
Peer reviewed version

Link to published version (if available):
10.1109/SiPS.2016.54

Link to publication record in Explore Bristol Research

PDF-document

This is the author accepted manuscript (AAM). The final published version (version of record) is available online via IEEE at http://ieeexplore.ieee.org/document/7780108. Please refer to any applicable terms of use of the publisher.

University of Bristol - Explore Bristol Research

General rights

This document is made available in accordance with publisher policies. Please cite only the published version using the reference above. Full terms of use are available: http://www.bristol.ac.uk/red/research-policy/pure/user-guides/ebr-terms/
Serving 22 Users in Real-Time with a 128-Antenna Massive MIMO Testbed

Paul Harris*, Wael Boukley Hasan*, Steffen Malkowsky†, Joao Vieira†, Siming Zhang*, Mark Beach*, Liang Liu†
Evangelos Mellios*, Andrew Nix*, Simon Armour*, Angela Doufexi*, Karl Nieman†, Nikhil Kundargi‡
*Communication Systems & Networks Group, University of Bristol, Bristol, UK
† Dept. of Electrical and Information Technology, Lund University, Sweden
‡Advanced Wireless Research Group, National Instruments, Austin, Texas, USA
Email: {paul.harris, wb14488, ss1659, m.a.beach, evangelos.mellios, andy.nix, simon.armour, angela.doufexi}@bristol.ac.uk
first_name.last_name@{eit.lth.se, ni.com}

Abstract—This paper presents preliminary results for a novel 128-antenna massive Multiple-Input, Multiple-Output (MIMO) testbed developed through Bristol Is Open in collaboration with National Instruments and Lund University. We believe that the results presented here validate the adoption of massive MIMO as a key enabling technology for 5G and pave the way for further pragmatic research by the massive MIMO community. The testbed operates in real-time with a Long-Term Evolution (LTE)-like PHY in Time Division Duplex (TDD) mode and supports up to 24 spatial streams, providing an excellent basis for comparison with existing standards and complimentary testbeds. Through line-of-sight (LOS) measurements at 3.51 GHz in an indoor atrium environment with 12 user clients, an uncoded system sum-rate of 1.59 Gbps was achieved in real-time using a single 20 MHz LTE band, equating to 79.4 bits/s/Hz. In a subsequent indoor trial, 22 user clients were successfully served, which would equate to 145.6 bits/s/Hz using the same frame schedule. To the best of the author’s knowledge, these are the highest spectral efficiencies achieved for any wireless system to date.

Index Terms—Massive MIMO, Testbed, Field Trial, Indoor, 5G

I. INTRODUCTION

Despite promising theoretical results that can be found well documented in [1], [2] and [3], massive MIMO introduces many new, challenging problems when it comes to a real-time system implementation. At the air interface, it is crucially important to coherently measure and track the channel conditions for all active users, where hundreds of signal paths need to be suitably managed and processed at the Base Station (BS) to ensure an acceptable turnaround time (latency) from channel estimation to Downlink (DL) precoding. Prototyping work for various aspects of massive MIMO technology is thus ongoing in many different parts of the world. The ArgosV2 testbed [4] [5] developed at Rice University is a TDD system based upon the Wireless Open-Access Research Platform (WARP) platform and some basic cell capacity predictions have been made by measuring the received Signal to Interference plus Noise Ratio (SINR). A field trail of a TDD massive MIMO prototype by ZTE was reported in [6] where 64 transceiver units served 8 TD-LTE commercial handsets located at different floor levels in a high rise building, achieving a 300 Mbps sum rate in 20 MHz of Bandwidth (BW). In Australia, the Frequency Division Duplex (FDD) Ngara demonstrator [7] uses practical low-cost hardware and was previously reported to have achieved a coded Uplink (UL) spectral efficiency of 67.26 bits/s/Hz in a lab environment at 638 MHz. The Titan massive MIMO platform [8] claims to provide a distributed solution for baseband processing with a large number of Radio Frequency (RF) chains through a two-layer linear processing method, but an implementation using the platform has not yet been reported. And finally, Lund University, also in collaboration with National Instruments (NI), have published numerous papers such as [9] describing their 100-antenna co-located massive MIMO testbed, which utilises the same underlying architecture presented here.

In this paper, we present preliminary results from two indoor field trials for the Bristol Is Open (BIO) massive MIMO research platform, previously introduced in [10], which has been developed within the Communication Systems & Networks (CSN) Research Group at the University of Bristol in close collaboration with both National Instruments and Lund University. The BS prototype consists of 64 NI Universal Software Radio Peripheral (USRP) Reconfigurable Input/Output (RIO) [11] Software-Defined Radios (SDRs) providing 128 RF chains, with a further 6 USRP RIOs acting as 12 single-antenna User Equipments (UEs). An additional 6 USRP clients were used in the second trial to allow the maximum number of users to be increased to 24. A linear array of 128 dipole antennas was deployed at the BS for the first trial and real-time channel measurements were taken simultaneously at 112 active BS antenna ports at 3.51 GHz for 12 transmitting UEs. In the second trial, we tested a new reconfigurable patch antenna array in a 4x32 form factor with alternate H & V polarisations for all 128 chains. The more deterministic orthogonality between spatial streams, also referred to as the channel hardening effect, is presented graphically, and the Singular Value Spread (SVS) for each scenario is discussed. Furthermore, with an UL
dominated frame schedule, we measured a real-time uncoded sum capacity of 1.59 Gbits/s for 12 UEs in the first trial, which equates to a spectral efficiency (SE) of 79.4 bits/s/Hz. For the second trial, we managed to separate 22 users and obtain clear 256-QAM constellations, which would equate to 2.91 Gbits/s and 145.6 bits/s/Hz when using the same frame schedule.

The remainder of the paper is organised as follows: Section II presents an overview of the testbed architecture and the novel Field-Programmable Gate Array (FPGA) implementation of the QR Decomposition and Minimum Mean Square Error (MMSE) detection. The measurement environment, test scenarios and methodologies are then described in Section III, before Section IV presents the throughput results and channel observations. Finally, Section V summarises the work and outlines our future measurement plans.

II. SYSTEM OVERVIEW

The top-level system specifications are provided in Table I. The testbed design has been based on the LTE specifications to a substantial extent; however, there are some major aspects for implementing a TDD based massive MIMO system in which it diverges from the LTE numerology. A TDD massive MIMO architecture exploits the reciprocity of the wireless propagation channel to avoid the large DL channel feedback overhead in an FDD implementation. Instead, orthogonal pilots are transmitted in the UL, and used to estimate the UL channel. This estimated channel is used to compute the DL precoder, which under perfect channel reciprocity is simply the conjugate of the UL decoder matrix.

A. Hardware and Software Architecture

The hardware architecture of the system is shown in Fig. 1. The architecture separates the various functional blocks of the massive MIMO base station into three units - the Remote Radio Head (RRH), the Base Band Unit (BBU) and the Software Controller.

---

TABLE I

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td># of BS Antennas</td>
<td>128</td>
</tr>
<tr>
<td># of UEs</td>
<td>12</td>
</tr>
<tr>
<td>Bandwidth</td>
<td>20 MHz</td>
</tr>
<tr>
<td>Sampling Frequency</td>
<td>30.72 MS/s</td>
</tr>
<tr>
<td>Subcarrier Spacing</td>
<td>15 kHz</td>
</tr>
<tr>
<td># of Subcarriers</td>
<td>2048</td>
</tr>
<tr>
<td># of Occupied Subcarriers</td>
<td>1200</td>
</tr>
<tr>
<td>Frame duration</td>
<td>10 ms</td>
</tr>
<tr>
<td>Subframe duration</td>
<td>1 ms</td>
</tr>
<tr>
<td>Slot duration</td>
<td>0.5 ms</td>
</tr>
<tr>
<td>TDD periodicity</td>
<td>1 slot</td>
</tr>
</tbody>
</table>

---

**RRH:** The RRH is implemented on a USRP RIO [11], each of which has two RF transceiver chains, and an FPGA for implementing the per-antenna processing such as Orthogonal Frequency Division Multiplexing (OFDM) modulation, demodulation, sample re-timing, reciprocity calibration, and IQ impairment correction.

**BBU:** Multiple Kintex 7 FPGAs manufactured by Xilinx are used to implement the BBU unit on NI FlexRIOs [12], which are FPGA coprocessors. The FlexRIO devices perform centralized MIMO processing including channel estimation, MIMO detection, MIMO precoding, and Quadrature Amplitude Modulation (QAM) symbol mapping/demapping.

**Software Controller:** A quad core Intel Xeon processor runs the Software Controller. The upper layer Medium Access Control (MAC) functions, system setup and configuration, FPGA bitfile management, bit source and sink for BBU and visualizations are implemented on the Software Controller. The software architecture of the system is shown in Fig. 2. A massive MIMO implementation imposes two major constraints on the system design - a) massive data transfer and b) massive and rapid matrix computations. This testbed has been designed to seamlessly tackle both of these constraints for a 128 antenna system with up to 20 MHz of channel bandwidth.

**Massive Data Transfer:** Efficient massive MIMO coding and decoding algorithms such as MMSE and Zero-Forcing (ZF) need availability of the per antenna samples at a single centralized location. The 128 antenna system generates 51.6 Gb/s of baseband samples and aggregation and disaggregation of this enormous amount of data into the FPGA coprocessors.
Massive Matrix Computation Massive MIMO performs UL/DL switching within the channel coherence interval to be able to use the UL pilots to compute the Downlink Channel State Information (CSI) at the transmitter. This channel reciprocity based approach needs an ultra fast computation of the channel estimate and generation and multiplication with the precoder matrix on a per subcarrier basis.

B. FPGA Implementation of MIMO Processing

To achieve high system throughput at the BS of 30.72 MS/s $\times$ 128 channels in UL and DL directions, baseband signal processing is performed on FPGAs. The MIMO processing architecture in the system has been designed to use a common MIMO core to produce a linear $12 \times 128$ decoder and $128 \times 12$ encoder for the UL and DL channels respectively. In this implementation, the linear encoder is the Hermitian transpose of the linear decoder. This MIMO core is capable of producing encoders and decoders based on several popular linear MIMO algorithms including MMSE, ZF, and Maximal Ratio Combining (MRC).

In [13], [14], the authors presented an FPGA based architecture to solve for $W_{MMSE}$, the 12x128 detection matrix, using QR decomposition. This testbed extends that algorithm to scale up significantly in throughput for a massive MIMO system. The QR Decomposition based MMSE is computed as shown in Fig. 3. ZF and MRC MIMO decoders are reformulated as special cases of the MMSE and implemented using additional logic that reconfigures the MIMO core to bypass orthogonalization in the MRC case and use a small $\sigma$ parameter in the ZF case.

Due to the high system throughputs and the relatively low FPGA clock rates, much of the processing must be performed in parallel by processing multiple input and/or producing multiple output samples every clock cycle. This type of processing is also known as wide data path processing. Depending on the clock rate and throughput, individual signal processing functions can be implemented in fully parallel (most resources, highest throughput), partial parallel (medium resources, medium throughput), or fully sequential (least resources, least throughput) forms. To relax throughput, the same linear decoder and encoder is used for each resource block (12 subcarriers). In this formulation, the $128 \times 12$ matrix throughput is 16.8 MS/s (LTE subcarrier rate) divided by $12 = 1.4$ Mmatrices/s. To meet this throughput, the matrices are processed at a 200 MHz clock rate using four FlexRIO FPGAs which each perform a partial-parallel (4 path) QR decomposition. An example of a $N \times 12$ systolic array QR decomposition based on modified Gram Schmidt for full-parallel, partial-parallel, and full-sequential is shown in Fig. 5 and described in [13], [14].

Fig. 5 shows three different systolic array formulations of the $N \times 12$ QR decomposition. Three different arrays are shown for full-parallel (12 path), partial-parallel (4 path), and full-sequential (1 path). Each formulation reuses three common nodes “orthogonalize”, “normalize”, and “orthogonalize and normalize”. To balance throughput requirements and resource utilization, the implementation for the $128 \times 12$ QR decomposition is chosen to be the four path partial parallel formulation. After the $W_{MMSE}$ and $W^H_{MMSE}$ decoder and encoder matrices are computed for the full bandwidth (100 resource blocks), UL subcarriers are decoded and DL symbols are encoded. Fig. 4 shows a partial parallel (4 path) implementation of the linear decoder, where a $12 \times 128$ by $128 \times 1$ matrix vector multiplication is performed in thirty two 200 MHz clock cycles by decomposing the operations into four parallel submatrices and subvectors, and summing the result.

C. Frame Schedule

The PHY frame schedule used for the system is shown in Fig. 6. In its default configuration, it is based closely upon the TD-LTE standard, but it can be completely customised at the OFDM symbol level to allow a range of configurations to be applied for different applications. For the throughput results shown in this paper, the schedule was configured to use UL only transmission with two pilots per radio frame.
Cabled synchronisation was used with a network of Octoclock distribution modules [15]. The BS utilises 8 of these to phase align the USRP Local Oscillators (LOs) by redistributing a 10 MHz clock with 50 parts-per-billion (ppb) accuracy provided by a PXie-6674T timing card [16]. To achieve sample alignment, a start trigger that signals the start of the first radio frame is fanned out to all radios by equal length cables. The user clients were connected to a 9th Octoclock by extended cable and the timing offsets were corrected for, but future deployments will use Over-the-air (OTA) synchronisation.

III. EXPERIMENT OVERVIEW

A. First Measurement Environment

The length of the lower atrium in the University of Bristol’s Merchant Venturers Building was used for three different LOS measurements between the BS and 12 UEs. UEs were grouped both in a straight line parallel to the BS and at a slant, with a distance of 3.3m, 12.5m or 18.1m to the nearest client in each scenario. At the BS side, a 128 element linear array of length 5.44m was used as shown in Fig. 7, providing half-wavelength spacing at 3.5 GHz. A floor plan of the experiment is shown in Fig. 8 with the UE locations for each of the three scenarios. Measurements were performed outside of university hours to ensure the environment remained as static as possible.

B. Second Measurement Environment

The upper level of the Merchant Venturers Building atrium was used with a patch panel antenna array to serve 22 user clients placed 24.8m away on the opposite balcony. The array was setup in a 4x32 configuration with alternate H & V polarisations for all 128 antennas. As with the first trial, the UEs were in LOS and placed in a straight line with 2.5 λ spacing. However, this environment was not so static, as it was a normal working day and students were present. An overview of the setup can be seen in Fig. 9.

C. System Configuration

As the environments were relatively static, the frame schedule was configured such that channel estimation was performed in 5ms intervals, and all remaining slots were assigned for UL data with a 256-QAM Modulation and Coding Scheme (MCS). ZF detection was used for all throughput measurements and equal transmit levels were fixed for all UEs. 100 channel data captures were recorded to disk for each scenario at an interval of approximately 200ms, resulting in a total measurement period of approximately 3 minutes for each run. In the second trial where the environment was a little more dynamic, the disk capture interval was reduced to 50ms.

D. Channel Processing

For each 12 user scenario, the capture files contain the raw 12x128 channel matrix $H$ for all 100 resource blocks of 180 kHz (12 subcarriers). When processing 24 users, the capture files contain a 24x128 matrix for 50 resource blocks of 360 kHz (24 subcarriers) due to the difference in pilot allocation.
To quantify the joint orthogonality between users, we obtain the SVS. This is found by first performing the Singular Value Decomposition (SVD)

$$H_{n,t} = U_{n,t} S_{n,t} V_{n,t}^H$$

where for resource block $n$ and acquisition time $t$, $U_{n,t}$ and $V_{n,t}$ represent the left and right unitary matrices, and $S_{n,t}$ is a diagonal matrix containing the singular values. The SVS is then defined as the ratio of the largest to the smallest singular values of $H$:

$$\gamma_{n,t} = \frac{\max \lambda_1, \lambda_2, ..., \lambda_i}{\min \lambda_1, \lambda_2, ..., \lambda_i}$$

and a single SVS value for each capture is obtained by averaging across resource blocks

$$\tilde{\gamma}_t = \frac{\sum_n \gamma_{n,t}}{N}$$

where $N$ is the total number of resource blocks. Cumulative Distribution Function (CDF) plots are then produced using the resource block averaged SVS values for all captures. This is performed in each scenario of trial one for 32, 64 and 112 elements by truncating the antenna dimension centrally. In addition to the SVS, the 12x12 and 24x24 $HH^H$ symmetric matrices were calculated and averaged across all subcarriers and captures for both trials, providing an alternative way of visualising the UE pairwise orthogonality.

IV. PRELIMINARY RESULTS

CDF plots for the resource block averaged SVS values obtained for each scenario in trial one with 32, 64 and 112 antenna elements are shown in Fig. 10. From 2.6 GHz measurements with a 128-element virtual linear array, [17] provides SVS statistics for closely spaced users in outdoor LOS conditions and compares these to the independent and identically distributed (iid) Rayleigh channel ideal of approximately 2.5 decibels (dB). For 32 antennas, the median SVS from these measurements with users spaced 13-17 wavelengths apart was approximately 10 dB. This most closely matches with scenario 2 where our UE separation was only 6 wavelengths. In the 128-antenna case, the median SVS in [17] only drops to 9 dB, where as all three of our measurement scenarios show lower values with only 112 antennas. The above two observations imply that we are experiencing a richer scattering environment indoors, even under LOS conditions. Furthermore, despite LOS conditions in the 3.3m scenario, the angular arrival spread between elements of the array for each UE will be far higher than the 18.1m scenario due to the extreme length of the linear array, allowing us to achieve a very low median SVS of 3.3 dB. Our client separation was only 2.5-6 wavelengths and 16 fewer antennas were used, indicating that extremely good spatial separation can be achieved using massive MIMO in indoor LOS settings, even when users are closely clustered. However, whilst both the 112 and 64 elements cases perform best in Scenario 1, the 32 antenna case appears in reverse, with Scenario 1 providing the worst median SVS by nearly 2 dB. This likely indicates that close range LOS separation is more challenging with the 32 element array as it cannot exploit the angular arrival spread to the same extent. In addition, the difference between median SVS for Scenario 2 and 3 in the 112 element case is 1.3 dB, whereas these are closely aligned in the 32 and 64 element cases. This shows that even at range, the 112 element array was able to exploit the greater spread in angular LOS arrival expected for Scenario 2 over Scenario 3. In trial two, the median SVS from 100 captures (5 seconds) was 6.5 dB, which agrees closely with the most distant scenario of the first trial (scenario 3). However, due to the greater dynamics of the environment, the SVS fluctuated between 6 and 7 dB during the course of the measurement.

To visualise the pair-wise orthogonality, Fig. 11 shows intensity plots of $HH^H$ in Scenario 2 for 32, 64 and 112 antennas, with a dB scale relative to the maximum value. Each plot is averaged across all subcarriers and acquisitions. On inspection of the plots, a strengthening diagonal can certainly be seen as the antenna numbers increase. Interestingly, each pair of streams appear to interfere with one another slightly, even when the antenna numbers are increased. This can be explained by the fact that those two streams correspond to the two antennas on each USRP that have a fixed separation explained by the fact that those two streams correspond to the two antennas on each USRP that have a fixed separation distance of 13cm. Thus, even with 50cm separation between USRPs, we should expect these to be more spatially correlated. As the number of antenna elements are increased, the cluster of user correlation seen in the top right for 32 elements begins to flatten out, and by 112 elements all the off-diagonal values excluding the user pairs are at least 10 dB below the diagonal. For some streams, the level of correlation is far below this, with user 3 experiencing a rejection of 25 dB with respect to users 1 and 5. In trial two where 22 users were served, the level of rejection was even higher, with the majority of the off diagonal elements falling to -30 dB relative to the eigenvalues or less as shown in Fig. 12. This could indicate that the alternate polarisations allowed multipath scattering to be exploited to a greater degree, but further measurements would be required to confirm this.

The received constellations for all twelve streams trial one (scenario 2) and the 22 streams in trial two are shown in
Fig. 11. \( \mathbf{H} \) intensity plots for increasing antenna numbers in the first trial. 12 UEs in LOS, 12.5m from BS with 50 cm separation.

Fig. 12. \( \mathbf{H} \) intensity plot for 22 UEs in LOS in the second trial: 24.8m from BS with 2.5 \( \lambda \) separation.

TABLE II

<table>
<thead>
<tr>
<th>Scenario 1</th>
<th>Scenario 2</th>
<th>Scenario 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Throughput</td>
<td>1.59 Gbps</td>
<td>1.58 Gbps</td>
</tr>
<tr>
<td>Spectral Efficiency</td>
<td>79.4 bits/s/Hz</td>
<td>79 bits/s/Hz</td>
</tr>
</tbody>
</table>

Fig. 13 and Fig. 14 respectively. These are each extremely clean, further illustrating the high level of orthogonality we observed when analysing the channel. Whilst not tested in this initial trial, it shows hope for the effective use of simple MRC detection and Maximal Ratio Transmission (MRT) precoding with more BS antennas as we leverage the natural pair-wise orthogonality in \( \mathbf{H} \).

For each scenario in trial one, the uncoded UL throughput was measured in real-time by transmitting a known pseudo-random byte sequence from each UE. The BS calculates the sum rate for a given radio frame using

\[
\text{Sum rate} = 100 \times \left( \text{Max} - \sum_{i=0}^{11} E_i \right) \text{bits/s} \quad (4)
\]

where \( E_i \) is the number of bit errors for stream \( i \) and Max is the aggregate bit total for all streams in an ideal error-free frame. For our frame schedule, \( \text{Max} = 1200 \text{ subcarriers} \times 8 \text{ bits} \times 138 \text{ symbols} \times 12 \text{ users} = 15897600 \text{ bits} \). 112 of the 128 elements were used in each case due to a hardware limitation at that time and the results are summarised in Table II. With the frame schedule applied, we could expect a maximum UL throughput of 1.589 Gbps, corresponding to 132.4 Mbps per stream. Scenario 1, where we have the lowest SVS, comfortably achieves this, and only drops slightly for the more distant scenarios. It is evident that the large number of antenna elements greatly improves user orthogonality for linear capacity growth. The absolute throughput could not be measured for 22 users as only decimated host detection was performed, but by observing the clarity of the constellations and the degree of spatial orthogonality achieved, an appropriate estimation could be made. With the same frame schedule applied in the 12 user case, the 22 user result would scale to 2.91 Gbps and 145.6 bits/s/Hz.

V. CONCLUSIONS AND FUTURE WORK

We have introduced a flexible 128-antenna massive MIMO testbed that can perform up to 128x24 MIMO processing in real-time on a 20 MHz LTE band. The performance in an indoor LOS environment was evaluated using both a physically large linear array and a patch panel array, and a spectral efficiency of 79.4 bits/s/Hz was achieved using uncoded transmissions. To the best of the author’s knowledge, this is the highest spectral efficiency reported thus far for any system in the world [18] [19]. Future work will include real-time DL performance evaluation, implementation of massive
Fig. 13. 256-QAM UL constellations from 12 UEs in trial one: 12.5m with 50cm client separation.

Fig. 14. 256-QAM UL constellations from 22 UEs in trial two: 24.8m with 50cm client separation.

MIMO optimised power control and OTA synchronisation, rooftop deployment and node distribution on the BIO citywide fibre network.

ACKNOWLEDGMENT

The authors wish to thank Bristol Is Open for access to the hardware facility. They also acknowledge the financial support of the Engineering and Physical Sciences Research Council (EPSRC) Centre for Doctoral Training (CDT) in Communications (EP/I028153/1), NEC and NI.

REFERENCES