RICE UNIVERSITY

Argos: Practical Base Stations for Large-scale Beamforming

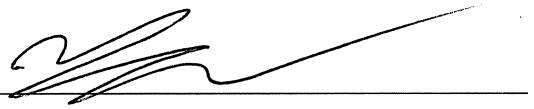
by

Clayton W. Shepard

A Thesis Submitted
in Partial Fulfillment of the
Requirements for the Degree

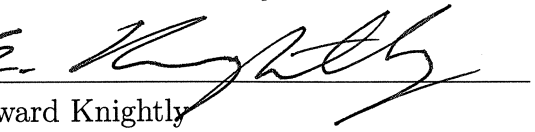
Master of Science

Approved, Thesis Committee:



Lin Zhong, Chair
Associate Professor
Electrical and Computer Engineering

Ashutosh Sabharwal
Associate Professor
Electrical and Computer Engineering



Edward Knightly
Professor
Electrical and Computer Engineering

Houston, Texas

April, 2012

ABSTRACT

Argos: Practical Base Stations for Large-scale Beamforming

by

Clayton W. Shepard

MU-MIMO theory predicts manyfold capacity gains by leveraging many antennas (e.g. $M \gg 10$) on wireless base stations to serve many users simultaneously through multi-user beamforming (MUBF). However, realizing such a large-scale design 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 large-scale base station that is capable of serving many (e.g., 10s of) terminals simultaneously through MUBF. Designed with extreme flexibility and scalability in mind, Argos exploits hierarchical and modular design principles, properly partitions baseband processing, and holistically considers real-time requirements of MUBF. To achieve unprecedented scalability, we devise a novel, completely distributed, beamforming technique, as well as an internal calibration procedure to enable implicit beamforming across large arrays. We implement a prototype with 64 antennas, and demonstrate that it can achieve up to 6.7 fold capacity gains while using a mere 1/64th the transmission power.

Acknowledgments

First and foremost I would like to thank my advisor, Dr. Lin Zhong, for all of his help, support, patience, and advice that made this project, and my graduate career, possible. I am also very grateful for the continued help from my colleagues, Narendra Anand and Hang Yu, who have directly contributed to the development and direction of this project.

I would also like to thank Dr. Richard Yang from Yale University, as well as Dr. Erran Li and the the other researchers at Bell Labs for introducing me to this project and their ongoing support. Dr. Ashutosh Sabharwal and Dr. Thomas Marzetta have been instrumental to my understanding of the relevant theoretical background; I greatly appreciate our incredibly enlightening discussions.

This project would not have been possible without the WARP platform, pioneered by Patrick Murphy. His technical support, as well as that provided by Siddharth Gupta, Gaurav Patel, and Chris Hunter, has been invaluable in prototyping Argos. I am very grateful to Nathan Zuege, Danny Eaton, Levi Ball, Paul Williams, Chris Harris, and Azalia Mirhoseini for their help in building the prototype. I would also like to thank the NSF, NDSEG, and ASEE for their financial support.

Words cannot express how grateful I am to my wonderful parents, who have provided my education and this opportunity.

Contents

Illustrations

Chapter 1

Introduction

Thanks to the popularization of smartphones, tablets and data-hungry applications, mobile data traffic is expected to grow 78% annually in the foreseeable future, an 18-fold increase within 5 years [1]. As a result, wireless operators are scrambling to acquire more spectrum resources and deploy more base stations to reduce cell sizes. However, there is a fundamental spectrum efficiency limit to existing and emerging cellular network architectures: they are *single-user* systems. That is, a base station only serves one terminal given a time slot, spectrum channel, or code sequence (resource block). Information theory shows that this limit can be overcome through multi-user multi-input, multi-output (MU-MIMO), or its special form called multi-user beamforming (MUBF). With MUBF, a base station employs many antennas to send independent data streams to multiple terminals in the same resource block, effectively improving spatial reuse. As the theory shows, the more base station antennas, the more terminals it can serve simultaneously resulting in higher spectral capacity. Not surprisingly, the theory community is envisioning large-scale 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., [2]. The key question to the proposal of MUBF base stations with many antennas remains: *is it practically*

feasible at all?

In this work, we answer this question with *Argos* [*], a flexible base station architecture that is scalable up to thousands of antennas and able to serve tens of terminals simultaneously through multi-user beamforming. Using commercial off-the-self software radio modules, i.e., the Rice WARP platform [3], we have realized an Argos prototype with 64 antennas that is capable of serving 15 terminals through zero-forcing and conjugate multi-user beamforming. Extensive experimental characterization using this prototype shows the spectral capacity can be boosted from 12.7 bps/Hz for a single-antenna base station to 85 bps/Hz for Argos employing zero-forcing beamforming, or 38 bps/Hz for Argos employing the much computationally simpler conjugate beamforming, while only using 1/64th of the original transmission power. We show that the spectral capacity grows nearly in proportion to the number of base station antennas and the number of simultaneously served terminals, as suggested by theory. At this moment, 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 ($M >> 10$). 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 principals. (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 reported prototype. As Argos scales up it selects the optimal beamforming algorithm by thoroughly analyz-

[*]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.

ing 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 are broadcasted 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 which allows localized weight computation at each antenna. We apply the transmission power normalization, which must be done centrally for standard conjugate beamforming, *locally* at each antenna, leveraging the fact that the channels to different terminals are statistically uncorrelated. This subtle 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, enabling real-time CSI estimation overhead to be independent of the number of base station antennas.

In summary, we make the following contributions to advance the state of the art of many-antenna multi-user beamforming:

- 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 of many-antenna MUBF and its capability to significantly improve spectral 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.

In the rest of this thesis, we provide the background and discuss related work in Chapter 2. We present the design and implementation of Argos in Chapters 3 and 4, respectively. In Chapter 5 we evaluate the real world performance of Argos, then conclude in Chapter 6.

Chapter 2

Background and Related Work

We first provide some background on multi-user beamforming and highlight the key benefits 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 transmissions. Due to constructive and destructive interferences of signals from multiple transmission antennas, the signal strength received at different directions varies spatially, leading to a *beampattern*, as shown in Figure 2.1. One can create various beam patterns by changing the beamforming *weights* applied to each antenna, effectively changing the amplitude and phase of the signal sent from that antenna. *Closed-loop* beamforming employs channel state information (CSI) to calculate the beamforming weights in order to maximize the signal strength at intended receivers and minimize the interference at unintended ones. In this work, we refer to closed-loop beamforming simply as beamforming unless otherwise indicated.

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

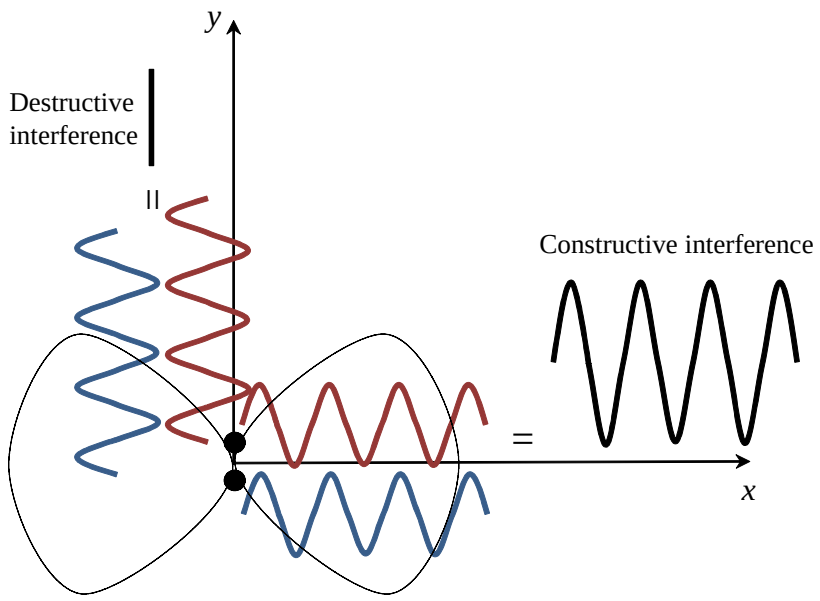


Figure 2.1 : Aerial view of the interference pattern created by two antennas emmitting a sine wave at the same frequency. Beamforming leverages multiple antennas to create constructive and destructive interference, thus forming a spatial beam-pattern.

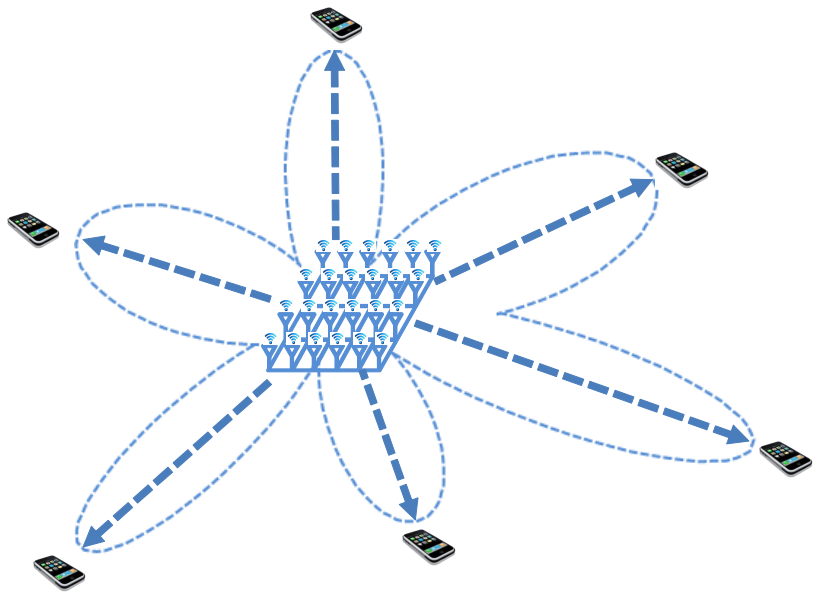


Figure 2.2 : Multi-user beamforming employs baseband precoding and many antennas to send independent data streams to multiple terminals at the same time.

conjugate of the CSI, while MUBF concurrently transmits multiple data streams, each to a different intended user as shown in Figure 2.2. Not surprisingly, information theoretical studies have shown that MUBF can improve spectral capacity manyfold. There are many baseband techniques to realize multi-user beamforming. 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 $\mathbf{W}$: $\mathbf{s}' = \mathbf{W}\mathbf{s}$. We refer to $\mathbf{W}$ as the beamforming weights.

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

Conjugate: $\mathbf{W} = \mathbf{W}_{conj} = \mathbf{H}^* \cdot c$, 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 multi-user beamforming is sub-optimal and may not perform well with a small M due to cross-terminal interference.

Zero-forcing: $\mathbf{W} = \mathbf{W}_{zf} = \mathbf{H}^* \left(\mathbf{H}^{\mathrm{T}}\mathbf{H}^*\right)^{-1} \cdot c$. 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 [4] with a small number of antennas (four) and terminals (four). Zero-forcing beamforming can keep inter-terminal interference to zero if $K \leq M$,

regardless of M. However, due to the required matrix inversion the computational overhead quickly becomes infeasible for real-time applications, as will be discussed in Chapter 2.4.2.

2.3 Benefits of Large-scale MUBF

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

First, MUBF can greatly improve the spectral capacity through spatial reuse. Roughly speaking, the spectral capacity gain from multi-user beamforming is $\min(M, K)$ [5]. 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 use to compensate for reduced transmission power; that is, a base station can achieve the same capacity with a much lower total transmission power. Moreover, multi-user beamforming essentially distributes the total transmission power to M antennas, leading to a much lower transmission power per antenna. The base station can therefore leverage much more efficient power amplifiers and simpler RF filters. This eliminates the need for active cooling, further reducing power consumption and total cost.

In Chapter 5, we will experimentally demonstrate both benefits using the Argos design: with 64 antennas and 15 terminals, the spectral capacity can be boosted from 12.7 bps/Hz to 85 bps/Hz and 38 bps/Hz for zero-forcing and conjugate beamforming respectively; while reducing the total transmission power to 1/64th of the original.

2.4 Challenges to Large-scale MUBF

Realizing the key benefits outlined above is, however, non-trivial. Any large-scale implementation of MUBF antennas faces fundamental timing constraints imposed by the coherence time of the physical wireless channel. MUBF must collect channel state information (CSI) for each terminal then use it to calculate beamforming weights within a fraction of the coherence time. Additionally, the computational complexity of MUBF beam weight calculation grows with the number of antennas, M, and the number of simultaneously served terminals, K. Both of these challenges have to be addressed by the Argos design.

2.4.1 CSI Estimation

Acquisition of CSI fundamentally limits the capacity of large scale MUBF. Beamforming with M antennas to serve K terminals requires CSI between every base station antenna and terminal, or MK channels. Importantly, all MK 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 terminals and environment move; in cellular systems this is typically on the order of a few milliseconds, but can drop below 500 microseconds with vehicular mobility at or near the terminals. 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 [6].

Traditionally, CSI is collected *explicitly*, that is, the basestation sends pilots to the the terminals which they use to estimate the channel. This channel estimate then has to be sent back to the basestation in order to perform downlink beamforming.

The reverse of this procedure is then used to find uplink CSI; however, the feedback is not necessary for maximum ratio combining at the base station (though it would be for uplink beamforming). This obviously does not scale up with M, as it requires $O(M + K)$ time to send pilots, and $O(MK)$ estimates which need to be sent back over-the-air. This suggests that a large scale system should leverage a TDD scheme to collect CSI *implicitly*. That is, terminals send uplink pilots which every basestation antenna listens to. This informs the basestation of uplink CSI, and, through channel reciprocity, the basestation can infer the downlink CSI as well. This limits CSI collection to $O(K)$ time, and eliminates the over-the-air transfer. While this reciprocity works in theory, in practical deployments the channels are *not* reciprocal. As shown in figure 3.2, the active RF components in the basestation and terminals form part of the channel, and are not reciprocal. In order to combat this, RF hardware has to be carefully calibrated, which is often time consuming, prone to error, or expensive. In Chapter 3.3, we present a simple and elegant calibration technique which overcomes this barrier.

2.4.2 Real-time Beam Weight Calculation

The computational complexity of MUBF beam weight calculation also grows with the number of base station antennas and the number of terminals. 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 that is $O(MK^2)$. While the incurred latency is acceptable at small-scales, the polynomial time nature of the inverse makes it very challenging for large-scale MUBF systems.

Additionally, both schemes incur high data transmission overhead because the beam weights have to be sent to and from the central controller from each of the modules. Zero-forcing cannot avoid this requirement because the inverse calculation requires the full H. Even the simplest beamforming algorithm, conjugate, requires full knowledge of H in order to appropriately scale the power of the steering weights. In Chapter 3.4, we present a novel method of localized conjugate beamforming in which we avoid the data dependency overhead of global power scaling.

2.5 Related Work

An information theoretic interpretation of our system would be the *MIMO Broadcast Channel (MIMO-BC)*. There have been many works since 2000 that study the capacity region of MIMO-BC [7, 8, 9, 10], and capacity-achieving coding techniques have been found, e.g., dirty paper coding [11]. The authors in [12] have analyzed the optimality of zero-forcing beamforming for broadcast scheduling. The authors of [13] similarly have studied the optimal zero-forcing beamforming with per-antenna power constraints. The authors of [14] studied a cellular architecture with unlimited number of base station antennas from an information theoretic perspective.

The Argos base station prototype is the most ambitious endeavor in multi-user beamforming for which there is no prior publicly reported realization with more than a few antennas, e.g., zero-forcing multi-user beamforming with four antennas in [4].

We note that most real-world cellular base stations already adopt multiple antennas, but are constrained to simple multi-antenna techniques such as antenna diversity, space-time coding, and sectorization. In particular, sectorization uses multiple antennas to form directional beams, each of which covers a range of directions and forms a sector. Terminals in different sectors can be simultaneously served. Therefore, sector-

ization improves the cell capacity via spatial reuse. It can be treated as a special case of multi-user beamforming where the number of simultaneous terminals is limited by the beamwidth of sectors.

Chapter 3

Design

The key question we ask in this Chapter is: *how do we design a multi-user beamforming base station that can flexibly optimize its architecture over a wide range of M and K?* Before proceeding to answer it, we need to highlight its practical interest: realistic wireless networks often have large a variation 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. While traditional base stations can only scale their transmission power or, equivalently, their cell size, Argos is presented with a unique opportunity to be able to scale cost-effectively with 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* multi-user beamforming scale up? Multi-user beamforming entails three distinct phases: 1) Channel Estimation, 2) Weight Calculation, and 3) Linear Precoding. We explore the feasibility and design implications of these as M

scales up.

3.1.1 Channel Estimation

Traditional channel estimation *does not* scale well with M or K, and typically requires $M + K$ pilots, as discussed in Section 2.4.1. However, we find that by exploiting channel reciprocity, which is enabled by our novel calibration scheme presented in Section 3.3, channel estimation overhead *can* remain constant as M scales up.

3.1.2 Beamforming Methods

Unfortunately, existing beamforming methods are distinctly unscaleable, as they all have centralized data requirements. To do any form of interference cancellation, such as zeroforcing, it is quite obvious that all of the K channel estimates from M radios must be taken in to account in order to correctly steer nulls. This implies that there must be a central computation point which supports a data rate that scales with $O(MK)$. Interference cancellation techniques, such as zeroforcing, also typically incur a huge computational overhead of O(MK^2). More subtly, even the simplest form of beamforming, conjugate, has a global power scaling factor which requires centralized computation using all of the channel estimates. In light of this, we propose a novel beamforming method which allows weights to be computed completely locally, at each basestation radio, described in Chapter 3.4. Leveraging this method, Argos can easily scale to an unprecedented number of basestation antennas, e.g. 1000. However, while this beamforming method performs well with a very large number, e.g., 100s, of basestation 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 emperically in our results, Chapter 5, where we find that zeroforcing results in up

to a 4x capacity increase over conjugate beamforming. We conclude that in order to scale optimally, Argos must support centralized beamforming techniques for smaller scale, though still many antenna, basesation 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 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 datarate requirements. Notably, there is still a need for some form a of central controller to demodulate the data once it has been completely recombined; however this operation is latency insensitive, and computationally trivial.

Thus we find that, yes, multi-user beamforming *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: 1) a central controller which handles modulation and demodulation, and 2) the M radio front-ends with 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 obviously does not scale well. On the other hand, we can daisy-chain all the radios serially. While scalability seems to be maximized, reliability and delay of the system is 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 which contains one or more radio front-ends. Next, in order to allow 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 switch, which allows multiple modules to be connected in parallel. Figure 3.1 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: 1) by adding more Argos switches, 2) by increasing

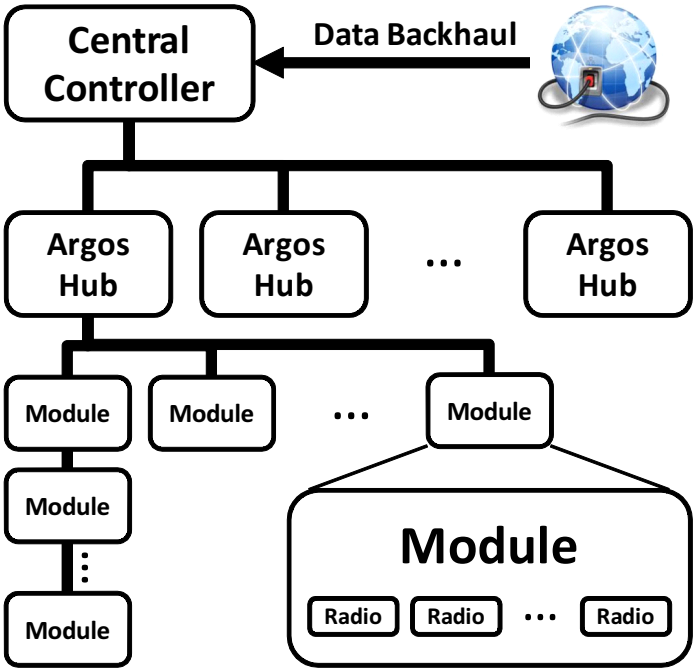


Figure 3.1 : Argos architecture: fat tree structure with daisy-chained leaf nodes

the length of the module chains, and 3) by increasing the number of antennas on a module. The hierarchal architecture facilitates large scale deployments to be flexibly distributed geographically by using a single link to an Argos switch, as well as small scale deployments where the switch can be omitted completely, and modules are simply chained together in series. Additionally, if chains become too long to meet latency requirements, Argos switch 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. As discussed in Section 2.4.1, traditional

methods of collecting CSI do not scale with M. To combat this scalablility issue Argos employs implicit beamforming 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, it 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 3.2, the uplink and downlink channels are *not* reciprocal due to the random phase and amplitude effects of 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 tx_i and rx_i.

The uplink and downlink channels between any two transcievers is a product of (*i*) the frequency response of the TX chains, (*ii*) the physical wireless channel, and (*iii*) the frequency response of the RX chains:

$$\hat{h}_{i\to j} = tx_i \cdot h_{i\to j} \cdot rx_j \tag{3.1}$$

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:

$$b_{i\to j} = \frac{\hat{h}_{i\to j}}{\hat{h}_{j\to i}} = \frac{tx_i \cdot h_{i\to j} \cdot rx_j}{rx_i \cdot h_{j\to i} \cdot tx_j} = \frac{tx_i \cdot rx_j}{rx_i \cdot tx_j} = \frac{1}{b_{j\to i}} \tag{3.2}$$

Notably, if both channels are measured within the coherence time then $h_{k\to m} = h_{m\to k}$ 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:

$$\hat{h}_{i\to j} = b_{i\to j} \cdot \hat{h}_{j\to i} \quad \text{or} \quad \hat{h}_{j\to i} = \frac{\hat{h}_{i\to j}}{b_{i\to j}} \tag{3.3}$$

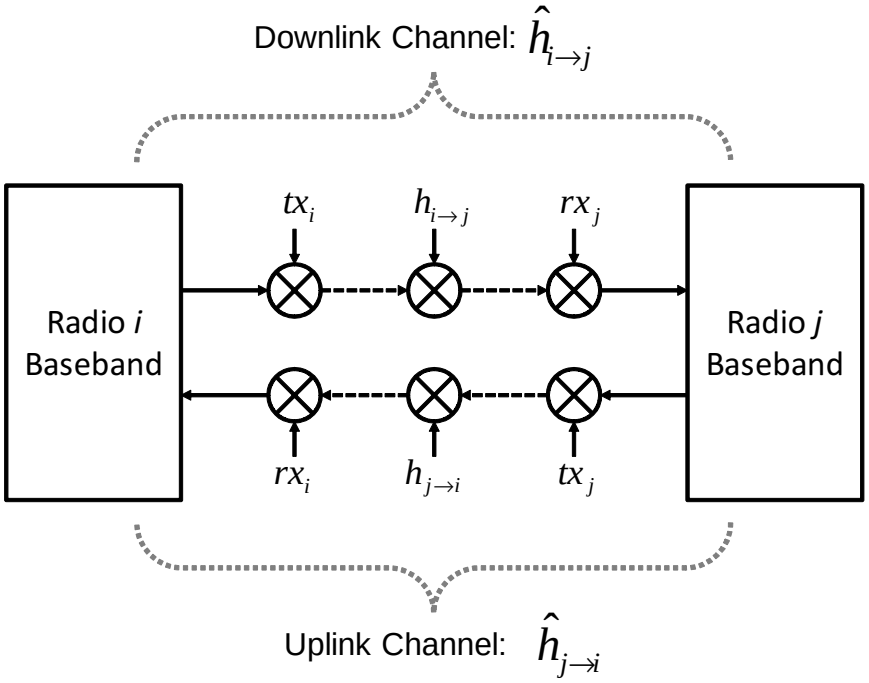


Figure 3.2 : Real channels are not reciprocal due to the differences in TX and RX hardware. Note that channel reciprocity indicates that within the channel coherence time the physical channel *is* reciprocal: $h_{i\to j} = h_{j\to i}$. Dashed lines indicate the channel is wireless.

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 TX and RX 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 $\hat{h}_{m\to k}$.

Internal Calibration However, 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 3.2. Note that these coefficients are in fact stable over long periods of time, as we show in 5, 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:

$$\frac{b_{i\to j}}{b_{i\to y}} = \frac{\frac{tx_i \cdot rx_j}{rx_i \cdot tx_j}}{\frac{tx_i \cdot rx_y}{rx_i \cdot tx_y}} = \frac{tx_y \cdot rx_j}{rx_y \cdot tx_j} = b_{y\to j} \tag{3.4}$$

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

$$\hat{h}_{k\to m} \cdot \frac{b_{1\to m}}{b_{1\to k}} = \hat{h}_{k\to m} \cdot b_{m\to k} = \hat{h}_{m\to k} \tag{3.5}$$

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. A similar approach has been proposed in [15] and [16].

Key Idea: Relative Calibration Our key idea in solving the 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 the CSI estimation of an base station antenna deviates from the real CSI by the same multiplicative factor as that of other base station antennas, multi-user beamforming will still result in the same beampattern. To visualize this, refer back to Figure 2.1; if both antennas were to experience the same phase offset, the resulting spatial beampattern would remain the same. Thus, we can assume $b_{1\to k} = 1$:

$$\hat{h}_{m\to k} = \hat{h}_{k\to m} \cdot \frac{b_{1\to k}}{b_{1\to m}} \quad \Rightarrow \quad \hat{h}'_{m\to k} = \frac{\hat{h}_{k\to m}}{b_{1\to m}} = \hat{h}_{k\to m} \cdot b_{m\to 1} \tag{3.6}$$

This means that we estimate *relative* downlink CSI, $\hat{h}'_{m\to k}$, by using *only* uplink pilots, without any feedback! To recapitulate, this 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 3.6.

4. Use $\hat{h}'_{m\to k}$ to calulculate the beam weights, then send the beamformed data.

Using this process we can effeciently 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 to realize a large-scale multi-user beamforming system.

Note that, the measurements of downlink and uplink has to be done within the channel coherence time in order for $h_{m\to 1} = h_{m\to 1}$. 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 contraints; 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, 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 BS radio exceeds its maximum power output (i.e., clips). For example, assuming a maximum radio transmit amplitude of 1, and in order to ensure at least one radio transmits at maximum power:

$$c = \left(\max \left(\sum_{k=1}^{K} \|\mathbf{h}_{m,k}\| \right) \right)^{-1} (m = 1, 2, ...M) \tag{3.7}$$

where c is the scaling factor used in the beamforming weight calculation ($\mathbf{W} = \mathbf{H}^* \cdot c$). 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 BS

antenna's weight for a given terminal, which ensures per-terminal transmission energy optimality, as proven in [17]. However, each BS 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 which closely approximates global normalization.

Argos leverages a key observation that for the different terminals in multiuser 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 BS antenna using only the CSI it measures:

$$c_m = \left(\sum_{k=1}^{K} \|\mathbf{h}_{m,k}\| \right)^{-1} \quad (m = 1, 2, ...M) \tag{3.8}$$

The conjugate beamforming weights are then scaled via:

$$\mathbf{W} = \mathbf{H}^* \cdot \mathrm{diag}(C) \tag{3.9}$$

Where C is the scaling vector given by $C_{local} = [c_1, c_2, ...c_M]$, from equation 3.8; notably the globally scaled conjugate can also be found in this form, using $C_{global} = [c, c, c, c...]$, from equation 3.7.

We have experimentally verified the effectiveness of such local power scaling and we observed that its performance is almost indistinguishable from the optimal global power scaling method (see Chapter 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 shown in the following section.

3.4.1 Proof of SNR Improvement

While our local conjugate method is suboptimal with regard to power, it can always achieve equal or greater SNRs at the terminals. This is because the global power scaling is constrained by the maximum transmit power of one radio, whereas our local method can fully utilize the transmit power of every radio, and thus send with a higher total transmit power. First we note beamforming over channel $\mathbf{H}$ using weights $\mathbf{W}$ results in signal power at terminal k defined as:

$$P_k = \left\| \left(\sum_{m=1}^{M} \mathbf{h}_{m,k} \cdot \mathbf{w}_{m,k} \right) \right\|^2 \tag{3.10}$$

If the beamweights are a conjugate of the channel matrix, as defined in equation 3.9, then this becomes:

$$P_k(C) = \left(\sum_{m=1}^{M} \|h_{m,k}\|^2 \cdot c_m \right)^2 \tag{3.11}$$

Rewriting equation 3.7 using equation 3.8 we find:

$$c = \min \left(\left(\sum_{k=1}^{K} \|\mathbf{h}_{m,k}\|^2 \right)^{-1} \right) = \min(c_m) \quad \Rightarrow c_m \geq c \qquad (m = 1, 2, ...M) \tag{3.12}$$

Thus:

$$P_k(C_{local}) = \left(\sum_{m=1}^{M} \|h_{m,k}\|^2 \cdot c_m \right)^2 \geq P_k(C_{global}) = \left(\sum_{m=1}^{M} \|h_{m,k}\|^2 \cdot c \right)^2 \quad \square \tag{3.13}$$

Q.E.D, our local conjugate method will always result in higher terminal SNRs. Of course, increasing transmit power increases inter-terminal interference, however since the SNR and interference to all terminals is increased proportionally, this does not significantly affect performance (as verified by our results).

Chapter 4

Implementation

In this section we provide a detailed report of our implementation of Argos which leverages WARP [3], commercially available clock distribution boards, a commodity PC, and an ethernet switch. Figure 4.1 shows an abstract representation of our implementation. As the first proof-of-concept prototype, our system includes a central controller, an Argos switch 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 switch 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 platform will be physically separated as well. Each radio module is a single WARP board with 4 radio daughtercards and 4 antennas. Figure 4.2 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 platform we are using, this bottleneck is the number of multipliers (328 on the Virtex 2 Pro xc2vp70) [18]. 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

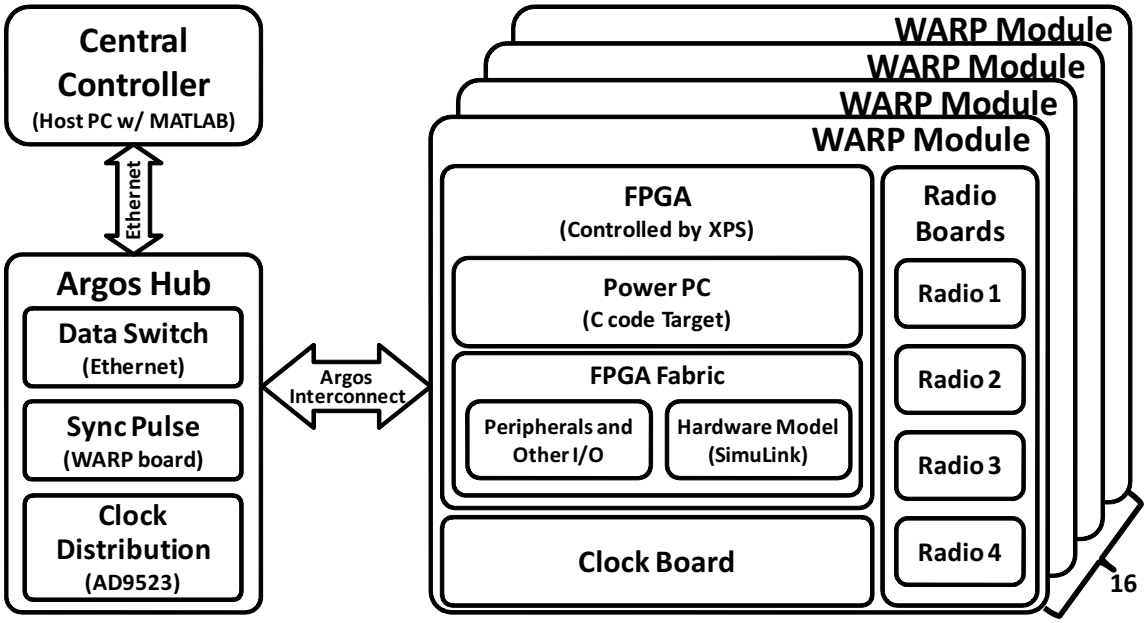


Figure 4.1 : Our implementation of Argos using WARP boards, a laptop, an ethernet switch, and an AD9523 based clock distribution board.

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 40Mhz sample (neglecting routing overhead and other functions that require mutlipliers), which would, obviously, alleviate this bottleneck [19].

To the best of our knowledge, our Argos prototype is the first publicly reported large-scale multi-user beamforming system with real-world feasibility. We next elaborate our implementation.

4.1 Hardware and Software Platform

WARP is a scalable and programmable wireless platform, built from the ground up, to prototype advanced wireless systems. Each WARP board allows up to four radio daughter cards to be connected and therefore can contribute up to four active anten-

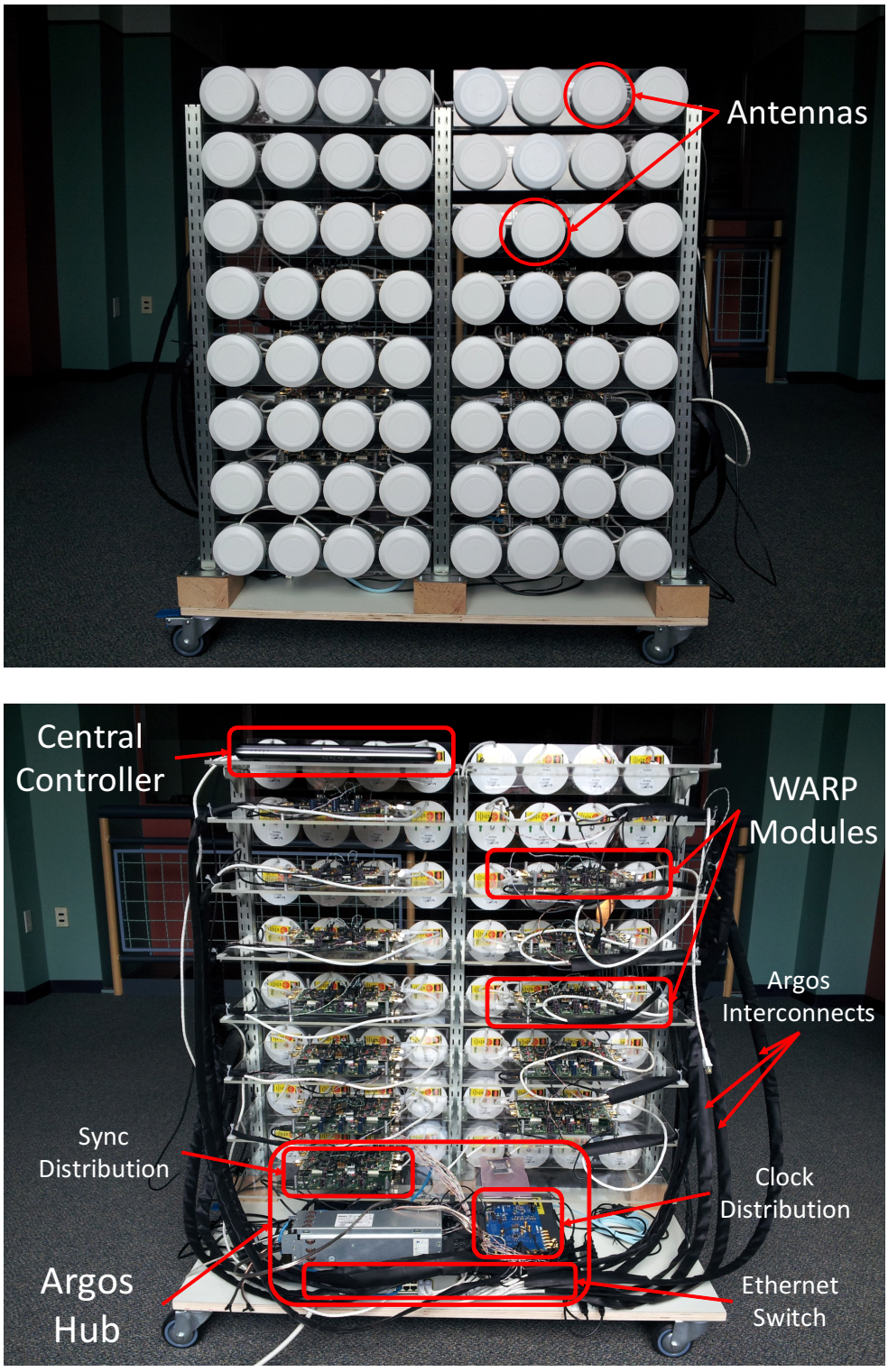


Figure 4.2 : Our prototype of Argos with 16 modules and 64 antennas. Top: front-side, showing antenna array; Bottom: back-side, depicting Argos architecture.

nas simultaneously to Argos. Each radio board includes a Maxim 2829 transceiver chip [20]. 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 4.1, WARPLab consists of four layers: 1) The underlying Simulink model which implements the custom hardware for controlling the FPGA board and radio boards, as well as linear precoding; 2) The Xilinx Platform Studio (XPS) project which integrates and connects all of the hardware components, including the Simulink model, the I/O cores for the serial port, Ethernet port, clocking, etc.; 3) The C code which runs on the PowerPC microprocessor, controls the hardware through memory mapped I/O, and acts as an interface to the Ethernet port; 4) The MATLAB interface which configures the boards, generates the transmit samples, and processes the receive samples.

We have extensively customized the WARPLab framework to enable 1) hardware multi-user beamforming, 2) transmission synchronization, 3) clock synchronization, and 4) indirect calibration among base station antennas. These functionalities are essential to for Argos to enable large-scale multi-user beamforming.

4.2 Hardware Multi-user Beamforming

A straightforward, and much easier approach to realize multi-user beamforming 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 transmissions, which increases linearly with the number of basestation antennas, and severely degrades performance. This is a result of the same scaling

problem discussed in Chapter 3.1. Therefore, we must modify the WARPLab hardware to enable hardware multi-user beamforming.

Mathematically, applying the beamforming weights consists of multiplying the baseband signal for each terminal by its corresponding beamforming weight, and then adding them together, or

$$s' = \sum_{k=1}^{K} \mathbf{w_k} * s_k, \tag{4.1}$$

where s' is the resultant beamformed signal vector; $\mathbf{w_k}$, s_k are the beamforming weight vector and modulated baseband signal for client k. 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, $\mathbf{w_k}(k = 1, 2, ...K)$, in the memory mapped registers. This is important since the PowerPC, and in turn, the MATLAB interface can directly control them.

4.3 Transmission Synchronization

WARPLab has a default function to enable transmission synchronization between multiple 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 an 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 switch, 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 cycles, the cables should be within λ length. 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 iteslf with an effective cable length of 0. This means the other cables must be less than 7.5 m, which is not a problem; in our current setup the length is 2 m. 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 architecture with our modular design. 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 in fact use a 40 MHz clock. Therefore, we can 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 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 switch), which would provide 17 more outputs. Alternatively, the existing modules can forward their clocks to additional modules, through Argos' multihop extension.

4.5 Indirect Calibration

For indirect calibration, we need to estimate $\frac{c_n}{b_n} * \frac{b_1}{c_1}$ for each antenna n with respect to the "reference antenna". 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 the 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 stability time.

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 calibration procedure's isolation of the radio hardware channel from the physical channel.

Chapter 5

Evaluation and Results

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 large-scale multi-user beamforming, using equal transmission power.* With 64 antennas and 15 terminals, the spectral capacity can be boosted from 12.7 bps/Hz to 85 bps/Hz for zero-forcing beamforming, and 38 bps/Hz for conjugate beamforming, while using a mere 1/64th of the original 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 proportionally 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 multi-user beamforming to 15 concurrent terminals. Since it is relatively easy to move our platform (see Figure 4.2), we moved the platform to various indoor locations (see Figure 5.1), allowed by the communication range of WARP, 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

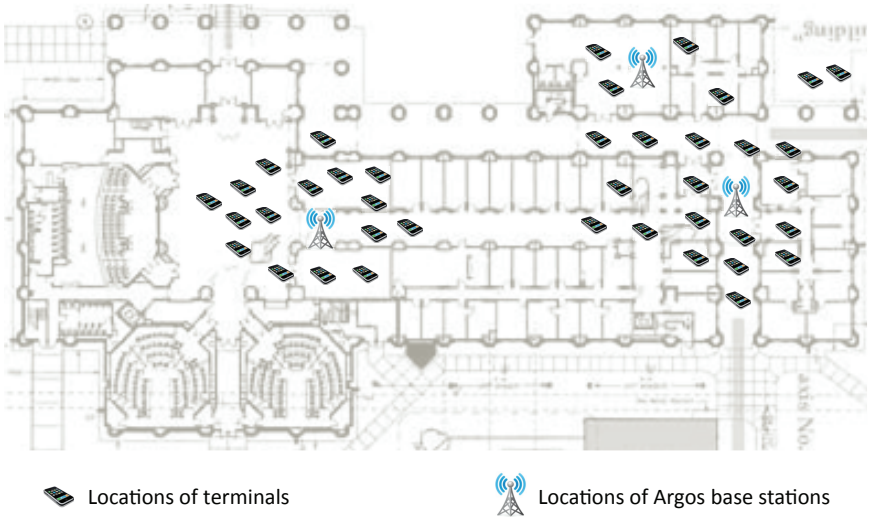


Figure 5.1 : Environments for our experiments. We have marked the locations of the base station and terminals for our measurements.

measurements at each location) to reliably average out the performance.

To obtain the network capacity, we aggregate the Shannon capacity for each terminal, or

$$C_{Network} = \sum_{k=1}^{K} \log(1 + SINR_k). \tag{5.1}$$

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 multi-user beamforming is a hardware 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

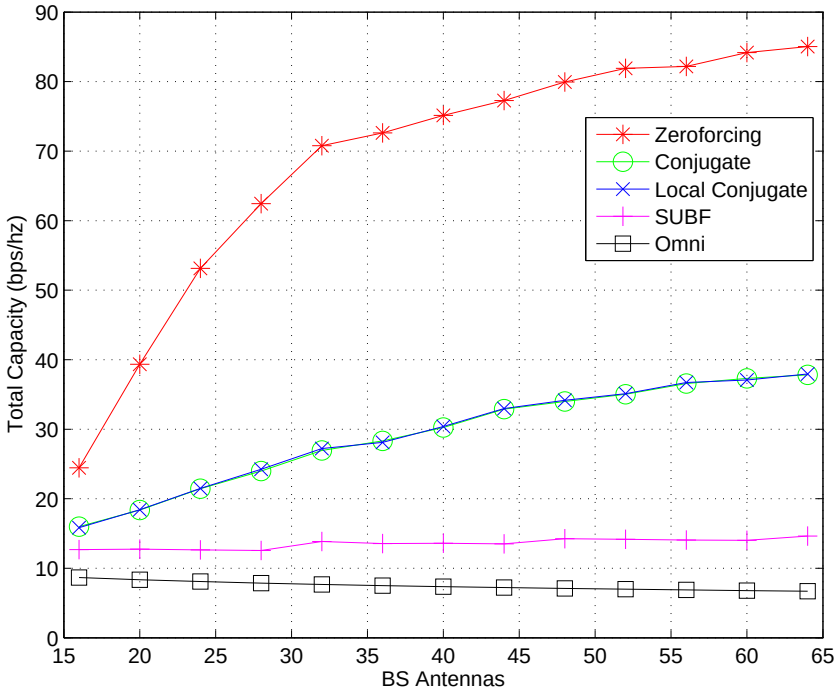


Figure 5.2 : Network capacity as the number of base station antennas (M) increases, with 15 terminals. Total transmission power is $1/M$.

acquire the SINR accordingly. To make sure the channel remains constant during the transmissions we conduct our experiments in an ultra-stable environment (late at night, without moving people and wireless traffic).

5.2 Improvement of Network Capacity

The primary purpose of our experiments is to inspect the capacity improvement of Argos, in order to ultimately answer the feasibility question to the many-antenna MUBF base station proposal from the theory community. To see how the network capacity improves, we first vary the number of base station antennas, M, assuming a fixed number of terminals $K = 15$. Figure 5.2 shows $C_{Network}$ as a function of M for

both conjugate beamforming and zero-forcing beamforming. We have the following key observations:

- When M is much larger than K, both conjugate and zero-forcing beamforming increase the network capacity as M scales up, *despite reducing transmission power proportionally with* M, as shown in Figure 5.2. The beamforming gain from the additional antennas compensates for the power reduction, as demonstrated by the flat performance of single-user beamforming (SUBF), while simultaneously increasing the natural orthogonality of the terminals. This reduces the inter-terminal interference of conjugate beamforming, and reduces the amount of power wasted to create nulls for zero-forcing beamforming. With $M = 64$ the improvement for conjugate and zero-forcing beamforming over a single antenna are 5.7x and 12.7x for equal power, or 3x and 6.7x for 1/64 power, respectively.

- 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. When $M = K$ this inefficiency can even result in conjugate beamforming out-performing zero-forcing.

We next fix M and vary the number of terminals, K, to see the network capacity change. For a fair comparison, the total transmission power is held constant, implying the power per terminal is scaled by $1/K$.

- When $M >> K$, as shown in Figure 5.3, capacity increases approximately linearly with the number of terminals for both conjugate and zero-forcing beam-

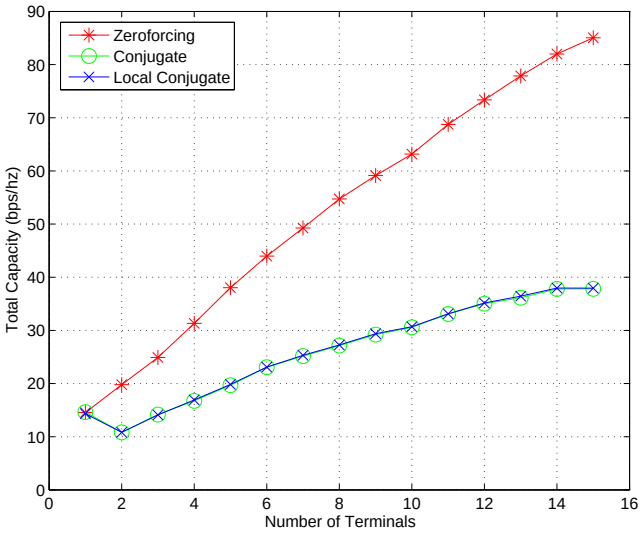


Figure 5.3 : Network capacity with 64 base station antennas as the number of terminals increases.

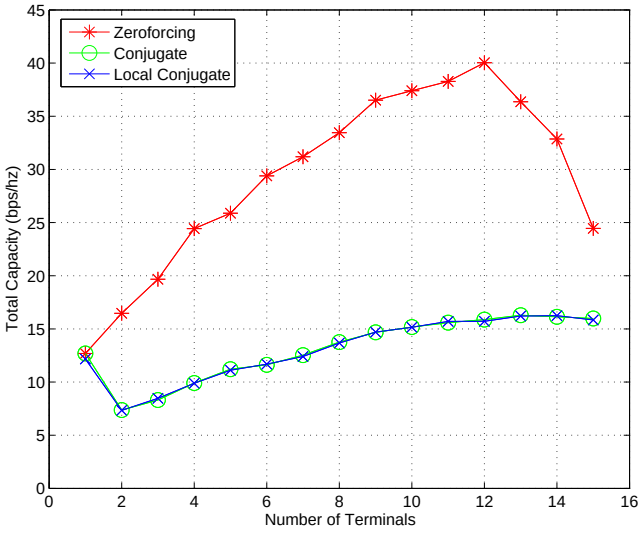


Figure 5.4 : Network capacity with 16 base station antennas as the number of terminals increases.

forming; this is attributable to the multiplexing gains from simultaneously serving K terminals.

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

- We find, however, in Figure 5.4, that as K approaches M the performance of zero-forcing drops sharply (for the same reasons described above). 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.

- As the transmission power is reduced, conjugate beamforming performs relatively better than zero-forcing, as shown in Figure 5.5. 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 Beamforming.

In order to verify the viability of our localized method for conjugate beamforming, we implement it in Argos and compare it to standard beamforming with global power control. As shown in Figure 5.6, we see that our local power control method results in

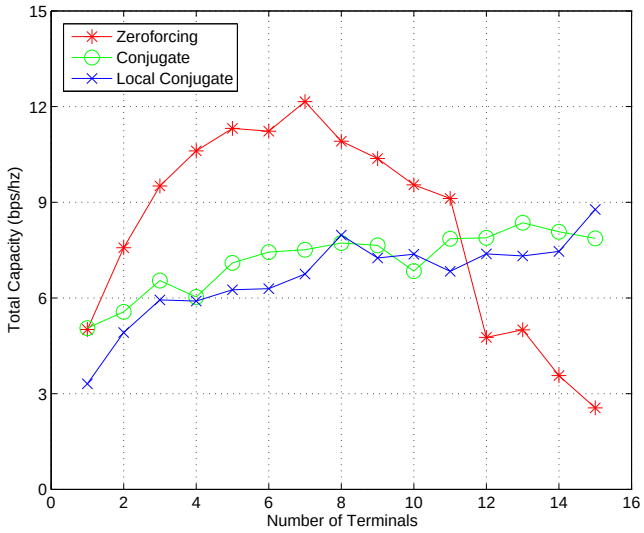


Figure 5.5 : Network capacity with 16 base station antennas and very low transmit power as the number of terminals increases.

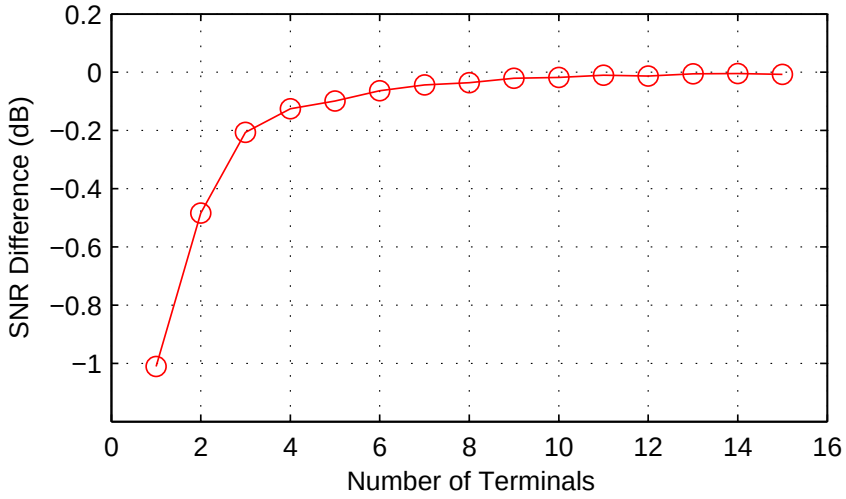


Figure 5.6 : The relative signal power between conjugate and our conjugate with local power scaling, sent at the same transmit power. Local conjugate performs within 1.5dB of global conjugate, and quickly converges to 0 dB as K increases.

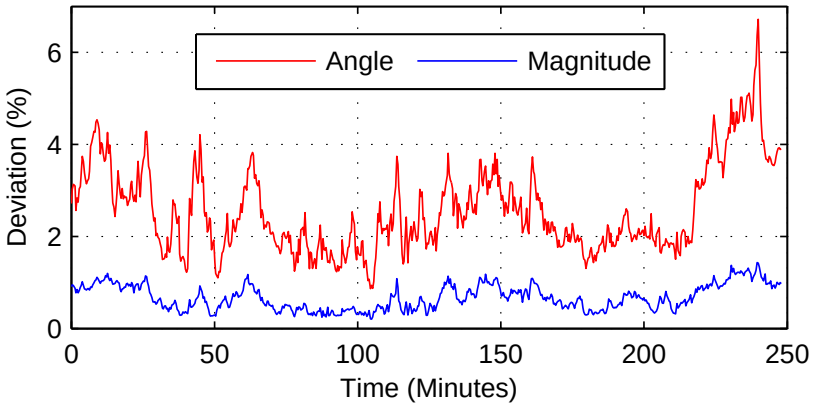


Figure 5.7 : Our calibration procedure exhibits an average instantaneous noise of less than 7% and remains stable indefinitely.

a signal power within 1.2dB of global power control, 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 control 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 5.2 to 5.5 that the performance difference between global scaling and local scaling 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 5.7 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 .7% (maximum 1.4%), over a period of 4 hours. Notably, these measurements were taken during the day

with normal movement around the basestation, indicating the calibration procedure is stable in real world environments. Angle deviation is calculated by difference in angle from average angle over pi, 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.

Chapter 6

Concluding Remarks

In this thesis, we present the design, realization, and evaluation of Argos, a base station architecture that can employ thousands of antennas to serve tens of terminals simultaneously through multi-user beamforming. In order to enable this unprecedented scaling in a practical environment we devise and evaluate a novel beamforming algorithm that is completely decentralized, as well as a new calibration method to facilitate implicit beamforming across large arrays, thus allowing CSI to be collected in constant time with regard to the number of base station antennas. Argos employs a hierarchal modular design which facilitates flexible, scalable, deployments while simultaneously constraining latency and providing fault tolerance.

Our experimental characterization of a base station 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. This work, however, already demonstrates the significant promise that large scale MUBF holds for the future of wireless communication.

Bibliography

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

[2] E. Aryafar, N. Anand, T. Salonidis, and E. Knightly, "Design and experimental evaluation of multi-user beamforming in wireless lans," in *Proceedings of MobiCom*, 2010.

[3] Rice University, "Rice University Wireless Open Access Research Platform." `http://warp.rice.edu/trac`.

[4] E. Aryafar, N. Anand, T. Salonidis, and E. W. Knightly, "Design and experimental evaluation of multi-user beamforming in wireless lans," in *Proc. Int. Conf. Mobile Computing and Networking (MobiCom)*, 2010.

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

[6] T. Marzetta, "How much training is required for multiuser MIMO?," in *Signals, Systems and Computers, 2006. ACSSC '06. Fortieth Asilomar Conference on*, pp. 359 –363, 29 2006-nov. 1 2006.

[7] S. Vishwanath, N. Jindal, and A. Goldsmith, "Duality, achievable rates, and sum-rate capacity of gaussian MIMO broadcast channels," *IEEE Transactions*

on Information Theory, vol. 49, pp. 2658–2668, October 2003.

[8] N. Jindal and A. Goldsmith, "Dirty-paper coding versus TDMA for MIMO broadcast channels," *IEEE Transactions on Information Theory*, vol. 51, pp. 1783–1794, May 2005.

[9] M. Sharif and B. Hassibi, "On the capacity of mimo broadcast channels with partial side information," *IEEE Transactions on Information Theory*, vol. 51, pp. 506–522, February 2005.

[10] H. Weingarten, Y. Steinberg, and S. Shamai, "The capacity region of the gaussian multiple-input multiple-output broadcast channel," *IEEE Transactions on Information Theory*, vol. 52, pp. 3936–3964, September 2006.

[11] M. Costa, "Writing on dirty paper (corresp.)," *IEEE Transactions on Information Theory*, vol. 29, pp. 439–441, may 1983.

[12] T. Yoo and A. Goldsmith, "On the optimality of multiantenna broadcast scheduling using zero-forcing beamforming," *IEEE Journal of Selected Areas in Communications*, vol. 24, pp. 528–541, March 2006.

[13] K. Karakayali, R. Yates, G. Foschini, and R. Valenzuela, "Optimum zero-forcing beamforming with per-antenna power constraints," in *IEEE International Symposium on Information Theory*, June 2007.

[14] T. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," *IEEE Transactions on Wireless Communications*, vol. 9, pp. 3590–3600, November 2010.

[15] A. Bourdoux, B. Come, and N. Khaled, "Non-reciprocal transceivers in OFDM/SDMA systems: impact and mitigation," in *Radio and Wireless Conference, 2003. RAWCON '03. Proceedings*, pp. 183 – 186, aug. 2003.

[16] M. Guillaud, D. Slock, and R. Knopp, "A practical method for wireless channel reciprocity exploitation through relative calibration," in *Signal Processing and Its Applications, 2005. Proceedings of the Eighth International Symposium on*, vol. 1, pp. 403 – 406, 28-31, 2005.

[17] T. Lo, "Maximum ratio transmission," *Communications, IEEE Transactions on*, vol. 47, pp. 1458 –1461, oct 1999.

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

[19] Xilinx Inc., "7 Series FPGAs Overview." `http://www.xilinx.com/support/documentation/data_sheets/ds180_7Series_Overview.pdf`.

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