, Hekaton can improve single-cell throughput by up to 2.5× over conventional MU-MIMO with a single antenna per RF chain, while using the same transmit power.

1. INTRODUCTION

Multi-User MIMO (MU-MIMO) systems allow a single transmitter with multiple antennas to serve multiple downlink users (*e.g.*, cellphones) concurrently using the same spectrum resource. Every single antenna is driven by one entire *RF chain* that consists of many PHY hardware components such as the baseband processor, the de/modulator, power amplifier and the ADC/DAC.

Theoretically, we have the potential to scale up the network throughput multiple-folds by increasing the number of RF chains. Several large-scale platforms, such as Argos [1] and BigStation [2], have recognized this potential and successfully demonstrated research platforms with a large number of antennas. These potential gains are critical for 5G cellular network and future WiFi network to meet the sharply increasing throughput demands. Unfortunately, we will encounter difficulties in achieving such gains in practice.

***(i)* Energy-Efficiency (bits-per-Joule).** To scale up the MU-MIMO downlink capacity, we add more RF chains (and thus more antennas) to the base station. However, the rank of the MU-MIMO channel in a particular environment is limited [3], *i.e.*, there is an upperbound for the multiplexing gain, which only depends on the environment, regardless of what kind of equipment is used. When we reach this limit, further increase in the number of antennas (and RF chains) only improves the beamforming gain. The rate of capacity scaling will thus fall behind the increasing rate of energy consumption, and the *energy-efficiency* will decrease. This is evident in MU-MIMO systems with large antenna number N, where serving N users can steeply degrade the performance due to the tightness of the degrees of freedom at the base station. In this case, the maximum downlink capacity is achieved when serving less than N users. This effect has been reported in existing MU-MIMO research works [1, 2, 4], where number of costly RF chains needs to be $1.5\times \sim 2\times$ the downlink user number to guarantee the channel orthogonality between these users.

***(ii)* High Overhead.** The *coordination overhead* of a MU-MIMO system increases linearly with the number of antennas. Before each MU-MIMO downlink beamforming, the base station must send a probe from each of its antennas, in turn, to obtain the channel state information (CSI) between each of its antennas and each user antenna involved in the transmission. Note that current LTE networks are typically Frequency Division Duplexing (FDD) networks. This means that different frequency bands are used for uplink and downlink transmissions. Hence, the implicit feedback [5] mechanisms that are used to reduce CSI feedback overhead cannot be employed here. As we increase the number of RF-chains/antennas on the base station, the overhead due to these measurements can easily overwhelm the capacity gains from having more antennas. Furthermore, the computational overhead of the commonly used Zero-Forcing Beamforming (ZFBF) mechanism can become prohibitive in existing large-scale MU-MIMO systems. Novel beamforming mechanism [1] or hardware design [2] are proposed to address this issue, which, however, deteriorate the backward compatibility.

***(iii)* No Standards Support.** Existing LTE and WiFi standards only support up to 8 MU-MIMO UEs/clients. This is a far cry from the hundreds of users envisioned for large-scale MU-MIMO networks [1]. Support for large-scale MU-MIMO systems is planned for 5G networks, but the 5G standard is not expected to be widely available before the year 2020. Given the complexities involved in upgrading LTE BSs or WiFi APs, it will be difficult to increase the number of RF-chains and antennas on the base station in a non-

MobiCom'15, September 7–11, 2015, Paris, France.
 ISBN 978-1-4503-3543-0/15/09 ...$15.00.
DOI: http://dx.doi.org/10.1145/2789168.2790116.

standards-compliant manner. In spite of these challenges, it is well-accepted that large-scale MU-MIMO systems offer the best capacity scaling potential for future wireless networks. The key question then is: *can we achieve a good fraction of these gains in existing systems in a standards-compliant manner?*

In this paper, we present Hekaton, a large-scale MU-MIMO framework for indoor environments (*e.g.*, LTE femtocell and WiFi). Hekaton stems from the key insight that performance gains in MU-MIMO networks consist of two components: multiplexing gains and beamforming gains. Multiplexing gains depend on the number of transmitted streams, each of which requires at least one separate RF chain[1]. On the other hand, beamforming gains depend on the number of antennas used to transmit each steam. In conventional MU-MIMO, every single antenna is driven by one entire RF chain, which enforces a tradeoff between multiplexing gain and beamformin gain under the base station's given number of RF chains. Hekaton eliminates this tradeoff by equipping every RF chain on the base station with a phased-array antenna that consists of multiple antenna elements each. As a result, the number of base station antennas is no longer bounded by the number of RF-chains. Hekaton can thus increase the beamforming gains independently of the number of RF chains.

Hekaton addresses the MU-MIMO scalability challenges in the following ways.

***(i)* Scaling up the beamforming gain.** Hekaton increases only the beamforming gains through the addition of phased-array antennas. Note that even though phased-arrays do require an energy source, the power consumed[2] is typically very small [6]. Hence, given the same energy-budget of the base station, Hekaton increases the capacity and thus, the energy efficiency over MU-MIMO under the same number of RF chains.

Furthermore, the multiplexing gain is practically limited by the *rank* of the channel [3]. Any additional RF chains beyond the channel rank can essentially be only used towards improving beamforming gains. Hekaton enables us to increase beamforming gains without unnecessary increase in the number of RF chains beyond the maximum channel rank.

***(ii)* Low coordination overhead.** A by-product of the improved beamforming gain is efficient inter-cell coordination. Hekaton needs only coarse-timescale analog beamforming to steer signal energy away from adjacent cells and maximize the Signal-to-Leakage Ratio (SLR). If Hekaton is not used, inter-cell coordination must occur over the fine-timescales that are used by conventional MU-MIMO-only networks.

Beamforming gains can be obtained using coarse-timescale control while multiplexing gains must be obtained using fine-timescale, frame-by-frame coordination. Inspired by this fact, we advocate a two-level two-time-scale beamforming architecture for Hekaton, which consists of a coarse-grained *analog beamforming* component that is performed by the phased-array antenna, and a *fine-grained* digital MU-MIMO precoding component, implemented by the existing RF-chains.

The per-frame per-antenna CSI measurement overhead is largely due to the multiplexing gains of the system. Hekaton maintains this overhead as it only increases the beamforming gains. The phased-array antenna only needs to be updated on a coarse timescale (*e.g.*, every 200ms in our implementation) and the effort required for each update is small.

[1]The RF chain consists of many PHY hardware components such as the baseband processor, the de/modulator, power amplifier and the ADC/DAC.

[2]The power is used to operate the codebook and program the phase shifters. It does not depend on the transmit power.

***(iii)* Retrofitted solution.** Hekaton extends existing architectures with phased-array antennas along with associated updates to the schedulers (in LTE) and software drivers (in WiFi). No changes to the LTE and WiFi standards, or changes to the wireless transceiver hardware is necessary. Hence, Hekaton is backward compatible with existing wireless infrastructure. For example, the deployment of Hekaton does not require approval from the 3GPP.

However, using phased-arrays in practice raises an important challenge: how does one steer beam of the analog phased-array efficiently? Unlike the case with traditional non-Hekaton MU-MIMO arrays, Hekaton cannot obtain channel information from each antenna element of the phased-array and therefore cannot directly compute the optimal downlink beam pattern from the CSI. Previous uses of phased-array antennas [7] exhaustively probe all possible beam directions and select the one that results in the highest SNR. Unfortunately, the overhead of such an approach scales *exponentially* with the number of phased-array antennas and the size of each antenna array.

Hekaton employs a novel compressive-sensing algorithm to steer the analog phased-array with only a *constant* overhead, regardless of the number of phased-array antennas or the size of each one. The downlink beam is steered towards the angle-of-arrival (AoA) of the strongest signal path from the target user. There are infinitely many possible AoAs over which to search for this direction. Fortunately, in real-world channels, the AoAs that result in strong signal energy at the user are clustered into 3-5 groups, and are thus *sparse*. We exploit this sparsity to select beam directions with a fraction of the overhead of other typical approaches [8]. We note that compressive sensing (CS) has been used previously for AoA measurements [9]. However, such approaches typically require one RF chain per antenna. Our novelty lies in our application of compressive sensing to phased-array antennas that are simply attached to existing RF chains. *This is the key algorithm that now enables phased-array antennas to be used together with existing basestations efficiently and effectively.*

We implemented and evaluated Hekaton on a WARP testbed. Hekaton , with two RF chains, achieves a throughput gain of $2.5\times$ over non-Hekaton MU-MIMO.

We will motivate our design in §2 and describe the key elements of Hekaton in §3. We then cover the design details of Hekaton in §4, §5 and §6. We discuss experimental evaluation in §7 and §8. Additional discussions are in §9 and related work are reviewed in §10. §11 contains the conclusion.

2. BACKGROUND AND MOTIVATION

2.1 MU-MIMO Background

Multi-User MIMO (MU-MIMO) enables one transmitter with multiple antennas to send concurrent data streams to *multiple* users with one or more antennas on each. As a comparison, in *single-user* MIMO (SU-MIMO), all receiving antennas rest on a single device. Assume the MU-MIMO transmitter has N antennas and each MU-MIMO receiver has one antenna, then the *total* number of MU-MIMO receivers M cannot exceed the number of antennas at the transmitter, i.e. $N \geq M$. The multiplexing gain depends on the total number of parallel streams of transmitted data, while the beamforming gain drives the capacity of each stream. In a typical MU-MIMO network, each antenna (on either the transmitter or the receiver) is driven by one RF chain.

2.2 MU-MIMO Capacity Scaling

In MU-MIMO, the increase in multiplexing gain without ensuring channel orthogonality between users may drastically affect the

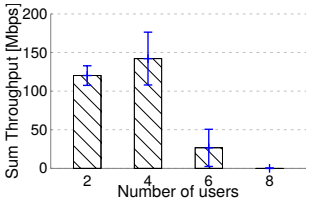


Figure 1: MU-MIMO DL throughput with increasing number of users (fixed 8 transmit antennas, 8 RF chains).

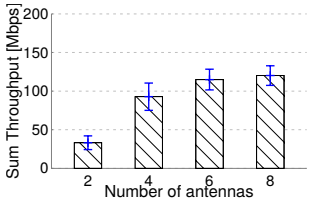


Figure 2: MU-MIMO DL throughput with increasing number of transmit antennas (fixed 2 users, 2 RF chains).

downlink performance [3]. As an example, Figure 1 shows the total throughput obtained using a MU-MIMO transmitter with eight antennas and a set of single antenna receivers. To obtain these results, we place WARP radios throughout a typical office environment, collect the MU-MIMO SINR, then map the SINR to the downlink throughput using the lookup table from [10]. Details of the MU-MIMO precoder used can be found in Section 7. It is observed that the total throughput increases as we increase the number of receivers from two to four. However, as we further increase the number of of receivers from four to eight, the total throughput actually decreases. Similar observations to Figure 1 have been reported in [2, 4]. Alternatively, we can increase wireless capacity by only increasing the beamforming gains. Figure 2 shows the total throughput with two single-antenna MU-MIMO receivers. As we increase the number of antennas at the transmitter from two to eight, the total network throughput strictly increases. Note that the total transmit power is kept constant as the number of antennas increases. Also, similar observations to Figure 2 have been reported in existing works like [1, 2].

2.3 The Benefits of Beamforming Gain

Our two-level hybrid beamforming architecture introduced in §1 scales up the beamforming gain and the downlink capacity under a given number of RF chains by equipping a phased-array antenna, instead of an omnidirectional antenna, to each RF chain.

Phased-array antennas ensure that more transmit power is projected towards desired downlink users. Hence, phased-arrays increase the channel orthogonality between users. This in turn reduces the condition number of the MU-MIMO channel matrix and improves the overall downlink capacity.

More specifically, users that are spatially correlated when using omnidirectional transmit antennas may no longer be correlated after the channel preconditioning of the phased-array antennas. Thus the system can even support more "usable" downlink streams (when maximizing the downlink capacity) without adding more RF chains. For example, in conventional MU-MIMO with 8 RF chains, serving only 4 users maximizes the downlink capacity as shown in Figure 1. However, if we equip each RF chain with a phased-array antenna, then with the same 8 RF chains, serving 8 downlink users can become the best choice as shown later in our evaluations (Figure 16a).

The two-level beamforming approach brings a multitude of benefits to large-scale networks: *(i) Energy-efficiency:* The additional phased-array antennas improve beamforming gain, with only an insignificant additional power outlay. As a result, with the same number of RF chains and hence the same energy consumed, the downlink capacity can be increased significantly (up to 2.5× in our empirical measurements); *(ii) Low coordination overhead:* Beamforming gains are achieved using the phased-array. Analog beamforming can only steer directional, coarse beams towards the UEs. Hence, it only needs the coarse-timescale, long-term CSI statistics

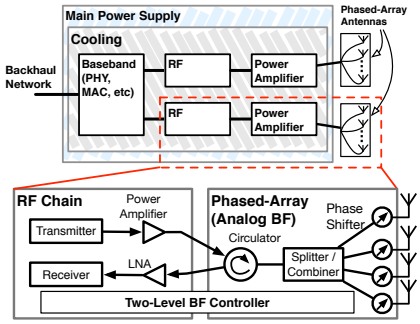


Figure 3: Hekaton basestation architecture.

of each UE. Multiplexing gains, on the other hand, must be obtained from the fine-timescale, frame-by-frame CSI information. Hence, by increasing only the beamforming gains, we only require additional coarse-grained control feedback, which incurs very minimal additional coordination overhead; *(iii) Retrofittable:* Phased-array antennas can be easily retrofitted onto existing MU-MIMO BSs, simply by replacing the existing omni-directional/directional antennas.

A Word on Multiplexing Gains. Our focus on beamforming gains does not trivialize the importance of multiplexing gains. Rather, we emphasize the fact that beamforming gains can be achieved *independently* of multiplexing gains, and without any additional energy and with a negligible coordination overhead. Hence, even for large-scale MU-MIMO systems [1, 2] that support a much larger number of RF chains than current LTE BSs, Hekaton can still be retrofitted to those systems to obtain even greater capacity via beamforming gain increases.

3. HEKATON OVERVIEW

3.1 Architecture

The radio of a node in typical wireless networks consists of a baseband component that contains the PHY, MAC and other upper layer protocols, and an RF component that modulates baseband I/Q signals into passband signals. A power amplifier increases the transmit power of these signals before they are transmitted over the channel via the antenna. All these components share the same cooling and power supply [11]. Figure 3 shows an example of a wireless transceiver with two RF chains, where Hekaton replaces the conventional antenna on each RF chain with an analog phased-array antenna.

Analog phased-array antenna. An analog phased-array antenna consists of multiple antenna elements arranged in a fixed, pre-defined pattern. Each antenna element is connected to a single analog phase shifter. Analog passband signals received from the RF chain is split equally amongst these antenna elements.

A set of unique phase shifting values for each antenna element defines a unique signal beam direction. This is known as a *codebook entry*. The set of all codebook entries supported by the phased array are arranged into a *codebook*. The codebook is typically preloaded into the phased array. During normal operations, the phased array switches between codebook entries to change the direction where the transmitted signal energy is focused.

Digital MU-MIMO Precoding. Hekaton is designed to be backward compatible with existing MU-MIMO systems like 802.11ac and LTE. Therefore its digital MU-MIMO precoding component reuses that on existing MU-MIMO systems. It does not require

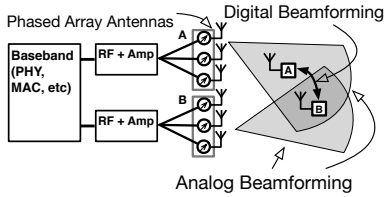


Figure 4: Hekaton operation with two RF chains.

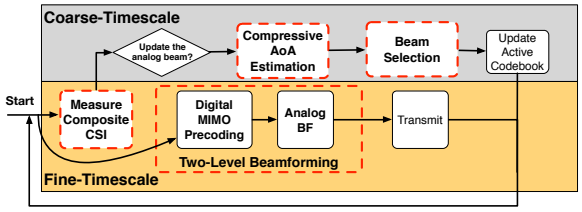


Figure 5: Hekaton two-timescale transmission.

new digital beamforming methods like in [1] to facilitate real-time large scale MU-MIMO.

Two-level beamforming. The core of Hekaton is a two-level beamforming architecture which consists of two main components: a coarse-grained *analog* beamforming enabled by the phased-array antennas, and a fine-grained *digital* MU-MIMO precoding performed by the RF chains.

Figure 4 illustrates the interaction between these two beamforming levels. Hekaton uses analog beamforming to steer signal energy in a coarse-grained fashion. Each phased-array antenna is always assigned to a unique user and it only steers the beam towards its assigned user. Digital MIMO precoding is then used to further reduce the cross-talk between users.

For example, in Figure 4, phased-array antennas A and B serve clients A and B with separate beams. The one-to-one mapping[3] between the phased-array antenna and the user is constructed to maximize the SLR, which is detailed in Section 6.1. However, since the coarse analog beamforming cannot achieve pinpoint focusing of transmitted signals, there will be residual interference between the different user. The digital MU-MIMO precoder is then used to cancel this residual interference between the downlink users.

3.2 Hekaton Operation

The two-level beamforming in Hekaton operates on two different timescales. Figure 5 shows how this two-level scheme works in practice. Each frame transmission follows a fast-timescale process, where the transmitted signal undergoes the two-level beamforming before being transmitted. The beam selection/update of the phased-array antenna occur over a slower, coarse-timescale process. The three-key components of Hekaton are:

1. Composite CSI Measurement. The Hekaton base station cannot access signal information from each of its phased-array antenna elements. Instead, given a codebook entry, the phased-array antenna combines the correspondingly phase-shifted signals from all antenna elements, and returns only this combined CSI signal to the RF chain.

2. Compressive AoA Estimation. The Hekaton base station selects the downlink beam direction (*i.e.*, the codebook entry) corresponding to the angle-of-arrival (AoA) direction with the strongest signals when it hears from the user. It is worth noting that although the channel reciprocity may not hold for an FDD system, the AoA direction of the user should still be similar across different frequencies. In fact there are many AoA-based indoor localization works using different frequency bands [8,12]. Hekaton is compatible with FDD systems as it uses codebook-based analog beamforming (8 beam patterns in our implementation), which creates a wide beam pattern and hence is robust to frequency diversity.

Unfortunately, conventional AoA estimation requires CSI measurement for each antenna element [7–9,12], which is infeasible for a Hekaton base station since it can only obtain composite CSI of all elements in a phased-array. Furthermore, an exhaustive search would require test transmissions over all possible AoAs and is too time-consuming to be practical. In Hekaton, we design a novel algorithm that can efficiently compute the AoA from just four composite CSI measurements. To the best of our knowledge, ours is the first algorithm that achieves a constant AoA measurement overhead, regardless of the number of phased-array antennas.

3. Downlink Beam Selection. Hekaton uses the AoA measurements to determine beam directions that both maximize the signal power at its intended user while minimizing the interference to users in adjacent cells. These beam directions are selected using signal-to-leakage ratio (SLR) as a metric.

We detail these components in §4, §5 and §6. We note that Hekaton can be implemented, via a two-level beamforming controller, as a minor firmware update in the downlink transmitter like a WiFi AP or an LTE eNodeB, without further hardware modifications.

4. COMPOSITE CSI MEASUREMENT

4.1 LTE Support for Composite CSI Measurement

Hekaton can be retrofitted onto existing LTE BSs. It uses three key features in current LTE BSes:

Coordinated MultiPoint (CoMP) synchronization. Each Hekaton-enabled LTE BS must measure the CSI of both its local UEs and UEs in neighbouring cells. Hence, because LTE operates using a TDMA schedule, uplink frames from neighbouring cells must be time-synchronized with those of the local cell. The LTE CoMP support ensures that this synchronization is achieved. Note that Hekaton only relies on CoMP features for inter-cell UE scheduling, and not cooperative transmissions. Hence, Hekaton does not require fast CSI/data exchange between BSes in different cells.

Coordinated scheduling. The LTE specification partitions the spectrum into multiple resource blocks (RBs), with each RB occupying a set of subcarriers for a period of time. Each UE is assigned one or more RBs for uplink transmission. With CoMP, the schedulers of different BSes coordinate to ensure that interfering uplink transmissions from different UEs are assigned non-overlapping RBs. This ensures that Hekaton can obtain a good CSI measurement from each UE.

Sounding Reference Signals (SRS). Hekaton uses the SRS that is transmitted at the end of a UE's frame for CSI measurement. CoMP support allows the BS to limit the bandwidth of the SRS to the RBs assigned by the UE. Hence, Hekaton can obtain CSI measurements from multiple non-interfering UEs concurrently.

4.2 Measuring the Composite CSI

The composite CSI is concurrently measured at all phased-arrays using uplink transmissions. Each transmission from the UE is received by all phased-arrays at the BS, and is used by each phased-array to measure the composite CSI. Figure 6 illustrates an example where Hekaton obtains upstream composite CSI from three separate UEs, A, B and C.

[3]This mapping may be suboptimal, a detailed discussion is in Section 9

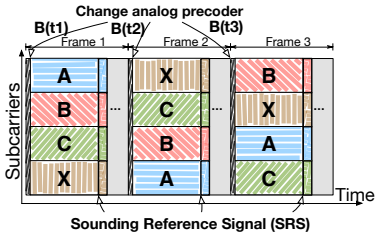


Figure 6: Obtaining the upstream CSI measurements from three UEs, A, B and C. Client X belongs to a neighbouring cell.

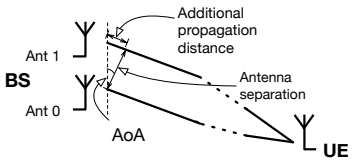


Figure 7: Arrival of a signal from the UE at two antenna elements (Ant1 and Ant2) at the BS.

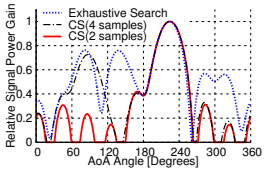


Figure 8: Azimuth signal power distribution estimated with compressive sensing algorithm.

At the first measurement frame, Hekaton randomly selects a codebook entry (corresponding to one beam direction) for each phased-array. The unmodified RF chain receives the combined signal from the phased-arrays, and computes the CSI for UEs A, B and C as per normal. Hekaton retrieves these composite CSI values from the RF chains. This process is repeated two more times, with a different random codebook entry selected for each measurement. At the end of the CSI collection step, Hekaton obtains three distinct composite CSI measurements for UEs A, B and C. Two points are worth noting here: (i) The CSI collection spans three frames. But meanwhile the data frame transmission can continue using the previous CSI since the channel coherence time usually spans multiple frames. (ii) The overhead of beam switching is very low. Modern phased-array antennas [13] can switch between codebook entries in as little as $1.2\mu s$, which is an order of magnitude smaller than the LTE subframe duration of $1ms$.

The above process is also used to measure the CSI to UEs in adjacent cells. With inter-cell time synchronization, Hekaton can also decode the SRS from UEs in adjacent cells, and obtain CSI measurements to them.

5. COMPRESSIVE AOA ESTIMATION

Hekaton employs a novel compressive-sensing based algorithm to estimate the AoA of the uplink probing signal from only a small number of composite CSI measurements. Existing AoA estimation approaches, such as those in [7, 8, 12], incur too large of a probing overhead and cannot be used efficiently in a two-level architecture. Even previous compressive-sensing approaches, such as [9], require per-antenna-element CSI and cannot operate with only the composite CSI.

5.1 What is the Angle-of-Arrival?

Because of multipath propagation, a single transmission from an antenna, even a phased-array one, will travel along multiple paths before arriving at the receiver. Each copy of the signal from the UE, travelling along a different path, arrives at the BS at a particular *angle-of-arrival (AoA)*, and will experience a different amount of attenuation and distortion. We refer to the magnitude of the arriving signal at a particular AoA as its *gain*. A distribution of signal gains, over all possible AoAs of $0 - 2\pi$ radians, is known as the *AoA distribution*. These signal paths are symmetric — a return signal transmitted from the BS to the UE along the same AoA will encounter the same gain.

5.2 How is the AoA Conventionally Estimated?

For the sake of clarity, we describe the conventional AoA estimation process using a Uniform Linear Array (ULA), where antennas are arranged in a line and with equal spacing between them. The extension to circular arrays (Hekaton uses circular arrays) is straightforward.

Consider a single signal path from a UE arriving at a phased-array at the BS. Figure 7 shows an example of this with two antenna elements at the phased-array. Since these two elements are spaced a distance of d apart, the additional propagation distance, travelled by the signal reaching the i^{th} antenna element is

$$\tau_i(\theta) = (i-1)\frac{d\sin\theta}{c} \tag{1}$$

where θ is the AoA and c is the speed of light. Note that for signals originating from UEs that are far away, the signals arriving the phased-array can be assumed to be parallel. The phase difference measured by all antennas is thus given by the column vector

$$\phi(\theta) = [1, e^{j2\pi f_c \tau_2(\theta)}, \ldots, e^{j2\pi f_c \tau_L(\theta)}]^T, \tag{2}$$

where f_c is the carrier frequency of the signal. For a single UE with P multipath signals $s_1, \ldots, s_P$ arriving at the BS, The received signal at the L antenna elements is thus

$$\mathbf{r} = [\phi(\theta_1), \ldots, \phi(\theta_P)][s_1, \ldots, s_P]^T + \mathbf{n} \tag{3}$$

where $\mathbf{n}$ is the channel noise energy.

In order to find the AoA distribution, we first discretize the angular space into D distinct, equally spaced, angles $\{\theta_1, \ldots, \theta_D\}$. We then scan through the discrete angular space of the phased-array and determine the gain at each discrete angle. This scanning process for l^{th} UE at the k^{th} phased-array antenna uses a correlation matrix $\mathbf{\Phi}$. The expression for the AoA distribution is

$$\mathbf{a}_l^k(\theta) = \mathbf{\Phi}\mathbf{r}. \tag{4}$$

5.3 How Does Hekaton Estimate the AoA?

Observe that with a phased-array antenna, because we cannot access each antenna element individually, the correlation procedure is replaced with D separate probes, each using a different row of the matrix $\mathbf{\Phi}$ as the active codebook entry.

Hekaton avoids this overhead by estimating the AoA from the CSI measurements of the channel. Consider, for the sake of simplicity, that Hekaton is estimating the AoA between a single phased-array antenna and the l^{th} UE in the cell. According to antenna theory [14], the measured CSI is related to the AoA distribution via an *Inverse Discrete Space Fourier Transform (IDSFT)*,

$$\mathbf{h}_l = \mathbf{F}^{-1}\,\mathbf{a}_l^k(\theta), \tag{5}$$

where $\mathbf{F}^{-1}$ is the IDSFT matrix and $\mathbf{h}$ is the (non-composite) CSI vector between the BS and the l^{th} UE.

Recall that Hekaton can only obtain the composite CSI, which is dependent on the active codebook entry (i.e. the weights of the phase-shifters). Hence, the actual CSI-AoA relationship that is used by Hekaton is

$$\hat{h}_l = \mathbf{b}^T\mathbf{h}_l = \mathbf{b}^T\mathbf{F}^{-1}\,\mathbf{a}_l^k(\theta), \tag{6}$$

where $\mathbf{b}$ is a vector specifying the weights of each of the phase-shifters, and $(\cdot)^T$ refers to a vector/matrix transpose. $\hat{h}_l$ is the composite CSI between the phased-array antenna and the l^{th} UE.

Unfortunately, the addition of the analog beamforming weights makes the CSI-AoA relation of (6) non-invertible as one can conceive of multiple AoA distributions that can map to the same composite CSI. This is a classic linear algebra problem — $\mathbf{a}_l^k(\theta)$ is a vector of length D and we can recover it precisely if we have D different measurements of $\hat{h}_l$, each taken using a different active codebook entry $\mathbf{b}$.

However, this is not a practical approach since D is typically very large ($D = 360$, in this case). Instead, Hekaton efficiently recovers the AoA distribution by exploiting its *sparsity* property.

5.4 AoA Sparsity

Empirical measurements [15] have shown that the AoA distribution is clustered. In a typical multipath environment, the dominant multipath components (i.e. those with the highest gain) arrive at the receiver from 3-5 distinct directions [15]. The signal gain at other AoAs are small and thus can be ignored as they do not contribute significantly to the fidelity of the received signal.

How is the AoA distribution sparse? If we discretize the angular space into $D = 360$ equally spaced angles, then $P < 5$ dominant signal paths is much smaller than the number of possible AoAs and is thus sparse. The signal gains along paths outside the P dominant ones are low and can be ignored as they do not contribute significantly to the decodability of the signal at the phased-array. The AoA measurement overhead is bounded by the number of multipath clusters. Because the AoA distribution is sparse, we can recover it from only a small, fixed number of composite CSI measurements.

5.5 Compressive AoA Estimation

We have seen how (a) the number of dominant AoA directions is sparse and (b) Hekaton obtains a transformed AoA distribution through CSI measurements. The objective of the compressive AoA estimation algorithm is to recover the AoA distribution $\mathbf{a}_l(\theta)$ from a fixed but small number of composite CSI measurements following (6).

Let $\hat{h}_{s,c}^{(1)}, \ldots, \hat{h}_{s,c}^{(V)}$ be a vector formed by V different composite CSI measurements, each taken with a different codebook entry $\mathbf{b}^{(1)}, \ldots, \mathbf{b}^{(V)}$. We can recover the AoA distribution using compressive sensing [16] via

$$\hat{\mathbf{a}}_l^k(\theta) = \arg\min \|\mathbf{a}_l^k(\theta)\|_1$$
$$\text{s.t. } \hat{h}_{s,c}^{(v)} = \mathbf{b}^{(v)}\mathbf{F}^{-1}\,\mathbf{a}_l^k(\theta),\ 1 \le v \le V \tag{7}$$

where $\|\cdot\|_1$ denotes the L1-norm.

How many CSI measurements are needed? Empirically, we have found that $V = 4$ is sufficient for recovering $\hat{\mathbf{a}}_{s,c}^k(\theta)$ accurately. This matches the measurements for the number of signal clusters for indoor environments [15]

We emphasize that unlike the naive, brute-force method, Hekaton's AoA estimation overhead is independent of the number of phased-arrays, or the number of antenna elements per phased-array.

5.6 AoA Estimation Example

Figure 8 shows the accuracy of the AoA estimation using the compressed sensing algorithm. We compare the AoA distribution estimated using 2 and 4 composite CSI measurements with a ground-truth distribution obtained using an exhaustive search of all possible AoA directions. Observe that even with only the few samples needed by the compressive-sensing algorithm, we can determine the AoA distribution peak accurately. The angle corresponding to this peak is the direction of the signal cluster from BS that results in the highest signal power gain at the UE.

6. DOWNLINK BEAM SELECTION

Hekaton can directly use a phased-array codebook entry that has a direction closest to the estimated AoA. However, a better approach is to chose a beam direction that maximizes the ratio of the signal energy to the intended UE to the total interfering signal energy to all *unintended UEs*. Give a phased-array antenna, an unintended UE is one that is either a UE in the local cell that is assigned to a different phased-array, or a UE in an adjacent cell. This metric is known as the *signal-to-leakage-power-ratio (SLR)*.

6.1 Selecting the SLR-maximizing Beam

Hekaton first determines the beam pattern between every phased-array antenna and every downstream user that maximizes the SLR. However, recall that each phased-array antenna in Hekaton only directs its beam towards a single user. Hence, Hekaton then assigns each user to the phased-array for which maximum SLR is achieved.

Hekaton uses the AoA distribution to users in both the local and adjacent cells to compute the SLR distribution. Simply put, the SLR distribution to a particular user l from a phased-array antenna k is defined as the element-wise ratio of the AoA distribution to this user, to the sum of all the AoA distributions from this same phased-array to concurrently active users in adjacent cells. Mathematically, this is expressed as

$$\eta_l^k(\theta) = \frac{\mathbf{a}_l^k(\theta)}{\sum_{d \in \mathcal{C}_U} \mathbf{a}_d^k(\theta)}, \tag{8}$$

where $\mathcal{C}_U$ is the set of unintended. Then we have the beam pattern from the phased-array to the user that maximizes the SLR:

$$\hat{\theta}_l^k = \arg\max_{\theta}\ \eta_l^k(\theta). \tag{9}$$

At this point, Hekaton has the SLR-maximizing beam direction between every phased-array antenna and every user. The next step is to construct a unique mapping between users and phased-array antennas, so that the SLR at the users are maximized. Hekaton uses a greedy algorithm for this.

For each user, Hekaton selects the phased-array antenna that can achieve the highest SLR when directing a beam to that user. Hekaton then pairs the phased-array antenna with that user. This is a one-to-one mapping: no user will be assigned to more than one phased-array antenna, and no phased-array antenna will be assigned more than one user. Hekaton then activates the codebook entry in each phased-array that corresponds to the beam direction which most closely matches its SLR-maximizing direction.

6.2 Downlink Beam Update Interval

The duration of the update interval of the analog beam has an impact on the performance of Hekaton. In environments with high mobility, the downlink beam must be constantly updated to track the user's direction of movement. Hence, a shorter update interval is necessary. Conversely, in a static environment, the beam update interval can be long.

Empirically, we have found that in an indoor office environment with typical mobility, Hekaton can outperform a comparable MU-MIMO transmission with the same number of RF chains if the analog update occurs every 10 frames.

7. IMPLEMENTATION

We prototype our two-level beamforming architecture using the WARPv3 software radio platform. Our prototype implementation

of Hekaton uses four WARPv3 boards, each with four antennas. This allows us to evaluate Hekaton in configurations utilizing up to 16 antennas in total (e.g. 4 RF streams each with a 4-antenna phased array). We rely on trace-driven emulations for larger-scale configurations that utilize more than 16 transmit antennas. The set of beams supported by our emulated phased-array is specified in a codebook. We employ another 2 WARPv3 boards to emulate up to 8 single-antenna clients.

Phased-Array Implementation. Hekaton design leverages analog phased-arrays to achieve efficient and practical large-scale MU-MIMO. However, fine-time-grained control over commercial phased-array antennas may require a custom FPGA controller[4]. For the sake of simplicity, we construct digital phased-array antennas using the WARPv3 platform.

The key difference between digital and analog phase-shifting is the phase resolution: digital phase-shifting is limited by the number of bits used to represent the phase angle, while its analog counterpart can shift the beam over a continuous set of angles. Fortunately, in practice, this resolution difference is very small and negligible.

In our implementation, we emulate circular phased arrays of up to eight antenna elements each, with corresponding codebooks of eight different beam patterns [17]. All phased-arrays used in any specific experiment have the same size. We also normalize the transmit power of each emulated phased-array to be equal to that of a single WARP RF chain.

We emulate each digital phased array by arranging up to eight antennas (each connecting to a WARP RF interface) in a circle. The radius of the circle is fixed at half the wavelength of the 2.4GHz carrier [18]. If multiple WARP devices are needed, they are synchronized via CM-MMCX clock modules [19] to eliminate sampling and frequency errors across the entire phased-array.

Phased-Array Codebook Design. The phased-array codebook is implemented using standard techniques [17] to obtain single-lobe beams with dominant directions equally spaced over 2π. The number of codebook entries is equal to the size of the phased-array antenna. All beamforming weights in the codebook are specified to the nearest one degree, so as to closely emulate the performance of real-world phased-arrays. The geometry of the phased array is a circle with radius $N\lambda/16$, where N is the total number of antenna elements in the codebook.

We note that our digital phased array implementation is different from conjugate beamforming. Unlike conjugate beamforming, the codebook only encodes phase, and not amplitude, changes.

Latency and Synchronization. Hekaton is optimized to minimize latency and maintain time/frequency-synchronizatoin across all its antenna elements.

(i) Low-Latency. We develop a C++ WARP software controller to enable simultaneous I/Q transfer between all WARP radios. OpenMP is used to accelerate PHY processing in parallel across multiple CPU cores. As a result, our implementation achieves an inter-frame PHY latency, in a 8×8 MU-MIMO transmission, of merely $23ms$ with an eight-core Intel i7 CPU. This latency is well below the empirical channel coherence time of $300ms$ in our office environment. We can thus perform real-time evaluation of Hekaton's performance under varying channel dynamics and mobility.

We note that the L1-norm minimization step required AoA estimation can take up to $200ms$, which is larger than the frame interval of our implementation. However, this does not impact our evaluations as the beam pattern of each phased-array antenna only needs to be updated asynchronously on a coarse timescale.

[4]The FCI-3710 [13] has a digital interface for use with FPGA controllers

(ii) Phased-Array Calibration. A key feature of phased-array antennas is that the transmit phases at each of its antenna elements is known to the phased-array. However, since we use different WARP radios for each antenna, hardware differences across the radios introduce varying phase-offsets between antennas. Hence, phase calibration mechanism must be applied to these antennas [1] to synchronize the phase of all antenna-elements in a phased array.

We synchronize all antenna-elements to a randomly chosen *primary antenna*. Recall that the exact placement of each antenna is fixed and known in advance. The primary antenna simply transmits an OFDM preamble. Each non-primary antenna uses the received phase information, along with its known physical location, to derive its true phase-offset from the primary antenna. It then compensates for this offset in all subsequent transmissions. This calibration is done for both uplink and downlink. We have empirically determined the phase-offsets between antennas remain unchanged for up to several hours (the duration of our experiments).

(iii) Scheduling. Cellular protocols are synchronous and rely on the presence of a global protocol clock. Since all WARP boards are connected to a central PC, we enforce this protocol clock without additional outside synchronization.

Digital Beamforming. We implement digital beamforming component of Hekaton using a 802.11-based OFDM PHY. The 20 MHz channel bandwidth is divided into 64 subcarriers, including 48 data subcarriers, 4 pilots, 1 DC and 11 guard band subcarriers. QAM-modulated input data symbols are precoded using zero-forcing beamforming (ZFBF).

8. EXPERIMENTAL EVALUATION

In this section, we demonstrate Hekaton's performance based on the aforementioned testbed implementation. Our evaluation answers three major questions: *(i)* Can Hekaton achieve capacity scaling while maintaining energy efficiency by simply increasing the number of passive antennas? *(ii)* Can each component of Hekaton effectively limit the coordination overhead, thus achieving throughput efficiency for large-scale MU-MIMO? *(iii)* Can Hekaton still manage interference in multi-cell networks through its two-level beamforming?

Hekaton is equally applicable to both indoor and outdoor environments. However, the indoor environment presents more challenging wireless channel conditions due to the abundance of multipath reflections. We thus conduct our experiments in a typical office environment (Figure 15) during normal office hours with around $300ms$ channel coherence time.

8.1 Experimental Setup

Hekaton configuration. Unless otherwise indicated, each Hekaton node in our experiments has two RF chains, each connected to an eight-antenna phased-array. Therefore each Hekaton node has two degrees-of-freedom and can serve up to two users concurrently. Trace-driven emulation is used for larger number of users.

Baseline MU-MIMO configuration. In our evaluation, we compare Hekaton with conventional MU-MIMO that has only a single omnidirectional antenna per RF chain. Unless otherwise indicated, the baseline MU-MIMO configuration shares the same user number and RF chain number with the Hekaton configuration. In Section 8.3, we also consider the oracle MU-MIMO configuration that exhaustively searches among all user subsets with all possible cardinalities to maximize the downlink capacity. To ensure a fair comparison, Hekaton and the baseline MU-MIMO configurations share the same center frequency, channel bandwidth and digital precoder specified in Section 7.

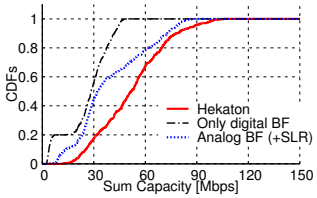


Figure 9: Combining digital and analog beamforming improves the performance.

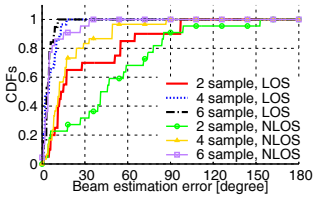


Figure 10: Accuracy of compressive AoA estimation.

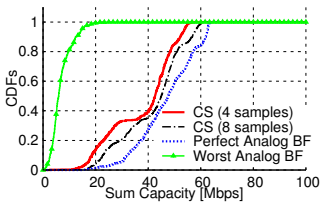


Figure 11: Data rate of AoA-based beamforming.

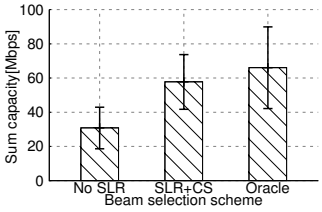


Figure 12: Capacity of SLR based analog beam selection.

Trace-driven emulation setup. Due to the limited number of WARP nodes available, we conduct trace-driven emulation for the scenarios that require more than 6 WARP nodes, *e.g.*, the multi-cell evaluation in Section 8.4. In order to collect these traces, we first note the antenna locations of the virtual phased array of the intended size, as described in §7. We then place the antennas of the physical transmitter on these locations in turn and collect the CSI between that antenna and multiple receivers via over-the-air transmission. The collected channel matrices are finally concatenated to emulate a large-scale multi-antenna transmitter serving the same user group.

During the trace collection, we ensure that the users and the environment are static (with around $300ms$ coherence time using 0.5 as the threshold for channel correlation). The trace collection *w.r.t.* a particular user group takes up to 40 minutes. This long measurement duration will reduce the overall spatial correlation between antennas. However, we note that this will affect both MU-MIMO and Hekaton similarly, and thus will not affect performance comparisons between MU-MIMO and Hekaton.

8.2 Hekaton Micro-Benchmarks

Analog Beamforming vs Digital Precoding. We compare the performance of Hekaton with two alternative architectural choices with the same number of RF chains: *(i) Standalone analog beamforming*: The base station has two 8-antenna phased-arrays, each connected to a separate RF chain. It beamforms to 2 users with minimized SLR but without digital precoding to eliminate the crosstalk; *(ii) Standalone digital precoding*: The base station has 2 omnidirectional antennas, each also connected to a separate RF chain. It employs a zero-forcing beamforming (ZFBF) precoder to serve 2 downlink users. Our Hekaton base station has the same configuration to the standalone analog beamforming except that it also uses digital precoding to eliminate the crosstalk between the two concurrent data streams.

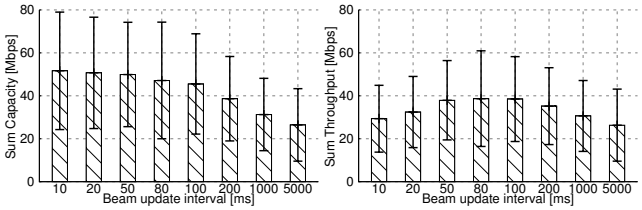
Figure 9 plots the CDFs of the sum downlink capacity under the above three architectures. It shows that Hekaton's median throughput is 66% higher than that of standalone analog beamforming. This is due to the distortion of the spatial beams in practice. Ideally, the two phased arrays can steer their beams towards orthogonal directions. However, multipath reflections indoor will easily distort the beams, resulting in crosstalk interference among different users. Hekaton overcomes this limitation by combining with digital ZFBF precoding to eliminate this residual interference after analog beamforming. On the other hand, compared with standalone digital precoding, the analog beamforming component of Hekaton essentially acts as a preconditioner, which makes the digital channel of the 2 users "seen" by the RF chain more orthogonal and reduces the condition number of the channel matrix. As a result, Hekaton also achieves substantial (median 76%) improvement over standalone digital precoding.

Compressive AoA Estimation. We now evaluate the accuracy of Hekaton's compressive AoA estimation algorithm using a single RF chain and a single eight-antenna phased-array. Note that due to the limited number of antennas in this setup, we can only make use of up to eight composite CSI measurements, thus limiting the number of recoverable multipath clusters to eight. However, this is already sufficient to identify the 3-5 clusters in typical indoor environments [15]. Based on the compressive sensing theory [9], we believe that our results in this evaluation are still valid for larger scale phased array under the same environment. With the eight-antenna phased array connecting to a single RF chain on the base station and one user sending the uplink probing signal, we first estimate the AoA of the strongest incoming signal path based on eight composite CSI measurements, and then study the error incurred when fewer composite CSI measurements is used. This comparison is made under 20 random line-of-sight (LOS) topologies and an equal number of non-line-of-sight (NLOS) channels.

Figure 10 plots the CDF of these AoA estimation errors of the strongest incoming signal path over all users in all topologies. Recall that Hekaton uses four CSI measurements for AoA estimation. We can observe that Hekaton has only around 6° of median error in LOS environment since the signal AoA is dominated by an LOS component. Meanwhile, the NLOS channel results in slightly higher error (median 12°) due to richer multi-path reflections. In both cases (LOS and NLOS), the AoA error should only have a marginal impact to Hekaton's beam selection from the eight-entry circular phased-array codebook (about 45° beam width).

To verify this, we further evaluate the impact of the compressive AoA estimation on Hekaton's beam selection and the resulting network capacity. In this micro-benchmark evaluation, We exclude the influence of the digital MU-MIMO precoder by running a Hekaton base station with only one RF chain connecting to an eight-antenna phased array. This modified Hekaton setup serves only one user. We compare the compressive approach using four or all eight composite CSI measurements with 2 other configurations in the same antenna and RF chain setup: *(i) Perfect analog BF*, which uses conjugate beamforming to steer the beams of each phased-array. Conjugate beamforming is equivalent to having a discrete codebook with infinitely fine resolution in theory (bounded by the hardware in practice), and thereby serves as a performance upper-bound. We emphasize that typical analog phased-array antennas are codebook-based and cannot perform conjugate beamforming; *(ii) Worst analog BF*, which intentionally chooses the worst analog beam pattern that leads to the lowest downlink capacity based on exhaustive search from the codebook, and thus acts as the performance lower bound.

Figure 11 plots the downlink capacity of these configurations, we see that Hekaton's compressive approach with four measurements only causes 5% median capacity loss compared to the one with eight measurements. Both schemes exhibit gains of more than $7\times$ over the worst-analog BF case. They also exhibit between 10% to 30% lower capacity than the perfect analog BF case due to the



(a) Without probing overhead. (b) With probing overhead.

Figure 13: The relationship between Hekaton performance and analog beam update interval.

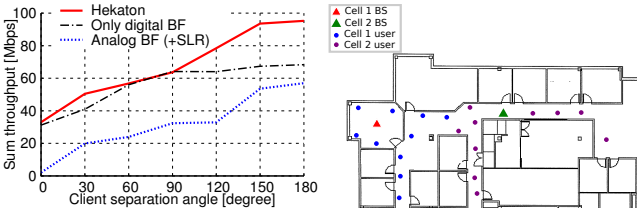


Figure 14: Spatial separation between users affects downlink driven beamforming.

Figure 15: Topology for trace-driven emulation.

discrete codebook they used, but this gap will naturally decrease as the number of antenna elements grows (which results in a codebook of finer resolution).

SLR-Based Beam Selection. Hekaton uses the SLR metric to select downlink beam for each phased array that minimize interference to non-intended users. In this micro-benchmark evaluation, we compare the performance of the SLR-based beam selection to two other configurations: *(i) Peak-AoA-based*, where we only use the peak AoA direction as the beam direction and *(ii) Oracle*, where we select the downlink beam direction assuming full non-composite CSI knowledge at the BS.

We test 30 topologies where a Hekaton base station with two eight-element phased-arrays serves two randomly selected users simultaneously. Figure 12 shows that the capacity under the SLR-based beam selection outperforms that of the AoA-based approach by almost 50% on average. Furthermore, the SLR-based approach achieves only 12% lower capacity than the oracle with optimal beam selection.

Update Interval vs Downlink Throughput. Hekaton must continuously update the beam direction so that it accurately "follows" the motion of mobile users. However, note that during a coarse-timescale beam update procedure (spanning multiple frames), random codebook entries are activated on the phased-array. This results in non-ideal analog beamforming for upstream and downstream transmissions for the duration of the update. Obviously, a non-stop update of the phased-array will have a detrimental impact on the achievable throughput.

We evaluate this trade-off in a dynamic topology where users in the testbed move at walking speed. We use the Hekaton base station with two 8-element phased arrays to serve two LOS users simultaneously. Sum downlink capacity of the two users are evaluated under different beam update intervals. Each experiment lasts for 5 minutes, and the resulting mean capacity is plotted in Figure 13a.

If we ignore the overhead due to beam update and probing, the results follow an expected pattern: the shorter the interval between updates, the higher the throughput, as seen in Figure 13a. However, due to the degraded channel during the analog beam update, we can see from Figure 13b that the best update interval is $80ms$, or once every 80 LTE subframes. Hence, Hekaton updates its active codebook entry every $80ms$.

Impact of Spatial Correlation. Since a Hekaton BS relies on analog beamforming as a first-level separation between concurrently-served users, its effectiveness is naturally affected by spatial separation of the users. We study this performance factor with our Hekaton base station and two downlink users. The Hekaton base station has 2 RF chains, each are connected to a circular phased array with 8 antenna elements. The angular separation between the two users (with respect to the BS) varies from 0 to 180 degrees. For each angular separation, we create 30 different topologies (all the while maintaining the same angular separation between the two users), and plot the mean downlink capacity in Figure 14. For comparison, we also repeat the experiments with standalone analog and digital beamforming.

In general, the performance of Hekaton and standalone analog beamforming keeps increasing with node separation since it becomes easier to form non-overlapping beams pointing towards the users. Standalone digital precoding is relatively less affected by node separation because it operates based on precise CSI information of each user. We also see that Hekaton and standalone digital precoding achieve similar performance under small node separation since the beams from the Hekaton phased-array antennas overlap with each other in this scenario and hence the interference cancellation is mainly left to the digital precoder. Hence, explicitly selecting distant users will improve Hekaton's overall performance.

8.3 Single-Cell Performance

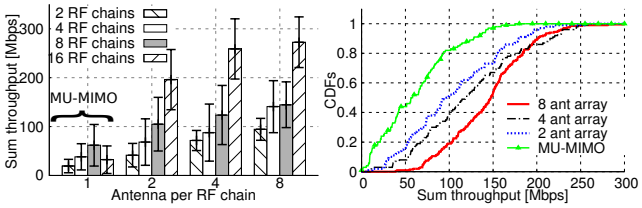
Throughput. We first examine how the throughput of Hekaton and conventional MU-MIMO scales in terms of the number of RF chains and antenna elements. We vary the number of RF chains in both configurations between 2 and 16. The number of users served by both cases is always equal to the number of RF chains used. For a given number of RF chains, we also vary the number of antenna elements in the phased-array connected to each RF chain to between 1 and 8 (1-antenna-per-RF-chain configuration corresponds to conventional MU-MIMO). Trace-driven emulation is used for the configurations where the total number of antennas exceeds the maximum number of 16 antennas supported by our implementation. The traces are collected using the Cell 1 topology in Figure 15 following the methodology detailed in Section 8.1.

Figure 16a shows the average downlink throughput and standard deviation across randomly selected downlink user combinations. We observe two interesting trends from the results.

First, when we only have a single antenna per RF chain, the sum throughput can decrease even if we increase the number of RF chains. For example, the sum throughput in the 16×16 configuration (16 RF-chains serving 16 users) is lower than the 8×8 configuration. This is because the channel orthogonality between users diminishes with increasing user number. We note that real-world MU-MIMO systems do not always operate in such a configuration to avoid this throughput scaling behavior.

Second, for a given number of RF chains, Hekaton scales up the network capacity by increasing the number of antenna elements per RF chain. The phased-array antennas effectively improves the channel orthogonality between users and reduces the channel condition number. Hence, sum throughput of Hekaton in 16×16 consistently outperforms the 8×8 Hekaton configuration.

In the results above, MU-MIMO cannot exploit beamforming gain since its user count is equal to its antenna count. However,



(a) Hekaton capacity scaling in a single cell.

(b) Transmitter with 8 RF chains serving 4 users.

Figure 16: Single cell performance comparison.

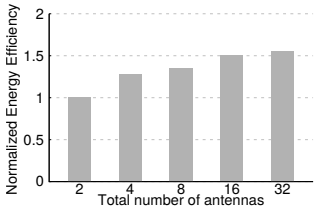


Figure 17: Energy-efficiency over conventional MU-MIMO with the same RF chain number.

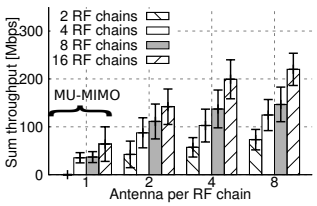


Figure 18: Throughput in 2-cell topologies.

Hekaton still achieves a significant throughput gain over MU-MIMO with fewer users. Here, we use MU-MIMO with 8 RF chains (i.e. 8 antennas) serving 4 users as the baseline. Note that this is the configuration that achieves the highest throughput in Figure 1. We compare MU-MIMO in this configuration against Hekaton with 2 RF chains and 2,4, and 8 antennas in each phased-array.

Figure 16b shows the throughput distribution over all topologies. We see that Hekaton achieves an impressive increase in throughput under the same transmit power — with 8 RF chains and 8-element phased-arrays serving 4 users, Hekaton achieves $2.5\times$ throughput gain over the 8×4 conventional MU-MIMO configuration.

Energy Efficiency. We have already seen that the throughput of Hekaton increases with the size of phased-array. We further demonstrate that the two-level beamforming architecture employed by Hekaton improves energy-efficiency alongside this throughput increase. We benchmark Hekaton with two RF chains and varying phased-array sizes, serving two users. The energy efficiency values (in bits-per-Joule) is normalized using 2×2 MU-MIMO as the baseline. Both Hekaton and the 2×2 MU-MIMO configuration have the same number of RF chains, although Hekaton uses phased arrays with each RF chain while MU-MIMO uses omnidirectional antennas. Note that since the power consumed by the phased array is typically very small, as discussed in Section 1, we assume that the power consumption of both of these configurations are equal. For the sake of clarity, we report results w.r.t. the total number of antennas used. Figure 17 shows this mean normalized energy efficiency. Since Hekaton supports more transmit antennas than conventional MU-MIMO under the same RF chain number thanks to its unique two-level beamforming architecture, we see that its energy-efficiency almost shows a monotonically increasing trend, and the energy efficiency gain is up to $1.67\times$ over conventional MU-MIMO with the same number of RF chains. We expect even greater efficiency gain with larger phased-array antennas.

8.4 Multi-Cell Performance

Hekaton uses two-level beamforming to simplify inter-cell coordination and reduce interference between adjacent cells. In this section, we compare its performance against conventional MU-MIMO in a two-cell network. Due to the large number of antennas required, our evaluation will rely primarily on trace-driven emulation. The locations of these two cells are chosen so induce inter-cell interference, and is shown in Figure 15.

We evaluate the sum throughput across both cells for the BSes with 2-, 4- and 8-RF chains. Hekaton uses phased-array antennas of varying sizes, similar to the configuration used in the single-cell experiments in §8.3. The number of users is always equal to the number of RF chains.

Figure 18 shows that Hekaton achieves significant throughput gain over conventional MU-MIMO. Specifically, due to the strong inter-cell interference, conventional MU-MIMO (1 antenna per RF chain) with only 2 RF chains achieves almost zero throughput. As we increase the multiplexing gains (by increasing the number of RF chains), MU-MIMO throughput increases marginally to a mean of 60 Mbps. On the other hand, as we increase beamforming gains in Hekaton, it can achieve throughput that is up to 300% that of MU-MIMO. Note that the transmit power in both Hekaton and MU-MIMO experiments is identical. Hence, this gain comes from (a) the ability of Hekaton to improve beamforming gain and better focus the signal energy on the intended users, and (b) the ability of Hekaton to direct signal energy *away* from unintended users. Hence, Hekaton can increase the throughput to its intended users, while simultaneously reducing inter-cell interference.

MOS Video Quality	PSNR Range (dB)
Excellent	>37
Good	31-37
Fair	25-31
Poor	20-25
Bad	< 20

Table 1: Mapping between video quality and PSNR.

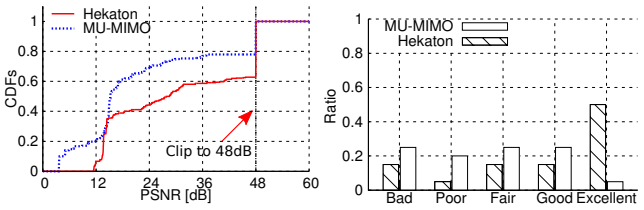
8.5 Impact of Hekaton on Application Performance

In this section, we combine the PHY layer implementation of Hekaton with application-layer simulation. Without loss of generality, our experiments compare two different large-scale MU-MIMO designs: *(i) Hekaton* BS with two phased-arrays (eight antenna elements each), and *(ii) conventional MU-MIMO*, with 2 RF chains. Both of these networks serve two single-antenna users concurrently.

8.5.1 Hekaton with H.264 Video Streaming

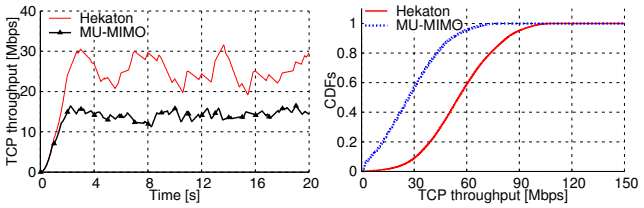
We first implement a H.264 video streaming emulator on top of our Hekaton testbed. The emulator runs at the BS that has two RF chains and transmits different H.264 video streams to two different users simultaneously. Upon receiving and decoding the video stream, each user computes its PSNR and user-perceived Mean Opinion Score (MOS) (the standard measures of video quality). The relationship between the PSNR and video quality [20] is summarized in Table 1. We evaluate both Hekaton and MU-MIMO across 10 different pairs of users using QPSK and 1/2-rate FEC, each test transmits 100 different 40-second videos.

Figure 19a shows the PSNR distribution over all clients and across all video frames (original PSNR of the videos is 48dB). Hekaton achieves *twice* the median PSNR, from 15 to 30dB, when compared to MU-MIMO. Figure 19b shows the corresponding MOS video quality. We see that Hekaton increases the ratio of excellent video streams by up to $10\times$ and reduces the ratio of all lower-quality video frames. With the base station's total transmit power unchanged, Hekaton's performance improvement comes by improv-



(a) PSNR distribution. (b) Video quality.

Figure 19: Video streaming evaluation.



(a) Time domain throughput comparison for one transmission. (b) Throughput CDF comparison.

Figure 20: TCP throughput comparison in an FTP application.

ing the beamforming gain and user channel orthogonality via the phased array.

8.5.2 *Hekaton with FTP File Transfer*

We then emulate the FTP protocol over TCP using PHY channel traces. These traces are obtained by running Hekaton over our testbed with the same antenna and RF chain setup in Section 8.5.1, and extracting the per-packet SNR (and achievable bit-rate) over time. The FTP data is fragmented into 1460-byte frames. The receiver window size, TCP maximum congestion window size, and round-trip delay are set to 200, 100 and 20ms, respectively. We assume the downlink and uplink have symmetrical rate.

Figure 20a shows the time-varying performance of one FTP transfer. We observe that even after taking into account the effects of slow start, signaling overhead, and the AIMD congestion control, Hekaton still outperforms convention MU-MIMO with respect to long-term throughput. Figure 20b further plots the CDF of TCP-flow throughput distributed across 10 pairs of clients. Hekaton improves the median TCP throughput of all TCP flows by more than $2\times$ over conventional MU-MIMO.

9. DISCUSSION

Is Hekaton useful with more users? One may assume that in networks with a large number of users, one will always be able to find users with uncorrelated channels due to the user diversity. Accordingly, the gain (and usefulness) of large-scale MU-MIMO is limited under large user population. However, such an assumption ignores the fact that the CSI measurement overhead needed to locate these uncorrelated users scales *linearly* with the total number of users. This overhead can easily negate any benefit of a larger users selection pool [21]. On the other hand, the additional overhead incurred by Hekaton is independent of the user population.

One- vs Two-Level Beamforming. Instead of increasing the number of RF chains, Hekaton adopts a two-level approach and relies on phased-array antennas to get higher total antenna number and hence lower channel condition number.

A phased-array antenna consists of multiple (*e.g.*, $K > 1$) antenna elements. However, a K-element phased-array antenna is not equivalent to K RF chains, each with an omni-directional antenna. Each antenna element is only equipped with an analog phase shifter that adjusts the phase but not the magnitude of the antenna signal. Hence, phased-array antennas cannot achieve the same beamforming gain or downlink capacity as a group of RF chains (with associated omni-directional antennas) of the same size.

However, increasing the number of MU-MIMO RF chains results in a proportional increase in the energy consumed. We emphasize that increasing the number of RF chains beyond the rank of the channel will only increase the beamforming gain. At this operating point, we can obtain a similar improvement of the beamforming gain with Hekaton, but without the associate increase in energy consumption.

Hekaton is a retrofittable solution that can improve the beamforming gain of existing wireless platforms without additional energy consumption. If backwards-compatibility is not a requirement, other energy-efficient operating points can be achieved by increasing RF chains instead of antenna elements, and employing one-level beamforming instead. We leave such studies to future work.

Beamforming in the 3-D space. In this paper, we use the standard codebook designed for the 2-D space [17], and thus the signal arriving path or AoA is correspondingly estimated in the 2-D space to facilitate the downlink beam selection. We emphasize that our design can also be used to estimate the signal arriving path in the 3-D space if codebook designed for 3-D space is available in the phased-array. There are on-going discussions on the 3-D beamforming in 3GPP. We expect that the knowledge of signal path in the 3-D space will allow us to better target the beams, especially in urban environments, and we leave this to future work.

Optimizing Hekaton. Currently, Hekaton does not exploit the full potential of two-level beamforming architecture because (a) it is limited to fixed number of users and (b) we enforce a one-to-one mapping between each user and phased array to simplify the problem formulation. It is worth noting that MU-MIMO schedulers [22,23] can optimize the users served, and similar approaches can also be applied to Hekaton . However, such extensions will increase the complexity of Hekaton. Given that our suboptimal Hekaton design already shows remarkable performance gains over conventional MU-MIMO, it is not immediately clear if the additional complexity will result in a commensurate performance improvement. We leave an in-depth study of this challenge to future work.

Hekaton with implicit and explicit CSI feedback. When explicit CSI feedback is used for MU-MIMO as in most existing wireless standards like 802.11ac and LTE, Hekaton shows an obvious advantage over existing large-scale MU-MIMO architecture [1, 2] due to the much smaller number of RF chains. However, when implicit CSI feedback is used, the probing process Hekaton used to find the SLR-optimal beam direction incurs extra overhead. Fortunately, it is mitigated by the fact that the additional probing overhead in Hekaton is a small constant number. We emphasize that this is not a constant scaling factor, but a constant of four probes due to the indoor signal sparsity [15]. Hence, when taken together with its two-level design, Hekaton is still an efficient, retrofittable design for large-scale MU-MIMO networks.

Hekaton with Mobility. Due to the limitation of our implementation, we only evaluate Hekaton in static environments with around $300ms$ channel coherence time. However, it is worth noting that the novel two-level beamforming architecture of Hekaton may improve the MU-MIMO performance under mobility if supported by proper hardware with sufficient processing capability. Since the analog beamforming component reduces the crosstalk interference between concurrent data streams and only needs to be updated in a coarse time manner, the system becomes more robust to the stale

CSI used in the digital precoder compared to a pure digital beamforming system.

10. RELATED WORK

Large-scale MU-MIMO. Several research platforms, such as Argos [1] and BigStation [2], have demonstrated the significant performance gain that can be achieved by large-scale MU-MIMO systems — Argos with 64 antennas achieves $12.7\times$ the capacity of a single-antenna system; BigStation [2], when using 12 antennas and 9 clients, achieves $6.8\times$ the capacity of a single-antenna-single-client setup. However, these platforms suffer from scalability issues. For example, the authors in [1] acknowledge that suboptimal conjugate beamforming is preferred in larger systems as ZFBF suffers from unacceptable complexity and overhead. In [2], parallel computing is used to reduce the processing time of ZFBF, but the solution requires powerful computing server that is not available in current base stations. Furthermore, when the number of RF chains goes well beyond the channel rank supported by current environment [3], further improving capacity by increasing the number of RF chains may reduce energy efficiency.

In contrast, Hekaton aims to achieve good energy efficiency, low hardware complexity, and affordable computational overhead while enjoying a good fraction of the potential performance gain of a large-scale MU-MIMO system by using phased arrays. We emphasize that Hekaton is designed to enhance the performance of MU-MIMO with the same number of RF chains, this design philosophy does not conflict with existing large-scale MU-MIMO architectures, it is possible to replace omni-directional antennas with phased-array ones in systems like Argos or BigStation to further boost the network performance.

Two-Level Beamforming. While the high-level concept of combining an analog beamformer with a digital RF chain is not new, Hekaton is the first design that efficiently adopts this architecture in a MU-MIMO communications system. ProBeam [7] is a comparable work that integrates a phased-array antenna with a non-MIMO WiMAX basestation. However, the optimal beam is chosen using an exhaustive search of all codebook entries. This overhead scales with the size of the phased-array antenna. Two-level digital beamforming is discussed in [24]. However, since both levels of beamforming are digital, to support a large number of antennas, such a system still requires the same number of RF chains. Two-level beamforming is also prevalent in 60 GHz networks [25, 26]. However, exhaustive beam searches are typically employed there too. Faster beam searches, such as those based on simulated annealing [27], can reduce the search time, but the overhead still increases with the size of the phased-array.

Joint optimization schemes [28, 29] have also been proposed for two-level beamforming. However, these algorithms require tight integration between the analog phased-array and digital RF chains. This level of integration is not feasible for a solution that is to be backwards compatible with existing BSs.

Two-level beamforming is also employed in other areas such as MIMO radar [30]. However, such systems are purpose-built for object tracking, not communications, and also require tight coordination between the analog and digital RF components.

Coordination Overhead. The distributed coordination scheme in [31] increases throughput through opportunistic use of degrees of freedom. However, it requires precise clock phase and frequency synchronization. Other centralized coordinated multipoint systems [32–36] demonstrate impressive gains from cooperative transmissions across access points, but also comes at the cost of significant synchronization and inter-cell CSI sharing overhead. Such an overhead is impractical for large-scale deployment in real-world cellular networks.

AoA Estimation. ArrayTrack [8] demonstrates a practical method of localizing indoor WiFi clients using MIMO techniques. This approach uses standard AoA estimation techniques, together with multipath smoothing to determine the direction of each client from the BS. However, ArrayTrack cannot be integrated into Hekaton as it requires one RF chain for each antenna. Furthermore, due to multipath effects, the downlink beam direction used by Hekaton is not necessarily the direction of the direct path from the BS to the UE.

PinPoint [12] has a similar goal of localizing wireless clients. However, it relies on the cyclostationary properties of many existing wireless protocols to estimate the AoA. It too, requires that each antenna is connected to a single RF-chain, and thus, cannot be used in Hekaton. The method in [9] is most similar to the technique in Hekaton as it too relies on a compressive-sensing approach to AoA estimation. However, its compressive-sensing method differs from that in Hekaton in one critical aspect: it requires one-antenna-per-RF-chain, and thus is unsuitable for use with phased-array antennas. Recall that we cannot get time-domain signal information directly from each antenna element. Hekaton uses a key relationship between CSI and AoA to extract AoA information from composite CSI data, and thus requires no modifications to the LTE protocol.

11. CONCLUSION

Large-scale MU-MIMO has the potential to introduce multi-fold capacity gains into wireless networks at high spectral efficiency. However, in practice, these gains are plagued by the diminishing user channel orthogonality, high coordination overhead and the lack of support in current standards. In this paper, we designed and evaluated Hekaton, a novel two-level, two-timescale beamforming architecture that brings a sizable fraction of these gains into existing networks in a standards-agnostic form. In our evaluation, Hekaton achieves up to $2.5\times$ capacity gain over conventional MU-MIMO in single-cell networks, without additional energy requirements.

Acknowledgement

We thank the anonymous shepherds for their helpful comments on the paper. This research was supported in part by the NSF under Grant CNS-1318292, CNS-1343363 and CNS-1350039.

12. REFERENCES

[1] C. Shepard, H. Yu, N. Anand, E. Li, T. Marzetta, R. Yang, and L. Zhong, "Argos: Practical Many-Antenna Base Stations," in *Proc. of ACM MobiCom*, 2012.

[2] Q. Yang, X. Li, H. Yao, J. Fang, K. Tan, W. Hu, J. Zhang, and Y. Zhang, "Bigstation: Enabling scalable real-time signal processing in large mu-mimo systems," in *SIGCOMM*, 2013.

[3] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge University Press, 2005.

[4] E. Aryafar, N. Anand, T. Salonidis, and E. W. Knightly, "Design and Experimental Evaluation of Multi-User Beamforming in Wireless LANs," in *MobiCom*, 2010.

[5] D. Love, R. Heath, V. Lau, D. Gesbert, B. Rao, and M. Andrews, "An overview of limited feedback in wireless communication systems," *IEEE JSAC*, 2008.

[6] S. Jeon, A. Babakhani, and A. Hajimiri, "Integrated phased arrays," in *Advanced Millimeter-Wave Technologies*. John Wiley and Sons, Ltd, 2009, ch. 14.

[7] J. Yoon, K. Sundaresan, M. A. Khojastepour, S. Rangarajan, and S. Banerjee, "Probeam: A practical multicell

beamforming system for ofdma small-cell networks," in *Proc. of ACM MobiHoc*, 2013.

[8] J. Xiong and K. Jamieson, "Arraytrack: A fine-grained indoor location system," in *NSDI*, 2013.

[9] S. Hong and S. Katti, "Cognitive spatial degrees of freedom estimation via compressive sensing," in *CoRoNet*, 2010.

[10] Cisco Systems Inc., "Wireless Mesh Access Points, Design and Deployment Guide," *Release 7.3*, 2012.

[11] O. Arnold, F. Richter, G. Fettweis, and O. Blume, "Power consumption modeling of different base station types in heterogeneous cellular networks," in *Future Network and Mobile Summit*, 2010.

[12] K. Joshi, S. Hong, and S. Katti, "Pinpoint: Localizing interfering radios," in *NSDI*, 2013.

[13] F. Comtech, "Fci-3740 phased array antenna," http://www.fidelity-comtech.com/products/phased-array-antennas.

[14] S. J. Orfanidis, *Electromagnetic Waves and Antennas*, 2014. [Online]. Available: http://eceweb1.rutgers.edu/ orfanidi/ewa/

[15] N. Czink, X. Yin, H. Ozcelik, M. Herdin, E. Bonek, and B. Fleury, "Cluster characteristics in a mimo indoor propagation environment," *Trans. on Wireless Comms*, 2007.

[16] E. Candes, J. Romberg, and T. Tao, "Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information," *Trans. on Information Theory*, 2006.

[17] W. Feng, Z. Xiao, D. Jin, and L. Zeng, "Circular-antenna-array-based codebook design and training method for 60ghz beamforming," in *IEEE WCNC*, 2013.

[18] R. Vescovo, "Array factor synthesis for circular antenna arrays," in *Antennas and Propagation Society International Symposium*, 1993.

[19] "CM-MMCX Clock module," http://warpproject.org/trac/wiki/HardwareUsersGuides/CM-MMCX.

[20] S. Sen, S. Gilani, S. Srinath, S. Schmitt, and S. Banerjee, "Design and implementation of an "approximate" communication system for wireless media applications," in *SIGCOMM*, 2010.

[21] X. Xie, X. Zhang, and K. Sundaresan, "Adaptive Feedback Compression for MIMO Networks," in *ACM MobiCom*, 2013.

[22] X. Xie, X. Zhang, and E. Chai, "Cross-Cell DoF Distribution: Combating Channel Hardening Effect in Multi-Cell MU-MIMO Networks," in *Proc. of ACM MobiHoc*, 2015.

[23] X. Xie and X. Zhang, "Scalable User Selection for MU-MIMO Networks," in *IEEE INFOCOM*, 2014.

[24] "Two-Layer Linear Processing for Massive MIMO on the TitanMIMO Platform," http://nutaq.com/en/library/whitepaper-news/new-paper-two-layer-linear-processing-massive-mimo-titanmimo-platform.

[25] S. Yoon, T. Jeon, and W. Lee, "Hybrid beam-forming and beam-switching for ofdm based wireless personal area networks," *JSAC*, 2009.

[26] J. Wang, Z. Lan, C.-W. Pyo, T. Baykas, C.-S. Sum, M. A. Rahman, J. Gao, R. Funada, F. Kojima, H. Harada *et al.*, "Beam codebook based beamforming protocol for multi-gbps millimeter-wave wpan systems," *IEEE Journal on Selected Areas in Communications (JSAC)*, 2009.

[27] B. Li, Z. Zhou, H. Zhang, and A. Nallanathan, "Efficient beamforming training for 60-ghz millimeter-wave communications: A novel numerical optimization framework," *IEEE Trans. on Vehicular Tech*, 2014.

[28] J. Nsenga, A. Bourdoux, and F. Horlin, "Mixed analog/digital beamforming for 60 ghz mimo frequency selective channels," in *IEEE ICC*, 2010.

[29] A. Alkhateeb, O. El Ayach, G. Leus, and R. W. Heath, "Hybrid precoding for millimeter wave cellular systems with partial channel knowledge," in *Information Theory and Applications Workshop (ITA), 2013*, 2013.

[30] D. Fuhrmann, J. Browning, and M. Rangaswamy, "Signaling strategies for the hybrid mimo phased-array radar," *Selected Topics in Signal Processing, IEEE Journal of*, 2010.

[31] K. C.-J. Lin, S. Gollakota, and D. Katabi, "Random Access Heterogeneous MIMO Networks," in *ACM SIGCOMM*, 2011.

[32] H. S. Rahul, S. Kumar, and D. Katabi, "Jmb: Scaling wireless capacity with user demands," in *SIGCOMM*, 2012.

[33] H. V. Balan, R. Rogalin, A. Michaloliakos, K. Psounis, and G. Caire, "Achieving High Data Rates in a Distributed MIMO System," in *Proc. of ACM MobiCom*, 2012.

[34] H. Dahrouj and W. Yu, "Coordinated Beamforming for the Multicell Multi-Antenna Wireless System," *IEEE Transactions on Wireless Communications*, vol. 9, no. 5, 2010.

[35] C.-B. Chae, I. Hwang, R. Heath, and V. Tarokh, "Interference Aware Coordinated Beamforming in a Multi-Cell System," *IEEE Transactions on Wireless Communications*, vol. 11, no. 10, 2012.

[36] Y. Noam and A. Goldsmith, "Exploiting Spatial Degrees of Freedom in MIMO Cognitive Radio Systems," in *Proc. of IEEE ICC*, 2012.