Argos: Practical Many-Antenna Base Stations

Clayton Shepard[1], Hang Yu[1], Narendra Anand[1], Li Erran Li[2],
Thomas Marzetta[2], Richard Yang[3], and Lin Zhong[1]

[1]Rice University, Houston, TX
{cws, hang.yu, nanand, lzhong}@rice.edu

[2]Bell Labs, Murray Hill, NJ
{erranlli, tlm}@research.bell-labs.com

[3]Yale University, New Haven, CT
yry@cs.yale.edu

ABSTRACT

Multi-user multiple-input multiple-output theory predicts manyfold capacity gains by leveraging many antennas on wireless base stations to serve multiple clients simultaneously through multi-user beamforming (MUBF). However, realizing a base station with a large number antennas is non-trivial, and has yet to be achieved in the real-world.

We present the design, realization, and evaluation of *Argos*, the first reported base station architecture that is capable of serving many terminals simultaneously through MUBF with a large number of antennas ($M \gg 10$). Designed for extreme flexibility and scalability, Argos exploits hierarchical and modular design principles, properly partitions baseband processing, and holistically considers real-time requirements of MUBF. Argos employs a novel, completely distributed, beamforming technique, as well as an internal calibration procedure to enable implicit beamforming with channel estimation cost independent of the number of base station antennas. We report an Argos prototype with 64 antennas and capable of serving 15 clients simultaneously. We experimentally demonstrate that by scaling from 1 to 64 antennas the prototype can achieve up to 6.7 fold capacity gains while using a mere 1/64th of the transmission power.

Categories and Subject Descriptors

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

Keywords

Large-scale Antenna Systems (LSAS), Many-Antenna, Massive MIMO, Multi-User MIMO, Beamforming, Conjugate, MRT, Zero-forcing

1. INTRODUCTION

Due to the popularization of smartphones, tablets and data-hungry applications, mobile data traffic is growing exponentially, with the expectation that it will increase 18-fold within 5 years [7]. In response, wireless operators are scrambling to acquire more spectrum resources and deploy more base stations to increase spatial reuse. However, there is a fundamental spectrum efficiency limit to existing cellular network architectures: they are *single-user* systems. That is, a base station serves only one terminal in a given resource block (i.e., time slot, spectrum channel, or code sequence). Information theory shows that this limit can be overcome through multi-user multiple-input multiple-output (MU-MIMO) [9]; one promising form of MU-MIMO is called multi-user beamforming (MUBF). With MUBF, a base station employs multiple antennas to send independent data streams to multiple terminals in the same resource block, effectively improving spatial reuse. As theory shows, the more antennas a base station has, the more terminals it can serve simultaneously, resulting in higher spectral capacity. Not surprisingly, the theory community is envisioning MUBF base stations with hundreds of antennas.

However, building a MUBF base station with many antennas is non-trivial. Scaling up baseband processing, clock distribution, transmission synchronization, and channel estimation raises serious system challenges. As a result, only testbeds with a few antennas have been reported in the literature, e.g., [5, 14]. Emerging wireless standards are similarly restricted to a small number of antennas and terminals. The key question to the proposal of MUBF base stations with many antennas remains: *is it practical at all*?

In this work, we answer this question affirmatively with *Argos*[1], a flexible base station architecture that is scalable up to thousands of antennas and able to serve tens of terminals simultaneously through MUBF. Using commercial off-the-self radio modules, i.e., the WARP platform [4], we have realized an Argos prototype with 64 antennas that is capable of serving 15 terminals through zero-forcing and conjugate MUBF. Extensive experimental characterization using this prototype shows that the spectral capacity increases from 12.7 bps/Hz when using a single-antenna to 85 bps/Hz for Argos employing zero-forcing MUBF, and to 38 bps/Hz for Argos employing the less computationally intensive conjugate MUBF, while using a mere 1/64th of the single-antenna transmission power. We show that the spectral capacity grows nearly linearly with the number of base station antennas and the number of simultaneously served terminals, as suggested by theory. The scale of our prototype and experimentation are only limited by the number of WARP boards that are available to us. To the best of our knowledge, Argos is the first publicly reported many-antenna MUBF base station design and realization ($M \gg 10$). Our work demonstrates the feasibility of the MUBF theory community's proposal, and presents key design principles for a scalable, flexible, and cost-effective realization.

Argos achieves its scalability and flexibility with four novel design principles. (i) First, Argos adopts a hierarchical and modular design. This allows it to scale up easily by incrementally adding modules, e.g., WARP boards in the

MobiCom'12, August 22–26, 2012, Istanbul, Turkey.

[1]Argos is a giant with 100 eyes in Greek mythology. The great vision of Argos is analogous to the improved capacity of our many-antenna base station.

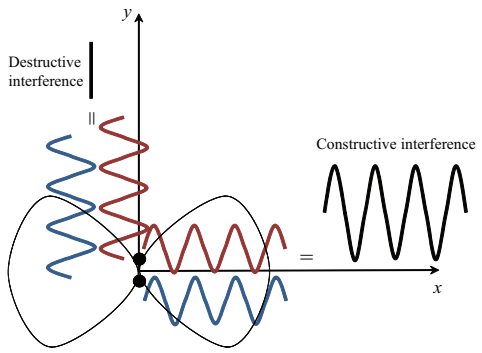

Figure 1: Aerial view of two antennas, represented by two dark dots, emitting identical sine waves at the same frequency. The two waves perfectly reinforce each other along the x axis (constructive interference) but completely cancel each other out along the y axis (destructive interference). Between the two axes the interference gradually varies, producing conical radiation know as a beam pattern.

reported prototype. As Argos scales up it can select the optimal beamforming algorithm by thoroughly analyzing the performance factors and data dependencies of various MUBF techniques. (*ii*) Second, Argos intelligently partitions computation tasks among the different modules in the hierarchy. In the downlink, data to multiple terminals is broadcast to all antennas. Each antenna locally applies its beamforming weights and transmits the combined signal to all terminals simultaneously. In the uplink, I and Q samples from each antenna are combined in upstream modules along the hierarchy. (*iii*) For very large scale operation, Argos leverages a modified version of conjugate beamforming that allows localized weight computation at each antenna. Specifically, traditional conjugate beamforming requires centralized transmission power normalization, while Argos conducts the normalization *locally* at each antenna. This modification allows Argos to scale almost indefinitely with regard to baseband complexity. (*iv*) Finally, Argos employs a novel *internal* calibration procedure that allows implicit beamforming across a large number of base station antennas without explicit channel state information (CSI) estimation, enabling the real-time CSI estimation overhead to become independent of the number of base station antennas. Notably, implicit beamforming requires time division duplex (TDD) operation, which is a substantial modification to the frequency division duplex (FDD) systems primarily used in cellular networks currently.

In summary, we make the following contributions to advance the state of the art of MUBF with many antennas:

- We design and realize Argos, a first-of-its-kind base station architecture that can scale up to thousands of antennas serving tens of terminals with either conjugate or zero-forcing MUBF. We report an Argos prototype with 64 antennas simultaneously serving 15 terminals;
- Using the Argos prototype, we experimentally demonstrate the real-world feasibility of base stations employing many-antenna MUBF and their capability to significantly improve capacity;
- The design of Argos contributes multiple novel techniques to address key challenges toward realizing base stations with a large number of antennas, including clock distribution, transmission synchronization, localized weight computation, and channel calibration.

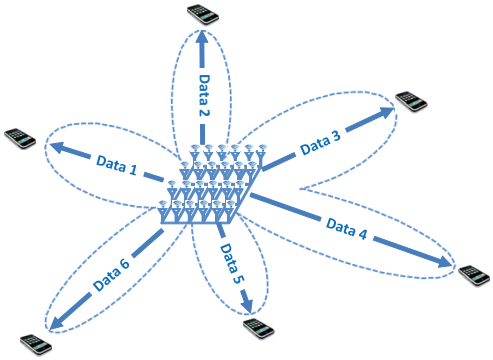

Figure 2: Multi-user beamforming employs baseband precoding and multiple antennas to send independent data streams to multiple terminals at the same time.

In the rest of this paper, we provide the background in Section 2. We present the design and implementation of Argos in Sections 3 and 4, respectively. In Section 5 we evaluate the real-world performance of Argos. In Sections 6 and 7 we discuss related and future work, respectively, and then conclude in Section 8.

2. BACKGROUND

We first provide some background on multi-user beamforming (MUBF) and highlight the key benefits and challenges of using a large number of antennas on base stations.

2.1 Beamforming Basics

Beamforming utilizes multiple antennas transmitting at the same frequency to realize directional transmission. Due to constructive and destructive interference of signals from multiple transmission antennas, the signal strength received at different directions varies spatially, leading to a *beam pattern*, as shown in Figure 1. This beam pattern can be altered by changing the beamforming *weights* applied to each antenna, effectively altering the amplitude and phase of the signal sent from that antenna. *Closed-loop* beamforming employs CSI to calculate the beamforming weights that maximize the signal strength at intended receivers and minimize the interference at unintended ones.

2.2 Single and Multi-user Beamforming

There are two major categories of closed-loop beamforming: *Single-user beamforming* (SUBF) and *Multi-user beamforming* (MUBF). SUBF maximizes the signal strength at a single intended receiver by using beamforming weights that are the complex conjugate of the CSI, which is also known as maximum ratio transmission [15]. MUBF concurrently transmits multiple data streams, each to a different intended receiver as shown in Figure 2. Not surprisingly, information theoretical studies have shown that MUBF can improve spectral capacity manyfold due to its spatial multiplexing gain.

There are many baseband techniques to realize MUBF. We focus on *linear precoding* since other methods are computationally infeasible for practical systems. Let $\mathbf{s}$ denote

a $K \times 1$ vector representing the data-bearing symbols to K users. Linear precoding creates a transmission vector $\mathbf{s}'$ for M antennas, by multiplying the original data vector $\mathbf{s}$ by a $M \times K$ matrix $\mathbf{W}$: $\mathbf{s}' = \mathbf{W} \cdot \mathbf{s}$. Where $\mathbf{W}$ consists of the beamforming weights.

In this work, we study two important forms of linear-precoding for MUBF: *conjugate* and *zero-forcing*. Let $\mathbf{H}$ denote the $M \times K$ channel matrix between the M base station antennas and K concurrent terminals. Let c denote a constant chosen to satisfy a transmission power constraint.

Conjugate: $\mathbf{W} = \mathbf{W}_{conj} = c \cdot \mathbf{H}^*$, where $\mathbf{H}^*$ is the complex conjugate of $\mathbf{H}$. In other words, conjugate beamforming simply takes the complex conjugate of each channel coefficient in $\mathbf{H}$ as the beamforming weight, normalized by c. Indeed, it can be viewed as simultaneous single-user beamforming to K terminals by aggregating the signals intended for these terminals. Conjugate MUBF is sub-optimal and may not perform well with a small M due to inter-terminal interference. This method has only been recently proposed for MUBF with a large number of antennas in [17].

Zero-forcing: $\mathbf{W} = c \cdot \mathbf{W}_{zf} = \mathbf{H}^* \left(\mathbf{H}^{\mathrm{T}}\mathbf{H}^*\right)^{-1}$. Zero-forcing beamforming employs the CSI to precode the data-bearing symbols so that they sum to zero, or a 'null', at unintended receivers. The effectiveness of zero-forcing has been experimentally demonstrated recently [5] with a small number of antennas (four) and terminals (four). Zero-forcing MUBF can keep inter-terminal interference to zero if $K \leq M$. However, due to the required matrix inversion the computational overhead quickly becomes infeasible for real-time applications, as will be discussed in Section 2.4.2.

2.3 Benefits of Many-Antenna MUBF

It is well known in information theory that MUBF with many antennas provide the following key benefits:

First, MUBF can greatly improve spectral capacity through spatial reuse. Roughly speaking, the spectral capacity gain from MUBF is $\min(M, K)$ [9]. A large M allows the base station to serve more terminals concurrently and therefore achieve higher spectral capacity.

Second, a very large M allows a more power-efficient and cost-effective base station. The directional gain from using a large M can be used to compensate for reduced transmission power; that is, a base station can achieve the same capacity with a much lower total transmission power. Under all conceivable propagation conditions doubling the number of base station antennas permits the total radiated power to be reduced by a factor-of-two with no degradation of performance. Only when the number of antennas grows so large that it begins to envelope the terminals or intervening scatterers will this effect cease. Moreover, multi-user beamforming distributes the total transmission power across M antennas, leading to a much lower transmission power per antenna. The base station can therefore leverage cheaper power amplifiers and simpler RF filters. This eliminates the need for active cooling, further reducing power consumption and total cost.

Finally, since power gains are reciprocal, the preceding benefit also applies to terminals. Specifically, it allows battery-constrained terminals to use much lower transmission power to achieve higher capacity.

In Section 5, we will experimentally demonstrate these benefits using the Argos design.

2.4 Challenges to Many-Antenna MUBF

Realizing the key benefits outlined above is, however, non-trivial. Any implementation of MUBF with many antennas faces fundamental timing constraints imposed by the coherence time of the physical wireless channel. MUBF must collect CSI for each terminal then use it to calculate the beamforming weights within a small fraction of the coherence time. Additionally, the computational complexity of MUBF weight calculation grows with the number of antennas, M, and the number of simultaneously served terminals, K. The Argos design has to address both challenges.

2.4.1 CSI Estimation

Acquisition of CSI fundamentally limits the capacity of MUBF with many antennas. MUBF with M antennas to serve K terminals requires CSI between every base station antenna and terminal, or $M \cdot K$ channels. Importantly, all $M \cdot K$ physical channels must be assessed within a period much shorter than the channel coherence time in order to be useful. The coherence time of a wireless channel depends on how quickly the terminal and environment move. In cellular systems this is typically on the order of a few milliseconds, but can drop below 500 μs [1] with vehicular mobility at or near the terminal. This results in a fundamental tradeoff between the time spent collecting CSI, which dictates how many users can be served simultaneously, and the time allocated to sending beamformed data to those users. This tradeoff is explored theoretically in [16].

Traditionally, CSI is estimated *explicitly*. That is, each base station antenna broadcasts a pilot to the terminal, where the latter then uses this pilot to estimate its channel to each of the base station antennas. In order for this channel estimation to be useful, it has to be fed back to the base station in order to perform downlink beamforming. The reverse of this procedure is then used to find uplink CSI, though feedback is unnecessary for maximum ratio combining at the base station. This method thus requires $O(M+K)$ time to send pilots (one pilot from each base station antenna and terminal) and $O(M \cdot K)$ estimates that need to be sent back over-the-air (M estimates from K terminals). This overhead is unavoidable in frequency division duplex (FDD) systems, since the physical channel is not reciprocal at different frequencies.

In time division duplex (TDD) systems the physical channel is reciprocal, and thus, theoretically, CSI could be estimated *implicitly*. That is, uplink pilots could be used to perform downlink beamforming, reducing channel estimation overhead to $O(K)$ and eliminating the required feedback. This is often called *implicit beamforming*. However, in practice, the uplink and downlink channels consist of not only the physical channel, but also the channels introduced by the active RF components in the transmit and receive hardware, as will be further discussed in Section 3.3.

2.4.2 Real-time Beamforming Weight Calculation

The computational complexity of MUBF weight calculation also grows with the number of base station antennas, M, and the number of terminals, K. For conjugate MUBF, the beam weight computation is trivial. In hardware, taking the complex conjugate of a signal only needs a bit-flip and an adder. Therefore, the delay introduced by weight calculation is negligible. However, zero-forcing requires the computation of a matrix inverse, a calculation with the com-

plexity of $O(M \cdot K^2)$. Moreover, the inverse algorithm has internal data-dependencies that limit its ability to be parallelized. While the incurred latency is acceptable at small scales, the polynomial time nature of the inverse makes it very challenging for MUBF systems with a large number of antennas. For example, we estimate that a *single* 15 by 15 matrix inverse would require approximately 150 μs using a specialized high performance FPGA implementation reported in [8]. 150 μs is already 30% of the 500 μs coherence time specified by the LTE channel model. Moreover, in a wideband system such as LTE, this inversion has to be performed for every 14 subcarriers [17]. Thus, while these computations may be pipelined, the true overall inversion time incurred will be far greater than 150 μs.

Additionally, existing beamforming techniques incur a high data transmission overhead because the channel estimates and beam weights have to be exchanged between a central controller and each of the antennas. Even using state-of-the art hardware, e.g., InfiniBand, such exchange incurs a sizable latency cost. The fastest InfiniBand bus has 1 μs overhead per hop and 40 Gbps transfer rate [3]; it will incur approximately 5 μs delay per subcarrier group in a 15 by 15 system. For a 20 MHz bandwidth this amounts to over a 700 μs delay. Zero-forcing cannot avoid this data exchange because the inverse calculation requires the full CSI matrix, **H**. More subtly, even the simplest beamforming algorithm, conjugate, requires full knowledge of **H** in order to appropriately scale the power of the steering weights. In Section 3.4, we present a novel localized conjugate beamforming method that eliminates the overhead due to data transfer between the central controller and antennas.

3. DESIGN

The key question we ask in this section is: *how do we design a MUBF base station that can flexibly optimize its architecture over a wide range of M and K?* Before proceeding to answer it, we want to highlight its practical interest: realistic wireless networks often have large variations in many of their properties, including the financial budget for the base stations, the terminal population within the coverage, and the data traffic volume from terminals. Traditional base stations can only scale their transmission power or, equivalently, their cell size, to cope with such variations. In contrast, Argos base stations can also scale the number of antennas to accommodate various deployment needs.

We argue that in order to meet these demands our many-antenna base station must: (*i*) be economically affordable with cost proportional with M, (*ii*) scale as both M and K become very large, and (*iii*) select the optimal beamforming technique given deployment requirements. We next present how our design of Argos accomplishes these attributes.

3.1 Scalability

The first question is: *can MUBF scale up with M, the number of base station antennas?* MUBF entails three distinct phases: channel estimation, weight calculation, and linear precoding. We explore the feasibility and design implications of these as M scales up.

3.1.1 Channel Estimation

Explicit channel estimation *does not* scale well with M or K. As discussed in Section 2.4.1, explicit channel estimation typically requires $M + K$ pilots to be sent, and $M \cdot K$ estimates to be fed back to the base station. This is clearly an unacceptable overhead for large-scale systems, and suggests that Argos *must* employ TDD reciprocity and implicit beamforming to reduce this overhead to K pilots and eliminate the feedback. In order to enable this, however, we must first overcome the asymmetries introduced by the RF hardware. To accomplish this we devise a novel *internal* calibration scheme, which we present in Section 3.3.

3.1.2 Beamforming Methods

Unfortunately, existing beamforming methods are distinctly unscalable, as they all have centralized data requirements and typically have polynomial time complexity, as discussed in Section 2.4.2.

In light of this, we propose a novel beamforming method that allows weights to be computed completely locally, at each base station radio, as described in Section 3.4. Leveraging this method allows additional radios to be added *without* requiring additional bandwidth, enabling Argos to easily scale up to an unprecedented number of base station antennas, e.g., 1000s.

However, while this beamforming method performs well with a very large number, e.g., 100s, of base station antennas serving 10s of terminals simultaneously, it is well known to be sub-optimal for smaller scale systems, e.g., $M = 30, K = 10$. We demonstrate this empirically in our results, Section 5, where we find that zero-forcing results in up to a 4 fold capacity increase over our method. However, this does not account for the data transport and computational overhead of zero-forcing, which becomes prohibitive with a large number of users or high mobility, as described in Section 2.4.2. Thus we conclude that in order to scale optimally Argos must support traditional, centralized beamforming techniques for smaller scale deployments.

3.1.3 Linear Precoding

Linear precoding requires each antenna to transmit a data stream that is the linear combination of K data streams with K beamforming weights. One design option is to apply these weights centrally. Since each antenna transmits a distinct data stream, this would require the central controller to deliver M I and Q sample streams to each of the individual radios. This approach, obviously, does not scale well, since it requires the central controller to have an output bandwidth proportional to M. As M increases to hundreds or even thousands, this becomes exorbitantly expensive and eventually intractable. Thus we conclude that in any efficient scalable design, the beamforming weights should be applied at the radio. This design choice conveniently allows all of the radios to share a common databus for downlink transmission. In contrast, for uplink transmission, the radio leverages the same linear precoding to apply K beamforming weights to the incoming I and Q samples. Since each radio has unique weights, this again results in M unique data streams (that are K wide)! Fortunately, linear precoding requires these streams to simply be added together; conveniently, this can be done anytime two streams merge in the architecture, thus, again, enabling a constant bandwidth databus. Indeed, we see that with careful design decisions linear precoding *can* scale up with constant data rate requirements. Notably, there is still a need for some form of central controller to demodulate the data once it has been

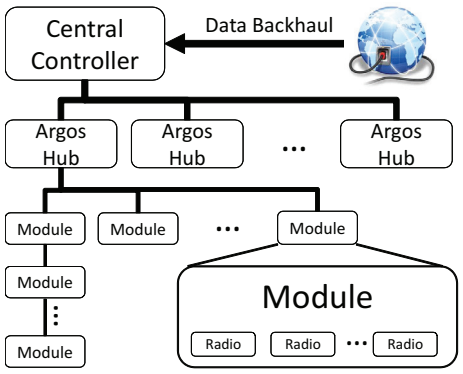

Figure 3: Argos architecture: fat tree structure with daisy-chained leaf nodes.

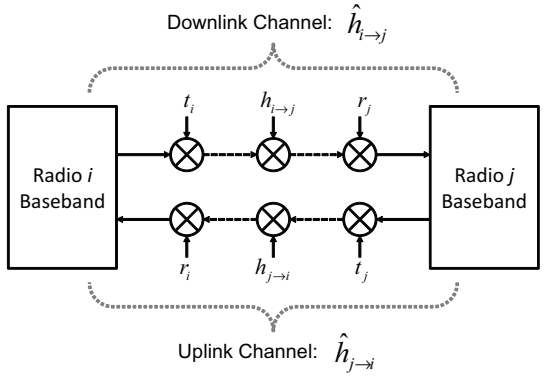

Figure 4: Real channels are not reciprocal due to the differences in transmit and receive hardware. Note that channel reciprocity indicates that within the channel coherence time the physical channel *is* reciprocal: $h_{i\to j} = h_{j\to i}$.

completely recombined; however this operation is latency insensitive, and computationally trivial.

Thus we find that, yes, MUBF *can* scale up with M, but only with careful design choices and new methods for weight calculation and channel estimation.

3.2 Architecture and Topology

The design choices to enable scalability presented above result in two distinct components: (i) a central controller, which handles modulation and demodulation, and (ii) the M radio front-ends that locally calculate beam weights and apply linear precoding. The immediate question we need to answer is: how do we interconnect the controller and the radios? On one hand, we can connect all the radios directly to the controller. This requires the controller to have at least M ports. Since M can be dynamic and very large, this would be a unscalable and inefficient design choice. On the other hand, we can daisy-chain all the radios serially. While scalability seems to be maximized, reliability and delay of the system are severely compromised.

Our solution is to add hierarchies to the base station to improve flexibility, and simultaneously achieve a balance between scalability, reliability, and delay. But, what type of hierarchical structure should we adopt? First we note that deploying M separate radios and antennas would be unwieldy, and cost ineffective to manufacture; thus we create our first level hierarchy: a module that contains one or more radio front-ends. Next, in order to achieve flexible, cost-effective, scaling we allow these modules to be connected serially; enabling additional modules to be added atomically with low overhead. Finally, in order to increase reliability and reduce end-to-end latency, we introduce the Argos hub, which allows multiple modules to be connected in parallel. Figure 3 depicts the Argos architecture.

The Argos base station enables unprecedented scalability and deployability, while fulfilling performance and cost constraints. This architecture enables the Argos base station to scale in three directions: by adding more Argos hubs, by increasing the length of the module chains, and by increasing the number of antennas on a module. The hierarchal architecture facilitates deployments of base stations with many antennas to be flexibly distributed geographically by using a single link to an Argos hub, as well as deployments of base stations with a small number of antennas where the hub can be omitted completely, and modules are simply chained together in series. Additionally, if chains become too long to meet latency requirements, Argos hub can simply be added to parallelize connections and reduce latency.

3.3 Channel Calibration

We devise a novel, completely internal, calibration procedure to enable implicit beamforming on many-antenna base stations through TDD channel reciprocity in order to collect CSI data in constant time with respect to M.

For an M antenna base station to multi-user beamform to K terminals, the base station must acquire the downlink channel state information, $\hat{h}_{m\to k}$, for all $m = 1, 2, ..., M$ and $k = 1, 2, ..., K$. The key challenge is to estimate the effective downlink CSI $\hat{h}_{m\to k}$ from the uplink CSI, $\hat{h}_{k\to m}$, acquired from the uplink pilot signals. However, as shown by Figure 4, the uplink and downlink channels are *not* reciprocal due to the random phase and amplitude differences in the RF hardware. This is caused by a combination of dynamic effects from internal clocking structures, such as dividers, multipliers, and PLLs, as well as static effects from manufacturing deviations. Indeed, we verify that simply resetting a given radio i, or even tuning to a different frequency, randomizes the phase effects of t_i and r_i.

The uplink and downlink channels between any two transceivers is a product of (i) the frequency response of the transmit hardware, (ii) the physical wireless channel, and (iii) the frequency response of the receive hardware:

$$\hat{h}_{i\to j} = t_i \cdot h_{i\to j} \cdot r_j \tag{1}$$

In order to estimate the reciprocal channel, $\hat{h}_{j\to i}$, we define a *calibration coefficient*, $b_{i\to j}$, between radios i and j as:

$$b_{i\to j} = \frac{\hat{h}_{i\to j}}{\hat{h}_{j\to i}} = \frac{t_i \cdot h_{i\to j} \cdot r_j}{r_i \cdot h_{j\to i} \cdot t_j} = \frac{t_i \cdot r_j}{r_i \cdot t_j} = \frac{1}{b_{j\to i}} \tag{2}$$

Notably, if both channels are measured within the coherence time then $h_{j\to i} = h_{i\to j}$ due to physical channel reciprocity. Clearly, if we know the calibration coefficient between two radios and one channel estimate, we can find the reciprocal channel:

$$\hat{h}_{i\to j} = b_{i\to j} \cdot \hat{h}_{j\to i} \quad \text{or} \quad \hat{h}_{j\to i} = \frac{\hat{h}_{i\to j}}{b_{i\to j}} \tag{3}$$

Now let's apply this to our scenario where we would like to estimate the downlink CSI from base station antenna m to

terminal k ($\hat{h}_{m\to k}$) from the uplink CSI ($\hat{h}_{k\to m}$). To do this we must know the M calibration coefficients between each base station antenna and the terminal, that is, all $b_{m\to k}$. These would be impractical to find in a real-system, as estimating $b_{m\to k}$ requires pilots to be sent between every base station antenna and terminal pair, as well as feedback from each terminal. Moreover, unless the terminal and base station share clocks, which is impossible in a wireless system, their hardware transmit and receive channels drift relatively over time, thus requiring this calibration to happen frequently. This approach would be counter-productive, since estimating $b_{m\to k}$ requires downlink pilots, which could be used to directly estimate $\hat{h}_{m\to k}$.

3.3.1 Internal Calibration

We find that it is possible to *internally* calibrate the base station relative to one of it's antennas, e.g., antenna 1. That is, we find all calibration coefficients $b_{m\to 1}$ (for $m = 2, 3, ... M$) using Equation 2. Note that these coefficients are in fact stable over long periods of time, as we show in Section 5.4, since all base station antennas share clocks. We also find that if we know the calibration coefficient between any two radios and a reference radio, then we can derive the direct calibration coefficient between them:

$$\frac{b_{i\to j}}{b_{i\to y}} = \frac{\frac{t_i \cdot r_j}{r_i \cdot t_j}}{\frac{t_i \cdot r_y}{r_i \cdot t_y}} = \frac{t_y \cdot r_j}{r_y \cdot t_j} = b_{y\to j} \tag{4}$$

Thus if we know the calibration coefficient between our reference antenna and terminal k, $b_{1\to k}$, we can find the downlink CSI:

$$\hat{h}_{k\to m} \cdot \frac{b_{1\to k}}{b_{1\to m}} = \hat{h}_{k\to m} \cdot b_{m\to k} = \hat{h}_{m\to k} \tag{5}$$

This suggests that full CSI can be found by simply sending one pilot from each of the terminals, then just one pilot from the base station's reference antenna! Unfortunately, however, to find $b_{1\to k}$ we must feedback the reference antenna's downlink channel estimate, $\hat{h}_{1\to k}$, from each of the k terminals. This significantly reduces the channel capacity, and quickly becomes infeasible for even a moderate K.

3.3.2 Key Idea: Relative Calibration

Our key idea in solving the calibration problem is that an *absolutely* accurate estimation of downlink CSI, $\hat{h}_{m\to k}$, is unnecessary. For all multi-user beamforming techniques using linear precoding, it is sufficient for beamforming antennas to have a *relatively* accurate estimation. That is, as long as each base station antenna's CSI estimation deviates from the real CSI by the same multiplicative factor, multi-user beamforming will still result in the same beam pattern. To visualize this, refer back to Figure 1; if both antennas were to experience the same phase offset, the resulting spatial beam pattern would remain the same. Thus, we can assume $b_{1\to k} = 1$:

$$\hat{h}_{m\to k} = \hat{h}_{k\to m} \cdot \frac{b_{1\to k}}{b_{1\to m}} \quad \Rightarrow \quad \hat{h}'_{m\to k} = \frac{\hat{h}_{k\to m}}{b_{1\to m}} = \hat{h}_{k\to m} \cdot b_{m\to 1} \tag{6}$$

This means that we estimate *relative* downlink CSI, $\hat{h}'_{m\to k}$, by using *only* uplink pilots, without any feedback! To recapitulate, the entire CSI collection process involves 4 steps:

1. Find all internal calibration coefficients, $b_{1\to m}$, offline by sending pilots to and from every base station antenna m and reference antenna 1.
2. Send K orthogonal pilots from each terminal and determine $\hat{h}_{k\to m}$.
3. Derive all $\hat{h}'_{m\to k}$ from 6.
4. Use $\hat{h}'_{m\to k}$ to calculate the beam weights, then send the beamformed data.

Using this process we can efficiently collect full channel state information at the base station by sending only K terminal pilots, without any feedback from the terminals. This enables us to scale M up without any additional channel estimation overhead, which is a critical feature in order to realize a MUBF system with many antennas.

Note that the measurements of downlink and uplink have to be done within the channel coherence time in order for $h_{m\to 1} = h_{1\to m}$. Since base station antennas do not move, the channel coherence time is much larger than typical base station to terminal coherence times. However, as we show in Section 4.5, this calibration can easily be done well within even highly mobile timing constraints; our prototype completes a single antenna pair calibration within 300 μs.

3.4 Decentralized Beamforming

In order to achieve scalable real-time beamforming weight calculation, Argos employs a novel method that allows weights to be calculated locally at each antenna, and therefore avoid the unscalable data-transport overhead required by existing beamforming techniques. As discussed in Section 2.4.2, to perform traditional conjugate beamforming, the weights must be globally normalized so that no base station radio exceeds its maximum power output. For example, assuming a maximum radio transmit amplitude of 1, and in order to ensure at least one radio transmits at maximum power:

$$c = \left(\max \left(\sum_{k=1}^{K} \|\hat{\mathbf{h}}_{m\to k}\| \right) \right)^{-1} (m = 1, 2, ...M) \tag{7}$$

where c is the scaling factor used in the beamforming weight calculation ($\mathbf{W} = c \cdot \mathbf{H}^*$). Global power scaling is characterized by using a single constant to scale all of the weights. This global scaling is necessary to maintain the ratio between each base station antenna's weight for a given terminal, which ensures per-terminal transmission energy optimality, as proven in [15]. However, each base station antenna must know either c (or $\mathbf{H}$) to properly scale its own beamforming weights. This requires full CSI to be transferred from each module to the central controller, nullifying the benefit from the aforementioned decentralization. To tackle this, we propose a local power scaling approach that closely approximates global normalization.

Argos leverages a key observation that for the different terminals in multi-user beamforming, *the channels corresponding to different terminals are uncorrelated and experience independent fading.* Therefore, statistically speaking, when the number of terminals is large, the actual transmission power at each antenna is very similar. Our solution simply normalizes the total transmission power locally at each base

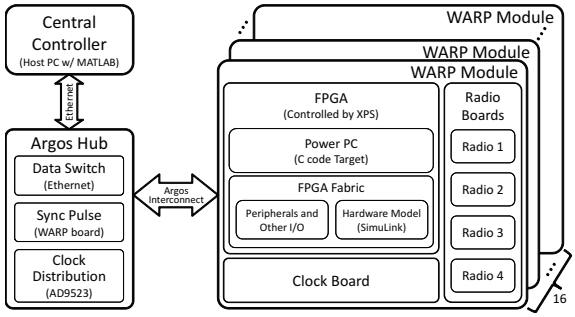

Figure 5: The implementation of Argos using WARP boards, a laptop, an ethernet switch, and an AD9523 based clock distribution board.

station antenna using only the CSI it measures:

$$c_m = \left(\sum_{k=1}^{K} \|\hat{\mathbf{h}}_{m \to k}\| \right)^{-1} \quad (m = 1, 2, ...M) \tag{8}$$

The conjugate beamforming weights are then scaled via:

$$\mathbf{W} = \mathbf{H}^* \cdot \text{diag}(C) \tag{9}$$

Where C is the scaling vector given by $C_{local} = [c_1, c_2, ...c_M]$, from Equation 8; notably the globally scaled conjugate can also be found in this form, using $C_{global} = [c, c, c, c...]$, from Equation 7.

We have experimentally verified the effectiveness of such local power scaling and observed that its performance is almost indistinguishable from the optimal global power scaling method (see Section 5), using equal transmit power for both methods. Moreover, in real deployments, since local power scaling ensures that each radio can utilize its full hardware power capacity, it can *always* achieve equal or greater SNR than global power scaling (since it can send with greater total transmit power), as proven in [22, p. 24]. Notably, if terminals are not approximately equidistant from the base station, then per-terminal power scaling is required to ensure fairness (preventing terminals closer to the base station from being allocated all of the transmission power), but this can be done at a much coarser time scale (i.e., seconds), thus not creating additional overhead or affecting performance.

4. IMPLEMENTATION

In this section we provide a detailed report of our implementation of Argos, which leverages WARP [4], commercially available clock distribution boards, a commodity PC, and an ethernet switch. Figure 5 shows an abstract representation of our implementation. As the first proof-of-concept prototype, our system includes a central controller, an Argos hub and 16 modules, each with 4 radios. The central controller consists of a single host PC, which uses MATLAB to send data, weights, and control commands to the radio modules. The Argos hub is comprised of a 24-port ethernet switch, a clock distribution board, and a WARP board, which uses its GPIO pins to provide transmission synchronization splitting/replication. Due to the limited availability of WARP boards, this board also serves as a radio module, however these roles are functionally separate, and in future generations of the Argos prototype they will be physically separated as well. Each radio module is a single WARP board with 4 radio daughter cards and 4 antennas. Figure 6 depicts the real system: the base station includes 16 WARP boards with 64 antennas that are compactly placed on a custom rack-mount platform. We note that the number of terminals supported by each module is fundamentally limited by its hardware capabilities. In the WARP boards we are using, this bottleneck is the number of multipliers (328 on the Virtex 2 Pro xc2vp70) [26]. We are able to use 240 of these multipliers to provide linear precoding for 15 terminals on the 4 antennas, which requires 60 complex multipliers. The remaining multipliers are used by other functions, and 4 are unusable due to routing constraints. However, the recently released Virtex 7 supports up to 3600 multipliers clocked at a rate of 741 MHz; with multiplexing this would enable 16,672 complex multiplies per 40 MHz sample (neglecting routing overhead and other functions that require multipliers), which would, obviously, alleviate this bottleneck [25].

To the best of our knowledge, our Argos prototype is the first publicly reported many-antenna MUBF system with real-world feasibility. We next elaborate on our implementation.

4.1 Hardware and Software Platform

WARP is a scalable and programmable wireless platform for prototyping advanced wireless systems. Each WARP board allows up to four radio daughter cards to be connected and therefore can contribute up to four active antennas simultaneously to Argos. Each radio board includes a Maxim 2829 transceiver chip [18], which operates at the 2.4 or 5 GHz ISM bands with a 20 MHz bandwidth. WARP conveniently provides a MATLAB-based framework, WARPLab, which allows MATLAB to control the WARP boards and process the transmit and receive data samples. As shown in Figure 5, WARPLab consists of four layers: (*i*) The underlying Simulink model that implements the custom hardware for controlling the FPGA board and radio boards; (*ii*) The Xilinx Platform Studio (XPS) project that integrates and connects all of the hardware components, including the Simulink model, the I/O cores for the serial port, Ethernet port, clocking, etc.; (*iii*) The C code that runs on the PowerPC microprocessor, controls the hardware through memory mapped I/O, and acts as an interface to the Ethernet port; (*iv*) The MATLAB interface that configures the boards, generates the transmit samples, and processes the receive samples.

We have extensively customized the WARPLab framework to enable hardware MUBF, transmission synchronization, clock synchronization, and indirect calibration among base station antennas. These functionalities are essential to for Argos to enable MUBF with many antennas.

4.2 Hardware MUBF

A straightforward, and much easier approach to realize MUBF in WARPLab is to implement it in software within the MATLAB interface; this, in fact, was our first implementation. In this approach the beamformed baseband signal can be directly delivered to the WARP boards without the need of linear-precoding in hardware. However, this method introduces major latency between the CSI collection and data transmission, which increases linearly with the number of base station antennas, and severely degrades performance. This is a result of the same scaling problem discussed in Section 3.1. Therefore, we modified the WARPLab hardware to enable hardware MUBF.

At each base station antenna, applying the beamform-

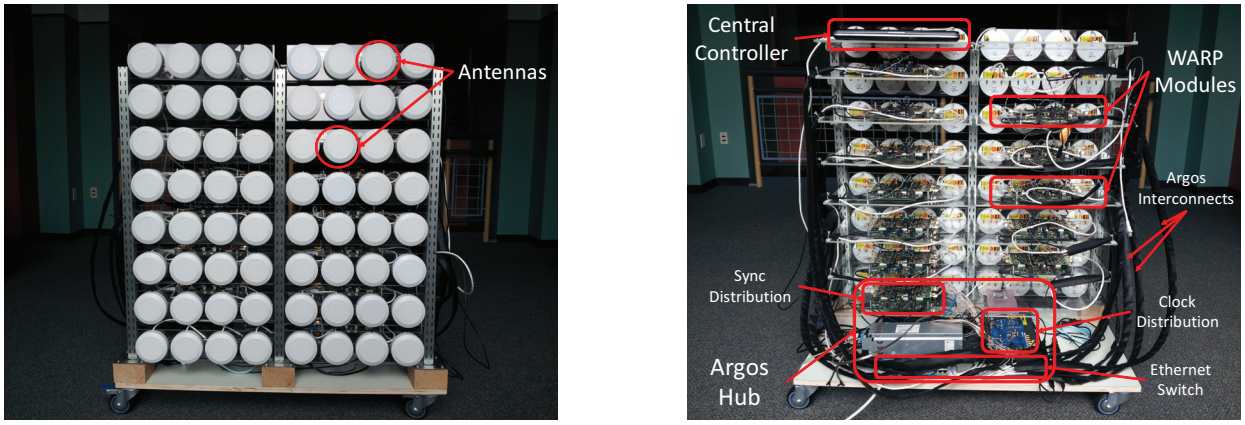

Figure 6: The prototype of Argos with 16 modules and 64 antennas. Left: front side. Right: back side.

ing weights consists of multiplying the baseband symbol intended for each terminal, s_k, by its corresponding beamforming weight, w_k, and then adding them together: $s'_m = \sum_{k=1}^{K} w_k \cdot s_k$ where s'_m is the resultant beamformed signal transmitted by antenna m. Multiplying the signal by a complex number is equivalent to rotating the phase and scaling the amplitude. In hardware, this requires K registers and K parallel complex multipliers (each complex multiplier needs 4 multipliers and 2 adders) in series with 2 K input adders. We store the beamforming weights, $w_k(k = 1, 2, ...K)$, in memory mapped registers. This is important since it enables the PowerPC, and in turn, the MATLAB interface to directly control them.

4.3 Transmission Synchronization

WARPLab has a default function to enable transmission synchronization between multiple WARP boards. It is achieved by using the built-in API command "sendsync()" in the MATLAB interface. However, due to the jitter introduced by the ethernet stack, switch, and cables, such synchronization can lead to a timing offset on the order of 20 samples, depending on the ethernet switch and cable lengths, which makes accurate CSI collection and beamforming impossible.

To address this challenge, we employ a WARP board to distribute the central controller's transmission synchronization signal. As part of the Argos hub, this WARP node leverages directly connected, registered, GPIO to reliably send the sync pulse to the radio modules. Notably, to ensure the modules receive the pulse within 1 clock cycle, the cable lengths should all be within one wavelength, λ. With a channel bandwidth of 20 MHz, λ is 7.5 meters (40 MHz sampling clock), which is a very easy constraint to meet. As stated above, this WARP node serves the dual role of sync distribution and module, thus it "distributes" the sync to itself with an effective cable length of 0. This means the other cables must be less than 7.5 meters, which is not a problem; in our current setup the length is 2 meters. While each board may have a slightly different clock phase, this phase offset is constant (due to the clock synchronization), and explicitly compensated for by the beamforming algorithm.

We have modified the Simulink model, the XPS project, and the C code to enable GPIO-based transmission synchronization. Specifically, we inserted appropriate gateways and registers into the Simulink model, re-mapped the GPIO pins to the appropriate signals in the XPS project, and disabled the traditional ethernet sync in the C code.

4.4 Clock Synchronization

Precise inter-board clock synchronization is critical for Argos, due to its distributed modular architecture. The WARP board requires two reference clocks: a 20 MHz RF clock and a 40 MHz logic/sampling clock. Both clocks can be either forwarded or driven by an external source. In addition, we discovered that the Maxim 2829 transceiver chip on the radio board can use a 40 MHz clock. Therefore, we were able to use a single external source to drive the logic clock, then forward the logic clock to the reference input for the RF clock. This way, inter-board clock synchronization can be achieved in an easily manageable and scalable way.

We leverage a commercial clock distribution evaluation board designed for LTE, the AD9523/PCBZ, to accomplish this. The AD9523 provides 18 clock outputs, which we leverage to drive all of the radio modules. Although we haven't exceeded the capacity of the AD9523, an additional clock distribution board could be connected (as part of an additional Argos hub), which would provide 17 more outputs. Alternatively, the existing modules can forward their clocks to additional modules, through Argos' multi-hop extension.

4.5 Indirect Calibration

For indirect calibration, we need to estimate $b_{m\to 1} = \frac{t_m \cdot r_1}{r_m \cdot t_1}$ for each antenna m with respect to the "reference antenna," as described in Section 3.3. Due to buffer constraints, we implement this in a per-module iterative fashion. First, the module containing the reference antenna calibrates internally; that is, the reference antenna sends a pilot while other antennas on the module listen, then each of those antennas sends a pilot, in turn, while the reference antenna listens. These channel estimates are then reported to the central controller so that the reference antenna's buffer can be overwritten. Next, the reference antenna sends a pilot sequence while all the antennas on another module listen, then each of those antennas transmits a pilot, in turn, while the reference antenna listens. Again, the channel estimates are reported to the central controller. The process is then repeated for each module. The calibration procedure is very latency sensitive, as the physical channel should not change between transmission and reception of pilots for any antenna pair. To address this, we implement the calibration locally on the PowerPC in C code and leverage Argos' transmission synchronization to coordinate the send and receive phases. The resulting calibration happens within 300 μs for each antenna pair, which is well within the channel coherence time.

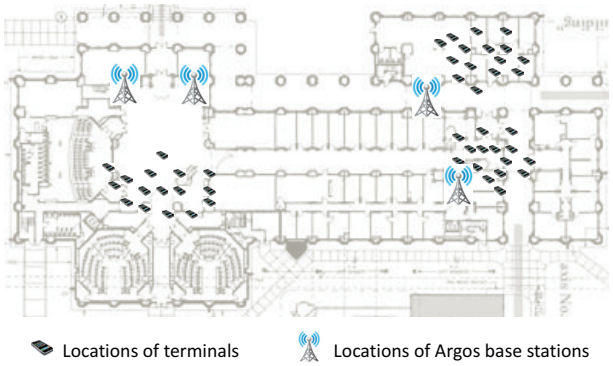

Figure 7: Environments and the locations of the base station and terminals for for the reported experiments. Note that the base station leverages directional antennas in order to serve one sector. Terminals have vertical separation as well, spanning up to three floors.

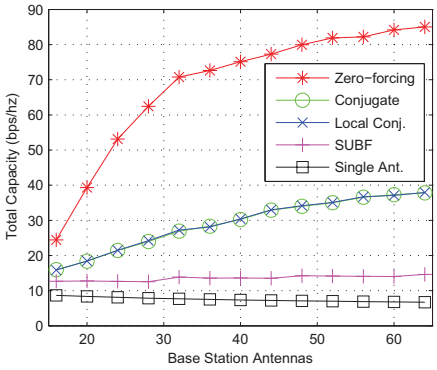

Figure 8: Cell capacity as the number of base station antennas (M) increases from 16 to 64, by 4, serving 15 terminals. In order to compensate for the beamforming gain, total transmission power is $1/M$, implying average power per-antenna for multi-antenna schemes is $1/M^2$.

Another challenge we encountered while performing our indirect calibration approach is the significant amplitude variation for the channels between the reference antenna 1 and other antennas. This is due to the grid-like configuration of our antenna array where different pairs of antennas can have very different antenna spacings. According to our measurement, the SNR difference can be as high as 40 dB, leading to a dilemma for us to properly choose the transmission power for the reference signal. To address this, we isolate the reference antenna from the others, and place it in a position so that its horizontal distance to the other antennas are approximately identical. Such placement of the reference antenna does not affect the calibration performance due to our calibration procedure's isolation of the radio hardware channel from the physical channel.

5. EVALUATION

Leveraging our prototype, we experimentally evaluate the feasibility of Argos in realistic environments. We have the following impressive observation: compared to using a single antenna, Argos can improve spectral capacity over *12 fold* leveraging MUBF with many antennas, using *equal total transmission power*. With 64 antennas and 15 terminals, the spectral capacity can be boosted from 12.7 bps/Hz to 85 bps/Hz (6.7x) for zero-forcing MUBF, and 38 bps/Hz (3x) for conjugate MUBF, while *using a mere* $1/64$*th of the total transmission power*. We find that Argos easily scales from 1 to 64 base station antennas serving 1 to 15 terminals, and that, in general, performance scales linearly with M and K. Finally, we experimentally validate the performance of our localized conjugate beamforming method, as well as our internal calibration procedure.

5.1 Experimental Setup

We employ all 64 antennas at the base station to perform MUBF to 15 concurrent terminals. We use the 2.4 GHz band with a 625 kHz carrier width to avoid frequency fading effects. Since it is relatively easy to move our platform (see Figure 6), we tested various indoor locations (see Figure 7) in order to collect data from diverse environments. There are both LOS and NLOS channels between the base station and terminals. We repeat our experiments multiple times, typically collecting over 3000 measurements at each location, to reliably average out performance.

To obtain the cell capacity, we aggregate the Shannon capacity for each terminal, or $C_{Cell} = \sum_{k=1}^{K} \log(1+SINR_k)$ where $SINR_k$ is the measured SINR at terminal k. We let the base station transmit dummy QPSK-modulated frames to the terminals, which is sufficient to validate the real-world feasibility of Argos since MUBF is a physical layer technique that is orthogonal to the MAC layer and above.

To accurately measure the terminal SINR, we use the RSSI indicator from the Maxim 2829 transceiver on the radio board to report the received signal strength for each transmission, as well as the noise floor after the transmission completes. Since the radio is unable to distinguish signal and interference strength, we slightly stagger the transmission to the intended terminal and that to the unintended terminals. This way we can separately measure the signal power and interference power, and acquire the SINR accordingly. To make sure the channel remains constant during the transmissions we conduct our experiments in an ultra-stable environment, i.e., late at night, without moving people and wireless traffic.

5.2 Improvement of Cell Capacity

The primary purpose of our experiments is to determine the capacity improvement of Argos, in order to ultimately answer the practicality of the many-antenna MUBF base station proposal from the theory community. We report two sets of experiments, which evaluate the scalability with regards to the number of base station antennas, M, and the number of terminals, K, respectively.

5.2.1 Scaling up with M

In the first set of experiments, we vary the number of base station antennas, M, assuming a fixed number of terminals, $K = 15$. Figure 8 shows C_{Cell} as a function of M for a base station with a single antenna (*Single Ant.*), SUBF (*SUBF*), conjugate MUBF (*Conjugate*), our modified localized conjugate MUBF (*Local Conj.*), and zero-forcing MUBF (*Zero-forcing*). In order to compensate for the beamforming gain, total transmission power is scaled by a factor of $1/M$ for all five cases. This enables our experiments to separate the orthogonality gain of scaling up from the well-known,

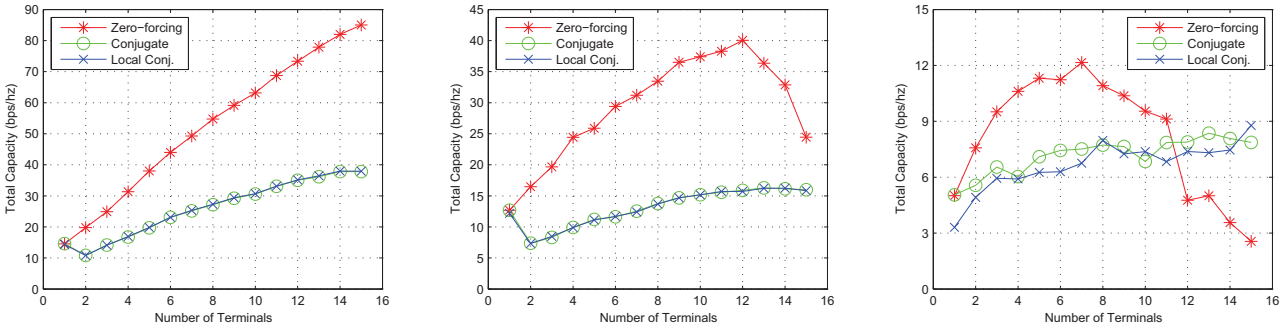

Figure 9: Cell capacity as the number of terminals K increases. Total transmission power is held constant within each plot. Left: $M = 64$; Middle: $M = 15$; Right: $M = 15$ with reduced transmission power.

predictable, beamforming gain. We have the following key observations:

First, when M is much larger than K, both conjugate and zero-forcing MUBF increase the cell capacity as M scales up, *despite reducing the total transmission power proportionally with M*. The beamforming gain from the additional antennas compensates for the power reduction, as demonstrated by the flat performance of SUBF, while simultaneously increasing the natural orthogonality of the terminals. This reduces the inter-terminal interference of conjugate MUBF, and reduces the amount of power wasted to create nulls for zero-forcing MUBF. With $M = 64$ the improvement for conjugate and zero-forcing MUBF over a single antenna are 5.7x and 12.7x for equal power, or 3x and 6.7x for 1/64 power, respectively.

Second, as M drops to K, i.e., $M \approx K = 15$, the performance of zero-forcing drops steeply. This is due to the tightness of the degrees of freedom at the base station; zero-forcing inevitably wastes the majority of transmission power for interference cancelation, leading to a much reduced signal power at the intended terminals. Later, we will show that when $M = K$ this inefficiency can even result in conjugate MUBF out-performing zero-forcing.

5.2.2 *Scaling up with K*

We next fix M and vary the number of terminals to see how capacity scales with K. In the experiments reported by Figures 9 Left and Middle, the total transmission power is scaled by $1/M$, similar to that in Figure 8, and is held constant regardless of K. Because the total power is split among the terminals, the power per terminal is therefore scaled by $1/K$. In the experiment shown by Figure 9 Right, we reduce the transmission power to the minimum WARP setting in order to demonstrate how the capacity of the three forms of MUBF are affected by power. We have the following observations:

First, when $M \gg K$, as shown in Figure 9 Left, capacity increases approximately linearly with the number of terminals for both conjugate and zero-forcing MUBF; this is attributable to the multiplexing gains from simultaneously serving K terminals.

Second, conjugate beamforming initially loses capacity as the number of terminals increases from 1 (SUBF) to 2 due to the addition of interference from the other terminal, and thus the overwhelming drop in SINR. This loss, however, is quickly compensated for by the multiplexing gains.

Third, as K approaches M, the performance of zero-forcing drops sharply as shown in Figure 9 Middle. This corroborates the second observation made in Section 5.2.1. Additionally, the performance of conjugate flattens, and even starts to decline, as the additional interference from more terminals causes the average SINR to approach 0 dB.

Finally, when the transmission power is reduced, conjugate MUBF performs relatively better than zero-forcing, as shown in Figure 9 Right. This is because the performance of conjugate is inherently limited by interference from other terminals, while the performance of zero-forcing is instead limited by noise, since the interference is explicitly canceled. It is not until the transmission power is reduced to a point where interference has the same magnitude as noise that there is a significant effect on the capacity improvement for conjugate.

5.3 Near-optimality of Localized Conjugate

In order to verify the viability of our localized method for conjugate MUBF, we implement it in Argos and compare it to standard conjugate MUBF with global power scaling. As shown in Figure 10, we see that our local power scaling method (*Local Conj.*) results in a signal power within 1.2 dB of global power scaling (*Conjugate*), but quickly approaches equivalent power as the number of terminals increases. For a fair comparison we ensure that both methods send with the same transmission power, however in a practical deployment our method will always transmit equal or more power. While local power scaling is less efficient for a given transmission power, it ensures that each base station radio is being fully utilized, thus more intelligently adapting to the constraints of real-world hardware. Furthermore, we see in Figures 8 and 9 that the performance difference between global power scaling (*Conjugate*) and local power scaling (*Local Conj.*) is almost indistinguishable.

5.4 Stability of Indirect Calibration

As described in the previous section, we implemented a novel reciprocal calibration method to enable implicit beamforming and efficient TDD operation. Figure 11 shows that this calibration deviates from the mean angle an average of less than 2.6% (maximum 6.7%), and from the mean amplitude less than 0.7% (maximum 1.4%), over a period of 4 hours. Notably, these measurements were taken during the day with normal movement around the base station, indicating the calibration procedure is stable in real-world environments. Angle deviation is calculated by difference in angle from average angle over π, i.e., 2.6% error is equivalent to 0.08 radians. This indicates that our internal calibration scheme can performed very infrequently, i.e., once a day, and thus has negligible performance overhead.

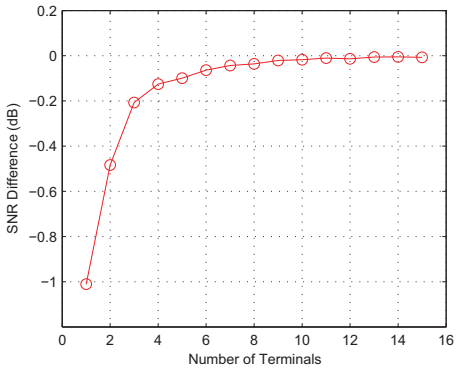

Figure 10: Relative signal power between conjugate and our conjugate with local power scaling, sent at the same transmit power. Local power scaling performs within 1.5 dB of global power scaling, and the difference quickly converges to 0 dB as K increases.

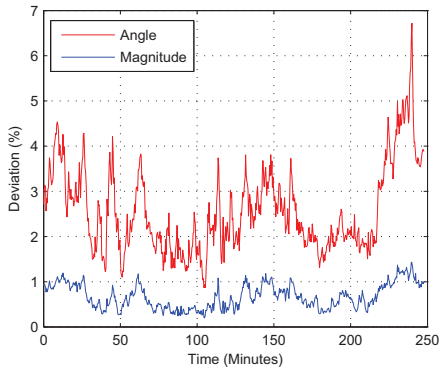

Figure 11: Our calibration procedure exhibits an average instantaneous noise of less than 7% and remains stable indefinitely.

6. RELATED WORK

Argos is directly motivated by multi-user beamforming (MUBF) theory. Recent theoretical works have demonstrated the exciting benefits of large-scale MUBF, or its general form, massive MIMO [12, 13, 20]. In [17], the realization of conjugate beamforming with an infinite number of antennas in a TDD multi-cell system is discussed. Leveraging conjugate beamforming for a distributed architecture is proposed in [20], however our paper explicitly addresses power-normalization, and shows that, in a statistical sense, each base station antenna can perform normalization independently of the others. Single-cell analyses of the performance of both conjugate and zero-forcing beamforming for finite numbers of antennas are obtained in [19]. However, these works are theoretical in nature, and do not address the system and implementation issues such as decentralized power scaling and TDD calibration. Argos is motivated by and built on top of this prior work, and complementarily addresses the architectural and system challenges to realize a many-antenna base station in the real-world.

Endeavors to push MUBF into practice have been observed recently. State-of-the-art wireless standards in cellular networks and WLAN, such as LTE, WiMAX, and 802.11, have all considered incorporating downlink MUBF in their emerging releases, e.g., 3GPP release 9 [1] and 802.11ac [2]. However, MUBF in these standards is restricted to a much smaller scale, e.g., up to eight antennas in 802.11ac. The most recent research efforts towards practical MUBF [5, 14] are limited to a small number of antennas. Argos, comparatively, is the most ambitious MUBF prototype, featuring a much greater number of antennas. As a result, our contribution differs from existing works in that we have identified and addressed a set of unique challenges regarding practicality and scalability. In addition, Argos is programmable and therefore can be used for any large-scale MU-MIMO technique, although we have focused on MUBF in this paper.

There are orthogonal techniques to improve the spatial reuse of spectrum such as sectorization and small cells. First, sectorization uses multiple antennas to form directional beams, each of which covers a range of directions as a *sector*. Terminals in different sectors can be served simultaneously. Therefore, sectorization can be treated as a special case of MUBF to physically separate terminals. Second, small cell techniques, such as femto-cells and pico-cells deploy base stations more compactly, with limited coverage to improve spatial reuse. In each sector or small cell, one can further apply large-scale MUBF to achieve even more spatial reuse and better energy efficiency.

Argos improves spatial reuse efficiency through by employing a large number of antennas on base stations. One can also add antennas at the terminal to further improve spectral efficiency and reduce inter-cell interference. The authors in [21] reported a system with multiple directional antennas on mobile terminals where only one antenna is active at a time to realize uplink directionality. The authors in [27] studied the feasibility of SUBF on terminals, and demonstrated its power efficiency and capacity benefit. Argos is completely complementary to these terminal-based solutions and provides orthogonal benefits.

A few prior works have offered solutions for efficient channel calibration in TDD systems, e.g., [6, 10, 11, 23, 24]. All prior solutions require terminal involvement and feedback in the calibration process, an unacceptable overhead in a large-scale MUBF system. In contrast, the relative calibration in Argos is done internally at the base station without such overhead.

7. DISCUSSION

This paper presents a scalable architecture for many-antenna base stations, a real-world implementation of this architecture, as well as compelling early experimental results. These results motivate further research in the area of many-antenna systems, and raise many practical challenges to be surmounted.

First, the size, cost, and power consumption of a many-antenna base station are significant impediments to adoption. However, we believe that advances in manufacturing combined with the use of specialized hardware in both the analog and digital domains can overcome these barriers.

While Argos supports multiple precoding techniques, an obvious question this work raises is: which precoding technique is optimal under a given scenario? This is a deceivingly hard question to answer, as there are many variables to optimize, such as power, fairness, and spectral efficiency, as well as many factors that impact this optimality, such as transmission power, propagation environment, terminal mobility, hardware capabilities, number of terminals, and number of base station antennas. Moreover, we conjecture that it will be advantageous to dynamically select the precoding technique, as many of these factors continuously change.

Finally, this many-antenna architecture also presents numerous network level challenges, including terminal paging, optimal grouping and scheduling of terminals, and handover between cells. This architecture also raises many opportunities for improving total network capacity; for example, as predicted by [17], the reduced transmit power will significantly improve inter-cell interference. Addressing these challenges and opportunities is the subject of our future work.

8. CONCLUDING REMARKS

We present the design, realization, and evaluation of Argos, a base station architecture that can potentially employ thousands of antennas to serve tens of terminals simultaneously through MUBF. In order to enable this unprecedented scaling in a practical environment Argos employs a hierarchal modular design that facilitates flexible, scalable deployments while simultaneously constraining latency and providing fault tolerance. It also features a novel beamforming algorithm that is completely decentralized and a new calibration method that allows CSI to be collected in constant time with regard to the number of base station antennas.

Our experimental characterization of an Argos prototype with 64-antennas clearly shows the practical benefits of MUBF base stations with many antennas, improving spectral and energy efficiency manyfold simultaneously. Our results are the first publicly reported evidence that many-antenna MIMO systems can produce significant benefits under real-world settings. The scale of our experiments is only limited by the number of Argos modules (WARP boards) currently available to us. The architecture of Argos, however, can easily accommodate many times more modules, each with more radios, potentially allowing thousands of antennas to serve tens of terminals through MUBF.

Acknowlegements

This work was supported in part by NSF grants CRI 0751173, MRI 0923479, NetSE 101283, MRI 1126478 and CNS 1018292. Clayton Shepard was supported by an NDSEG fellowship and grant CNS 1018292. We would like to thank Ashutosh Sabharwal, Edward Knightly, Patrick Murphy, Reinaldo Valenzuela, Cuong Tran, our reviewers, and our shepherd, Sunghyu Choi, for their input and support.

References

[1] 3GPP Release 9. `www.3gpp.org/Release-9`.

[2] IEEE 802.11ac. `mentor.ieee.org`.

[3] InfiniBand. `www.infinibandta.org`.

[4] Rice University Wireless Open Access Research Platform. `warp.rice.edu`.

[5] E. Aryafar, N. Anand, T. Salonidis, and E. Knightly. Design and experimental evaluation of multi-user beamforming in Wireless LANs. In *Proc. ACM MobiCom*, Chicago, Illinois, Sept. 2010.

[6] A. Bourdoux, B. Come, and N. Khaled. Non-reciprocal transceivers in OFDM/SDMA systems: impact and mitigation. In *Proc. IEEE RAWCON*, 2003.

[7] Cisco Inc. Cisco visual networking index: Global mobile data traffic forecast update, 2011-2016. `cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper_c11-520862.html`.

[8] C. Dick, F. Harris, M. Pajic, and D. Vuletic. Implementing a real-time beamformer on an FPGA platform. *Xcell J.*, 2007.

[9] D. Gesbert, M. Kountouris, R.W. Heath, C. Chae, and T. Salzer. Shifting the MIMO paradigm. *IEEE Signal Processing Magazine*, 2007.

[10] M. Guillaud, D.T.M. Slock, and R. Knopp. A practical method for wireless channel reciprocity exploitation through relative calibration. In *Proc. IEEE ISSPA*, 2005.

[11] Y. Hara, Y. Yano, and H. Kubo. Antenna array calibration using frequency selection in OFDMA/TDD systems. In *Proc. IEEE GLOBECOM*, 2008.

[12] J. Hoydis, S. ten Brink, and M. Debbah. Massive MIMO: How many antennas do we need? *arXiv:1107.1709v2 [cs.IT]*, 2011.

[13] H. Huh, G. Caire, H.C. Papadopoulos, and S.A. Ramprashad. Achieving large spectral efficiency with TDD and not-so-many base-station antennas. *IEEE-APS Topical Conf. on APWC*, 2011.

[14] J. Koppenborg, H. Halbauer, S. Saur, and C. Hoek. 3D beamforming trials with an active antenna array. In *Int. Workshop on Smart Antennas*, 2012.

[15] T.K.Y. Lo. Maximum ratio transmission. *IEEE Trans. on Communications*, 1999.

[16] T.L. Marzetta. How much training is required for multiuser MIMO? In *Proc. IEEE ACSSC*, 2006.

[17] T.L. Marzetta. Noncooperative cellular wireless with unlimited numbers of base station antennas. *IEEE Trans. on Wireless Communications*, 2010.

[18] Maxim. Single-/dual-band 802.11a/b/g world-band transceiver ICs. `datasheets.maxim-ic.com/en/ds/MAX2828-MAX2829.pdf`.

[19] H. Ngo, E. Larsson, and T.L. Marzetta. Energy and spectral efficiency of very large multiuser MIMO systems. *IEEE Trans. on Communications*, 2011.

[20] F. Rusek, D. Persson, B. Lau, E. Larsson, T.L. Marzetta, O. Edfors, and F. Tufvesson. Scaling up MIMO: Opportunities and challenges with very large arrays. *arXiv:1201.3210v1 [cs.IT]*, 2011.

[21] A. Amiri Sani, L. Zhong, and A. Sabharwal. Directional antenna diversity for mobile devices: Characterizations and solutions. In *Proc. ACM MobiCom*, 2010.

[22] C. Shepard. Argos: Practical base stations with large-scale multi-user beamforming. Master's thesis, Rice University, April 2012. Available at: `clay.rice.edu/pubs/MasterThesis.pdf`.

[23] J. Shi, Q. Luo, and M. You. An efficient method for enhancing TDD over the air reciprocity calibration. In *Proc. IEEE WCNC*, 2011.

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

[25] Xilinx Inc. 7 Series FPGAs Overview. `xilinx.com/support/documentation/data_sheets/ds180_7Series_Overview.pdf`.

[26] Xilinx Inc. Virtex-II Pro and Virtex-II Pro X Platform FPGAs: Introduction and Overview. `xilinx.com/support/documentation/data_sheets/ds083.pdf`.

[27] H. Yu, L. Zhong, A. Sabharwal, and D. Kao. Beamforming on mobile devices: A first study. In *Proc. ACM MobiCom*, 2011.