Design and Demonstration of a Scalable Massive MIMO Uplink at E-Band

Greg LaCaille, James Dunn, Antonio Puglielli, Lorenzo Iotti, Sameet Ramakrishnan, Lucas Calderin
Zhenghan Lin, Emily Naviasky, Borivoje Nikolic, Ali Niknejad, Elad Alon
Department of EECS, Berkeley Wireless Research Center
University of California at Berkeley
Berkeley, USA
greglac@berkeley.edu, jhdunn@berkeley.edu, apuglielli@berkeley.edu, lorenzo.iotti@berkeley.edu,
sameetr@berkeley.edu, lcalderin@berkeley.edu, zhenghan_lin@berkeley.edu,
enaviasky@berkeley.edu, bora@berkeley.edu, niknejad@berkeley.edu, elad@berkeley.edu

***Abstract*—In this paper we present a real-time, multi-user massive MIMO hardware testbed operating in E-Band. We propose a system architecture that scales efficiently to large arrays at sampling rates associated with mm-wave bandwidths. The architecture implements two-stage beam-forming algorithms on distributed hardware, as well as per-antenna channel estimation. The features of this architecture address practical issues associated with interconnecting modular radios and steering the beams of highly directive arrays, while providing minimal hardware overhead. The hardware testbed is built with custom radio boards, off-the shelf mm-wave radio components, and FPGAs. Single and multi-user measurements are performed on the hardware to validate the effectiveness of the proposed architecture.**

***Index Terms*—Massive MIMO, E-Band, mm-wave, Beamforming, 5G**

I. Introduction

As mobile handset data consumption continues to increase, and with new wireless applications such as industrial control, autonomous vehicles, and virtual reality on the horizon, the demand to extend wireless channel capacity by orders of magnitude shows no sign of ceasing. The massive MIMO paradigm, in which a large number of base station antenna elements spatially multiplex user streams, has emerged as one of the most promising means of increasing channel capacity to projected demand [5]. A recent direction of research in massive MIMO systems is toward implementation at mm-wave carrier frequencies. The reduced size of mm-wave antennas allows large arrays to be implemented in more compact, lightweight, and inexpensive form factors. Available bandwidth in the E-Band (60-90GHz) when sub-6GHz spectrum is nearly fully allocated makes mm-wave implementation more desirable. An additional consideration for network operators is flexibility. To realize the ubiquitous connectivity envisioned in 5G and beyond, network operators will deploy base stations in a wide variety of settings. For example, base stations targeting a streetlamp post versus a telecommunications tower will need to serve different numbers of users, with different degrees of inter and intra-cell interference. Realizing massive, scalable arrays at mm-wave brings many challenges [1], the most fundamental of which is architecture. The organization of baseband processing elements in a massive MIMO system dictates interconnect, computational complexity, and ultimately power consumption. Current architectures for massive MIMO baseband processing are well-suited to systems with narrow channel bandwidths [8] [9] [10]. In order to realize large arrays and bandwidths for gigabit communication, new architectures must be developed.

In this paper we explore an architecture for a scalable massive MIMO uplink operating at mm-wave. The architecture is well-suited for a system that can aggregate smaller sub-arrays to form a larger system. The signal processing techniques required to deal with non-idealities such as timing skew, phase noise, and interference are all incorporated into the design. We construct a prototype array consisting of 20 base station antennas and demonstrate it with two users, operating at 75GHz with 250MHz of channel bandwidth. Data from this prototype array is used to demonstrate a full MIMO uplink signal processing chain.

The paper is organized as follows. Section II discusses the challenges with signal processing for large-scale arrays and potential architectural trade-offs. Section III details the prototype hardware modules and presents an uplink architecture that can leverage trade-offs discussed in Section II. Section IV demonstrates measurements of system functionality and performance with the proposed uplink architecture. Finally, Section V summarizes results and proposes future work.

II. Energy-Efficient Signal Processing for Scalable Arrays

It has been shown that for sufficiently large arrays, the effectiveness of linear spatial processing techniques approaches optimal [7]. Additionally, any linear spatial processing methods can be expressed as a set of frequency-dependant matrix operations. Most challenges in implementing large-scale arrays arise from two core issues: first, mapping the required matrix

This work was supported in part by an NSF EARS MIMO grant. This work was supported in part by tasks 2778.026 and 2778.007 of the ComSenTer Research Center, one of six centers in JUMP, a Semiconductor Research Corporation (SRC) program sponsored by DARPA.

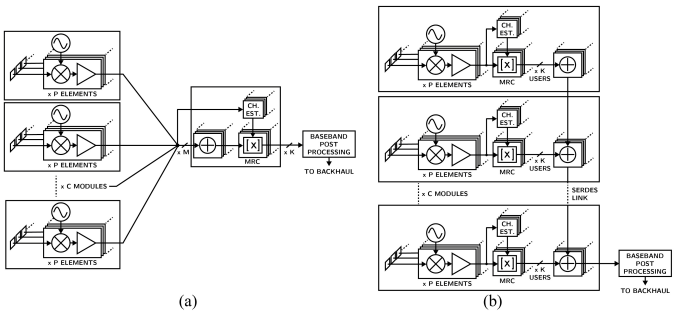

Fig. 1. Centralized (a) and distributed (b) beam-forming architecture.

operations to hardware in a practical and energy-efficient manner; second, determining the required coefficients for the spatial processing matrices.

A. Data Aggregation of Sub-Arrays

In a fully integrated single module array [6], the implementation of matrix operations is relatively straightforward. However, to achieve large array sizes, inevitably the array will need to be split into a series of front-end modules that are aggregated in some. This presents an important area of consideration in implementation of an architecture, as the power required to combine the signals from different modules can dominate the total power consumption of the system.

To efficiently map the processing algorithms, it is useful to break linear spatial processing methods such as zero-forcing (ZF) and minimum mean square error (MMSE) into two phases: beam-forming and decorrelation of users. [4] presented an analysis of separate beam-forming and decorrelation stages. The first stage uses complex conjugate operations to maximize SNR and form a beam to each user across the antenna array, while the decorrelation stage removes inter-user interference within the beam space and increases SINR at the cost of SNR.

In Figure 1-a, we show an architecture in which C modules each pass samples directly to a central element which aggregates data, performs maximum-ratio combining (MRC) based beam-forming, and decorrelates. In this case, interconnect bandwidth and dimension of matrix multiplication operations at the the central element scale with $M = C * P$, the total number of antennas.

In contrast, a distributed architecture is shown in Figure 1-b. In this case, C modules perform MRC beam-forming across their respective P antennas, reducing data order to K. Data is accumulated among neighboring modules in sequence until the final distributed module sends data to a central element for decorrelation and post-processing. Data interconnect scales with K, and dimension of matrix multiplication operations scales with the greater of K and P. As K is necessarily much less than M in a massive MIMO system, and P may be made arbitrarily smaller than M based on the number of modules, interconnect and computational requirements are relaxed, with significant power savings. Uniform modules may be added to increase array size and therefore channel capacity, allowing for more flexible implementations across the design space of system power and cost.

B. Channel Estimation

The highly directional nature of large arrays creates an inherent need for an algorithm to steer the direction of the beams. This equates to determining coefficients for the spatial processing matrix. Two classifications of strategy typically exist for this process: open-loop beam steering and closed-loop channel estimation.

Open-loop beam steering relies on pre-calibrated lookup tables that store the necessary matrix coefficients for each element for a desired beam pattern. Typically, the desired beam pattern is selected by some sort of beam search algorithm. While this requires little in terms of physical hardware, it can take the beam search algorithm a large amount of time to converge relative to the channel coherence time for mobile scenarios, as it requires feedback over the downlink on which beam to use [2]. Additionally, this search time grows with the number of elements M as the beam becomes more narrow. This long search time can limit the practicality of this technique to stationary point-to-point links. Open-loop tables may also be sensitive to process and temperature variation and typically require individual units to be calibrated before deployment.

Closed-loop channel estimation relies on the ability to measure the channel from each user to each base station antenna. This method can converge much faster than open-loop searches, since no feedback across the wireless channel via the downlink is required. Channel estimation has the additional benefit of absorbing any phase or gain variations in the transceiver into the channel, eliminating the need for precise calibration of the variation across elements.

The greatest disadvantage of per-antenna channel estimation-based techniques is the lack of SNR gains associated with beam-forming. This creates an SNR gap between the channel estimation and payload section of the the uplink frame. This gap can be closed by using a sufficient amount of coding gain or averaging for the pilot signals. However, this can lead to a substantial hardware demand, as pilot correlators are required at each antenna. Selection of the pilot scheme can keep the estimation hardware requirement at an acceptable level. The use of Golay pilots aligns well with Rician channels that have strong line-of-sight components, typical of most mm-wave channels [3]. Golay pilots have the added benefit of an efficient hardware implementation that scales logarithmically rather than linearly with the length of the code. By time-interleaving the pilots, only one Golay correlator is required per antenna. Time interleaving does increase the coherence time overhead associated with the pilots by a factor of K, but since no feedback is required over the downlink, the total overhead can still remain small relative to open-loop beam searches.

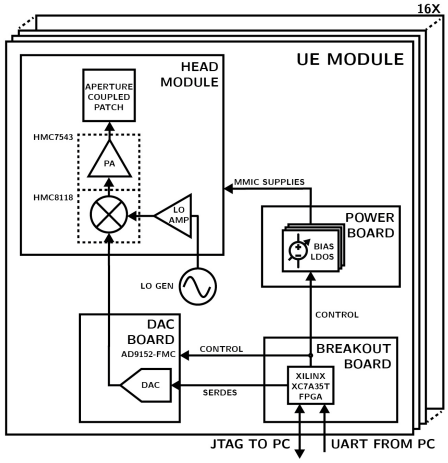

Fig. 2. Block diagram of user equipment module.

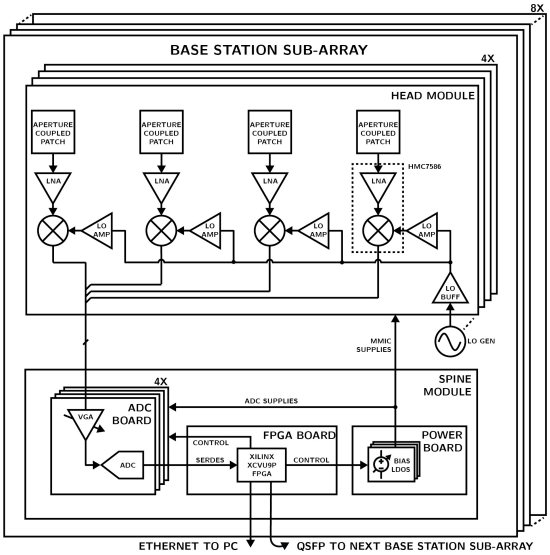

Fig. 3. Block diagram of base-station.

III. Proof-of-Concept Hardware

To explore the feasibility of different mm-wave array architectures, a system that pairs simple mm-wave front ends with data converters to capture long frames of receive samples has been constructed. This testbed enables measurements of real-time data from multiple antennas, allowing us to verify assumptions about the mm-wave channels and the impacts of analog non-idealities. Additionally, different uplink signal processing techniques can be investigated, since any of the captured signals can be processed in software.

A. User Equipment

Since most massive MIMO schemes assume the handset has minimal processing and low directionality, the user equipment (UE) has been designed as a single direct-conversion transmitter driving an aperture-coupled patch antenna. The antenna has gain greater than 0 dBi over at least ± 50 degrees for angles along both the azimuth and elevation to accommodate a large field of view. The UE transmitter is driven by a baseband DAC which repeats a 200µs frame consisting of both channel estimation pilots and single-carrier payload symbols. This frame is stored on a small FPGA, which also handles the configuration of the DAC and transmitter via a JTAG bridge from a PC.

A block diagram of a single UE module is shown in Figure 2.

B. Base Station

The base station receiver array has been designed with the goal of implementing a large-scale digital array using off-the-shelf components. Off-the-shelf E-band radios are expensive, require advanced packaging techniques which may result in low yield, and are comparable in size to their wavelength of $\sim$4mm. These issues all impact the design of larger-scale arrays that use these off-the-shelf components. When designing a PCB for a modular transceiver array, a trade-off in yield and form factor exists. Putting a small number of elements per-board leads to a large area overhead for connectors and peripheral components, which limits the spacing that can be achieved between elements. Alternatively, putting a large number of elements on the PCB can help reduce area overhead, but is more susceptible to yield issues since it will become more and more likely that the packaging/assembly of at least one radio in the module fails.

The unit receiver module designed for this array consists of 4 elements each spaced at 8mm, which corresponds to 2λ spacing at 75 GHz. The same aperture-coupled antenna from the user equipment module is re-used for base station receivers. 16 of these modules have been tiled in a linear array that would nominally consist of 64 elements. However, due to yield issues only 20 elements in the array are functioning, resulting in a sparse, non-uniform array with spacing much greater than the ideal of $\lambda/2$ to implement a beam pattern without grating lobes. While not ideal, this setup still allows for the investigation of different processing techniques and modeling assumptions.

Each direct-conversion receiver is connected to a dual I/Q channel ADC with 8 bits of resolution. These ADCs are grouped into sets of 16 and connected to FPGAs capable of signal processing and capturing an entire 200µs frame from each channel. Each FPGA is daisy-chained to its neighbor via high speed serial lanes, enabling a variety of distributed signal processing algorithms. Each FPGA also has an Ethernet connection to a PC so that a single frame from the entire array can be captured, and any form of distributed or centralized processing techniques can be performed in software on a PC.

A block diagram of the base station is shown in Figure 3.

C. Uplink Processing

The processing algorithms utilized in the base station can be grouped into two sections. A distributed processing section is implemented so that reasonably sized sub-arrays can be daisy-chained in the manner described in Section II-A. After

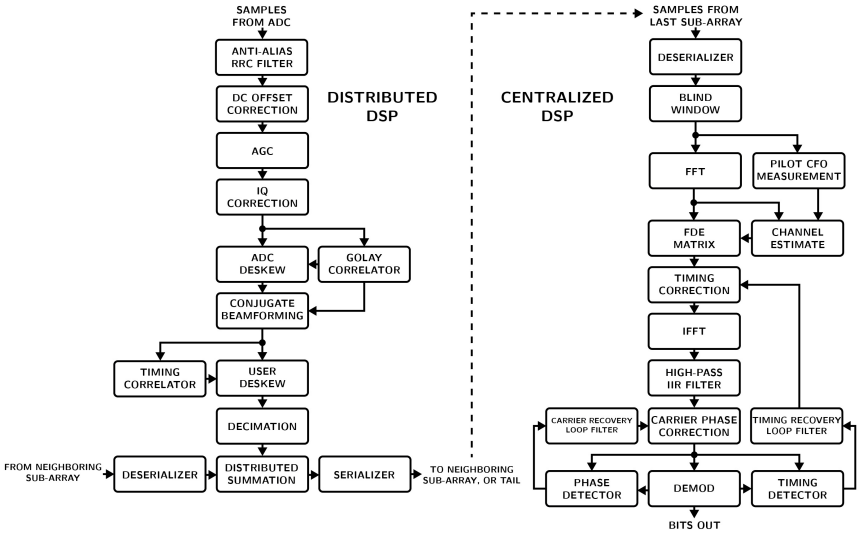

Fig. 4. Distributed sub-array and centralized final stage signal processing algorithms

beam-formed data from every sub-array has been combined to form a set of effective beams for the entire array, a centralized processing section removes inter-user interference, equalizes, and performs carrier/timing recovery.

The ADCs are run at $2\times$ the Nyquist frequency to allow for digital filtering of nearby blockers that may be out-of-band but in-beam. The signal processing blocks before the distributed summation are also ran at $2\times$ oversampling to allow easy implementation of sampling-time deskewing FIR filters.

The distributed processing block, shown in Figure 4, starts with a root-raised-cosine FIR filter to prevent aliasing of out-of-band blockers and noise. The next blocks are the DC offset correction and IQ correction. These blocks are only intended to remove errors from the base station antenna that they follow, and are updated very slowly compared to the update rate of the beam-forming coefficients. After IQ correction, the signals are measured by a Golay correlator to extract magnitude, timing, and phase information about the effective channel by finding the peak cursor in the time domain. Timing skews are averaged across several frames for all users to determine an effective ADC deskewing measurement, which is corrected by a programmable integer sample delay and a 5-tap FIR fractional delay filter. The magnitude and phase information from the correlator are used to determine the coefficients for a frequency-flat MRC sub-array beam-forming matrix. Golay codes and a frequency-flat beam-former were selected to trade off performance in channels with a wide spread of multi-path components for large reductions in hardware. This trade-off is favorable at mm-wave frequencies where multi-path components tend to be heavily attenuated.

After the beam-former, signals are translated from the antenna domain to the user domain. A second set of Golay pilots are used to measure the effective sampling skew of each user. This is a necessary step to properly align the beam-formed signals from the local array to the incoming signals from the the neighboring array. Using pilots to perform this step avoids the need for any measurement/calibration of the delay accumulated in the daisy chain of the array, as well as absorbing any delay spread associated with true time delay across the aperture of the array. This relaxes the need for frequency-dependent beam-forming at high bandwidths, since the criteria for channel delay spread relative to a baseband sample is reduced by a factor equal to the number of sub-arrays. Once the signals are aligned in time using an integer delay and FIR fractional delay filter, they are summed with the data from the neighboring array and passed along the daisy chain.

The centralized processing algorithm which operates on the effective $M \times K$ MRC beam-formed data is shown in Figure 4 as well. The payload symbols for the uplink utilize a frequency domain equalizer (FDE) that requires a block-level cyclic prefix (CP) for the symbols, but encodes the constellations into the time domain symbols as opposed to the frequency domain sub-carriers as used in OFDM. The use of OFDM was avoided primarily because it limits the bandwidth of any carrier recovery algorithms to a maximum of the sub-carrier bandwidth. At mm-wave frequencies, it is easier to remove phase-noise with high-bandwidth carrier recovery loops in the signal processing path than it is to spend the required power in LO synthesis and distribution to make phase noise sufficiently low.

The core of the centralized processing algorithm is the FDE-based zero-forcing matrix. The matrix acts to both eliminate any inter-user interference and equalize any inter-symbol interference associated with the channel. It is important that the zero-forcing matrix is frequency-dependant, as there is no guarantee on the timing and delay spread associated with inter-user interference. The matrix operates on a 256-sample FFT with a 32-sample CP. The window of the FFT is selected blindly, which slightly increases the required length of the CP.

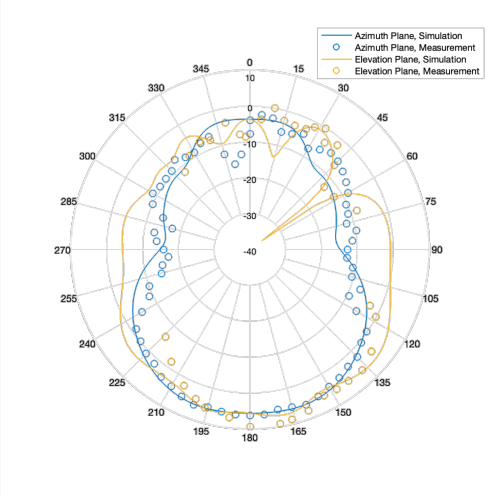

Fig. 5. Simulated vs. measured characteristics for aperture-coupled patch antenna.

This is done because it is not feasible to constrain the skew across the channel between users within a single baseband sample, which means that selecting a truly optimum window across all users is difficult, and it is more reasonable to constrain their synchronization skew within half a cyclic prefix and blindly window. Pilots for the equalizer are sparsely encoded in the sub-carriers and measured by a channel estimator for several FFT blocks at the start of the frame. Time-domain auto-correlation is done on these pilots to measure any impact from carrier frequency offset (CFO) in the pilots, which is taken into account by the channel estimator.

The analog signal path contains an AC-coupling point with a cut-off that is higher in frequency than an FFT bin. Since equalizing the lowest few FFT bins would cause large amounts of noise enhancement due to this cut-off, a high pass IIR filter is implemented after equalization to reduce this noise.

Decision-directed timing and carrier recovery loops are implemented after zero-forcing, since these corrections are time-variant and would reduce the effectiveness of interference cancellation by altering the effective channel. The carrier recovery loop removes any CFO as well as filtering a significant portion of phase noise from the signal. The baseband timing jitter is significantly smaller than a baseband sampling period, so the timing recovery filter can use a relatively small bandwidth and is exclusively used to account for sampling frequency offset (SFO). The correction for the SFO is fed back to the frequency domain, since delays manifest as cyclic within an FFT block due to the CP. After demodulation, BER and SINR measurements are calculated against the expected payload.

IV. Measurement and Validation

A. User Equipment Measurements

The radio characteristics of individual UE modules were measured to establish their transmission specifications. Simulated and measured antenna gain values of the aperture-coupled patch antenna for azimuth and elevation sweeps are shown in Figure 5. We find gain over a large field of view that closely matches with simulated results. Additionally, the transmit power of the UEs was measured with an E-Band power meter. EIRP was measured as a function of distance to verify the measurements were taken at sufficiently far-field distances. After de-embeding path loss and the gain of the power meter antenna, the UEs have an EIRP at the peak of their antenna pattern ranging between 5 and 12dBm, depending on the unit.

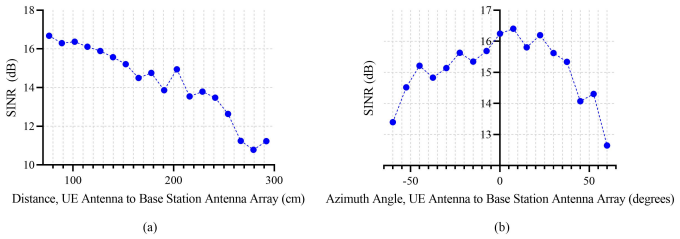

Fig. 6. SINR vs. distance, single UE antenna to base station antenna array. Angle is 0 degrees (broadside).

B. Uplink Measurements

The 20 functioning base station elements are used to implement the architecture described in Section III-B. The functioning elements are randomly distributed across 4 sub-array groupings. Measurements with a single UE demonstrate the automatic beam steering capabilities of the proposed architecture, and two UE measurements show the ability to successfully cancel interference in the presence of real world analog nonidealities and actual mm-wave channels.

Figure 6-a shows SINR for a single UE as a function distance from the base station at a broadside angle. QPSK symbols with a SINR over 10dB are successfully demodulated at a distance of over 3m for a UE only transmitting with 12 dBm of EIRP. Considering that the 20 elements provide up to 13dB of SNR improvement from antenna gain, this means that the pilots are able to provide enough coding gain, even with SNRs below 0 dB at individual antennas.

Figure 6-b shows SINR for a single UE as a function of azimuth angle to the base station panel at a fixed distance of 1.3m. The loss in SINR at this distance follows gain of the individual base station antennas over the ± 50 degree field-of-view, showcasing the ability of the distributed array and per-antenna channel estimation scheme to steer the beam.

Uplink measurements using two UEs have been taken in a variety of positions, and a set of three channels with ranging amounts of isolation have been selected to demonstrate the performance of the system. A sample of frequency domain channel measurements at the input to the FDE-based zero-forcer, which characterise the beam-to-beam isolation and residual multi-path, are shown in Figure 7. The zero-forcer is able to remove the interference and obtain an acceptable SINR for demodulation as shown by the BER values in the

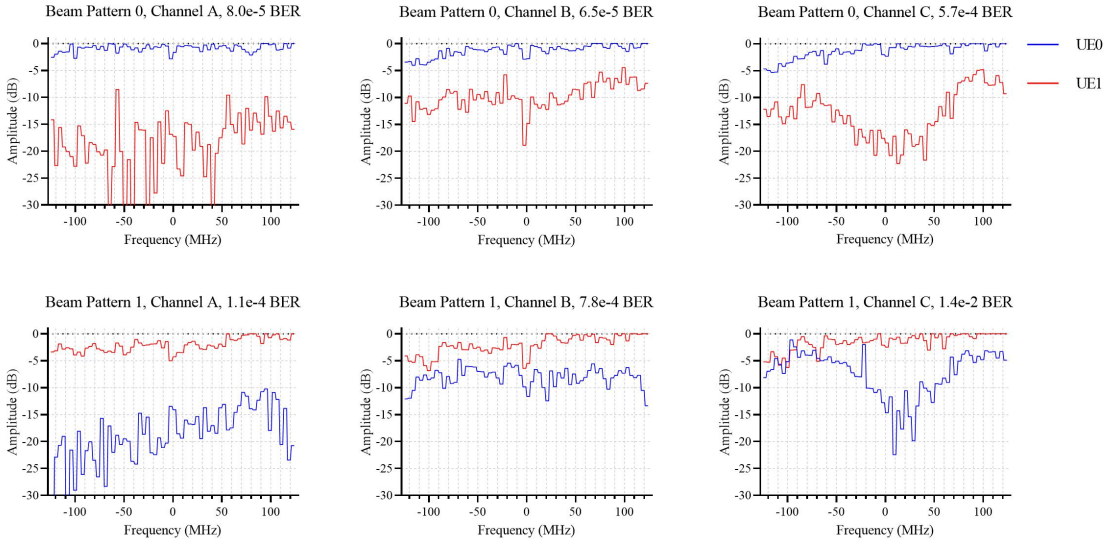

Fig. 7. Channel measurements before interference cancellation for three different channels with the associated BER after cancellation

figure. While two UEs is short of the number of users one would want to classify a system as true massive MIMO, the ability for these algorithms to function successfully on actual hardware in the presence of nonidealities such as phase noise, CFO, SFO, IQ imbalance, inter-user interference, inter-symbol interference, timing skew, and transceiver non-linearities, while having SNRs at individual antennas as low as 0dB is important, since it has been shown using signal processing algorithms designed to map efficiently to practical hardware implementations. Additionally, the fact that this result is achieved with a sparse, non-uniform array and no foreground calibrations to correct for the transciever variations across individual elements highlights the highly scalable and robust nature of the architecture.

V. Conclusion

In this paper, we have analyzed a scalable massive MIMO architecture, demonstrating why distributed signal processing on modular hardware is necessary to realize flexible, energy efficient, multi-gigabit systems. We have implemented proof-of-concept hardware for such a system, with channel measurements for 2 users and 20 base station antennas. The measurements on the system have shown that the architecture is capable of handling a wide variety of real world non-idealities. An extension of this work will implement up and downlink communication. Additionally, we plan to re-implement the system with higher reliability transceivers, and believe that this will allow implementation of up to a 16-user, 128-antenna system supported by this architecture.

Acknowledgment

The authors would like to thank Xilinx and Analog Devices for the donation of hardware. The authors also acknowledge the students, staff, faculty, and sponsors of the Berkeley Wireless Research Center.

References

[1] R. W. Heath, N. González-Prelcic, S. Rangan, W. Roh, and A. M. Sayeed. An overview of signal processing techniques for millimeter wave mimo systems. *IEEE Journal of Selected Topics in Signal Processing*, 10(3):436–453, April 2016.

[2] Junyi Wang, Zhou Lan, Chang-woo Pyo, T. Baykas, Chin-sean Sum, M. A. Rahman, Jing Gao, R. Funada, F. Kojima, H. Harada, and S. Kato. Beam codebook based beamforming protocol for multi-gbps millimeter-wave wpan systems. *IEEE Journal on Selected Areas in Communications*, 27(8):1390–1399, October 2009.

[3] R. Kimura, R. Funada, Y. Nishiguchi, M. Lei, T. Baykas, C. Sum, J. Wang, A. Rahman, Y. Shoji, H. Harada, and S. Kato. Golay sequence aided channel estimation for millimeter-wave wpan systems. In *2008 IEEE 19th International Symposium on Personal, Indoor and Mobile Radio Communications*, pages 1–5, Sep. 2008.

[4] L. Liang, W. Xu, and X. Dong. Low-complexity hybrid precoding in massive multiuser mimo systems. *IEEE Wireless Communications Letters*, 3(6):653–656, Dec 2014.

[5] T. L. Marzetta. Noncooperative cellular wireless with unlimited numbers of base station antennas. *IEEE Transactions on Wireless Communications*, 9(11):3590–3600, November 2010.

[6] H. Prabhu, J. N. Rodrigues, L. Liu, and O. Edfors. 3.6 a 60pj/b 300mb/s 128×8 massive mimo precoder-detector in 28nm fd-soi. In *2017 IEEE International Solid-State Circuits Conference (ISSCC)*, pages 60–61, Feb 2017.

[7] F. Rusek, D. Persson, B. K. Lau, E. G. Larsson, T. L. Marzetta, O. Edfors, and F. Tufvesson. Scaling up mimo: Opportunities and challenges with very large arrays. *IEEE Signal Processing Magazine*, 30(1):40–60, Jan 2013.

[8] Clayton Shepard, Hang Yu, Narendra Anand, Erran Li, Thomas Marzetta, Richard Yang, and Lin Zhong. Argos: Practical many-antenna base stations. In *Proceedings of the 18th Annual International Conference on Mobile Computing and Networking*, Mobicom '12, page 53–64, New York, NY, USA, 2012. Association for Computing Machinery.

[9] Clayton Shepard, Hang Yu, and Lin Zhong. Argosv2: a flexible many-antenna research platform. pages 163–166, 09 2013.

[10] J. Vieira, S. Malkowsky, K. Nieman, Z. Miers, N. Kundargi, L. Liu, I. Wong, V. Öwall, O. Edfors, and F. Tufvesson. A flexible 100-antenna testbed for massive mimo. In *2014 IEEE Globecom Workshops (GC Wkshps)*, pages 287–293, Dec 2014.