On the Achievable Rates of Decentralized Equalization in Massive MU-MIMO Systems

Charles Jeon, Kaipeng Li, Joseph R. Cavallaro, and Christoph Studer

***Abstract*—Massive multi-user (MU) multiple-input multiple-output (MIMO) promises significant gains in spectral efficiency compared to traditional, small-scale MIMO technology. Linear equalization algorithms, such as zero forcing (ZF) or minimum mean-square error (MMSE)-based methods, typically rely on centralized processing at the base station (BS), which results in (i) excessively high interconnect and chip input/output data rates, and (ii) high computational complexity. In this paper, we investigate the achievable rates of decentralized equalization that mitigates both of these issues. We consider two distinct BS architectures that partition the antenna array into clusters, each associated with independent radio-frequency chains and signal processing hardware, and the results of each cluster are fused in a feedforward network. For both architectures, we consider ZF, MMSE, and a novel, non-linear equalization algorithm that builds upon approximate message passing (AMP), and we theoretically analyze the achievable rates of these methods. Our results demonstrate that decentralized equalization with our AMP-based methods incurs no or only a negligible loss in terms of achievable rates compared to that of centralized solutions.**

I. Introduction

Massive MU-MIMO is widely believed to be a key technology for next-generation wireless systems [1]. By equipping the BS with hundreds or thousands of antennas and serving tens or hundreds of users simultaneously in the same time-frequency resource, massive MU-MIMO enables orders-of-magnitude improvements in spectral efficiency compared to traditional, small-scale MIMO [2]. However, the presence of hundreds or thousands of antenna elements at the BS causes significant implementation challenges of this technology.

One of the most critical implementation challenges is the excessively high amount of data that must be transferred between the BS antenna array and the baseband processing unit. For example, the raw baseband data rates (from or to the RF chains) exceed 200 Gbit/s for a MU-MIMO system with 128 BS antennas each using two 10 bit analog-to-digital converters operating with 40 MHz bandwidth. Such high data rates cannot be sustained by existing interconnect technology and chip input/output (I/O) interfaces. Furthermore, existing conventional linear equalization algorithms, such as ZF and MMSE-based methods, rely on centralized processing and require excessively high computational complexity [3]. Hence, existing massive MU-MIMO testbeds either distribute baseband processing in the frequency domain [4], i.e., perform parallel computations per subcarrier, or use maximum ratio combining (MRC) [5], which enables fully distributed processing at the antenna elements. Frequency distribution, however, requires that each frequency cluster needs access to all the BS antennas, which prevents a scaling to thousands of antennas. MRC is known to result in poor spectral efficiency for realistic antenna configurations [6]. Consequently, a practical deployment of massive MU-MIMO necessitates novel equalization methods that reduce the interconnect bandwidth and baseband processing complexity while maximizing the spectral efficiency.

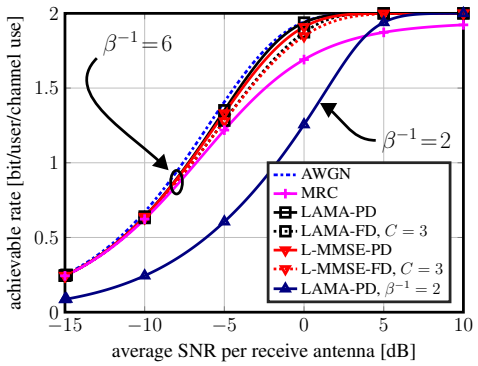


Fig. 1. Achievable rates for QPSK in the large-system limit for $\beta^{-1} = B/U = 6$. Fully decentralized LAMA and L-MMSE with $C = 3$ clusters only suffer a minimal rate loss compared to the fundamental AWGN performance limit. MRC, which is decentralized by nature, performs poorly at higher rates.

A. Contributions

We propose two decentralized BS architectures, namely partially decentralized (PD) and fully decentralized (FD) equalization, which mitigate the interconnect, I/O, and computation bottlenecks. For both of these architectures, we investigate the performance of ZF, linear MMSE (L-MMSE), and a novel, non-linear decentralized equalization method that builds upon our recently proposed large MIMO approximate message passing (LAMA) algorithm [7]. We develop a state-evolution (SE) framework that enables a precise analysis of the achievable rates and error-rate performance of decentralized equalization in the large-system limit, and we show simulation results for realistic, finite-dimensional systems that agree with our theory.

Fig. 1 demonstrates that FD equalization with $C = 3$ antenna clusters in combination with LAMA achieves rates that are close to that of an interference-free AWGN channel even for moderate BS-to-user-antenna ratios. FD equalization achieves significantly higher rates than PD equalization with a lower BS-to-user-antenna ratio, which demonstrates that higher spectral efficiency can be achieved through decentralized architectures that reduce the interconnect and chip I/O bandwidths.

CJ and CS are with the School of ECE at Cornell University, Ithaca, NY; e-mails: jeon@csl.cornell.edu, studer@cornell.edu.

KL and JRC are with the Department of ECE at Rice University, Houston, TX; e-mails: kl33@rice.edu, cavallar@rice.edu.

The work of CJ and CS was supported by Xilinx, Inc. and by the US NSF under grants ECCS-1408006, CCF-1420328, and CAREER CCF-1652065. The work of KL and JRC was supported by Xilinx, Inc. and by the US NSF under grants CNS-1265332, ECCS-1232274, and ECCS-1408370.

B. Relevant prior art

Architectures that perform decentralized processing in the spatial domain have been proposed in [3], [8].[1] The idea is to partition the BS antenna array into C independent clusters, each associated with local computing hardware. Equalization and beamforming is then carried out in an iterative fashion by exchanging consensus information among the clusters. While these iterative methods significantly reduce the raw baseband data rates and the computation bottlenecks, their performance has not been analyzed and the throughput suffers from interconnect latency. In contrast, we focus on decentralized *feedforward* architectures whose performance can be analyzed theoretically and is less susceptible to interconnect latency.

One of the proposed decentralized equalization algorithms in this paper builds upon AMP [11], [12]. Centralized equalization via AMP was shown to achieve near individually-optimal (IO) performance for realistic massive MU-MIMO systems [7], [13]. A distributed version of AMP was proposed recently in [14]. The key differences to this method are as follows: (i) we consider feedforward architectures, which are key for low-latency processing as required by next-generation massive MU-MIMO systems [2], and (ii) we analyze the achievable rates and error-rate performance in massive MU-MIMO systems.

C. Notation

Lowercase and uppercase boldface letters designate vectors and matrices, respectively; uppercase calligraphic letters denote sets. The transpose and Hermitian of the matrix $\mathbf{H}$ are represented by $\mathbf{H}^{\mathrm{T}}$ and $\mathbf{H}^{\mathrm{H}}$. We define $\langle \mathbf{x} \rangle = \frac{1}{N}\sum_{k=1}^{N} x_k$. The multivariate complex-valued Gaussian probability density function (pdf) with mean $\mathbf{m}$ and covariance $\mathbf{K}$ is denoted by $\mathcal{CN}(\mathbf{m}, \mathbf{K})$. $\mathbb{E}_X[\cdot]$ and $\mathbb{V}\mathrm{ar}_X[\cdot]$ represent the mean and variance with respect to the random variable X, respectively.

II. Decentralized Equalization Architectures

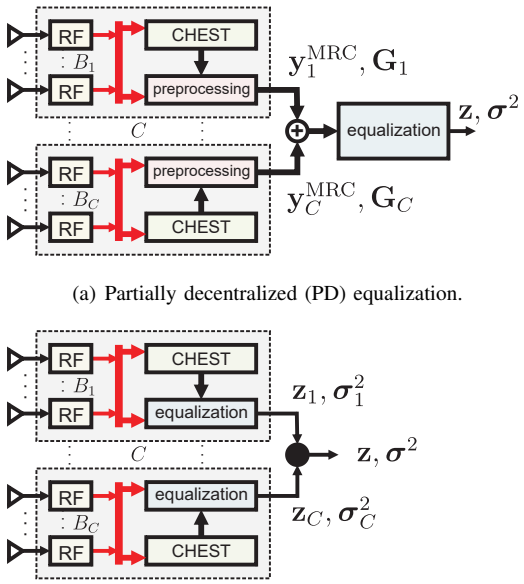
We start by proposing two feedforward architectures depicted in Fig. 2 that enable decentralized equalization and achieve (often significantly) higher spectral efficiency than MRC-based architectures that naturally enable distributed processing [2].

A. System model for decentralized equalization

We model the input-output relation of a massive MU-MIMO uplink system by $\mathbf{y} = \mathbf{H}\mathbf{s}_0 + \mathbf{n}$. Here, $\mathbf{y} \in \mathbb{C}^B$ is the receive vector and B denotes the number of BS antennas, $\mathbf{H} \in \mathbb{C}^{B \times U}$ is the known MIMO system matrix where each element of $\mathbf{H}$ is distributed $\mathcal{CN}(0, 1/B)$ and $U \leq B$ denotes the number of users, $\mathbf{s}_0 \in \mathcal{O}^U$ contains the transmit symbols for each user, $\mathcal{O}$ is the constellation set (e.g., QPSK), and $\mathbf{n} \in \mathbb{C}^B$ is i.i.d. circularly symmetric complex Gaussian noise with variance N_0 per complex entry. We assume an i.i.d. prior $p(\mathbf{s}_0) = \prod_{\ell=1}^{U} p(s_{0\ell})$ and each symbol is distributed as:

$$p(s_{0\ell}) = \tfrac{1}{|\mathcal{O}|}\textstyle\sum_{a \in \mathcal{O}} \delta(s_{0\ell} - a), \qquad (1)$$

[1]Distributed processing is also a key component of coordinated multipoint (CoMP) [9] and cloud radio access networks (CRANs) [10] for multi-cell transmission. The decentralized architectures and algorithms in [3], [8] are specifically designed for massive MU-MIMO systems in which the computing hardware is collocated near the antenna array and within a single cell.



(a) Partially decentralized (PD) equalization.

(b) Fully decentralized (FD) equalization.

Fig. 2. Partially decentralized (PD) and fully decentralized (FD) equalization architectures for the massive MU-MIMO uplink with C clusters. (a) PD performs decentralized channel estimation (CHEST) and preprocessing; equalization is performed in a centralized fashion and operates on low-dimensional data. (b) FD performs CHEST, preprocessing, and equalization in a decentralized manner. The $\oplus$ operator in (a) denotes matrix/vector-additions and $\bullet$ in (b) denotes a weighted vector addition (see Section IV for the details).

where $|\mathcal{O}|$ is the cardinality of $\mathcal{O}$ and $\delta(\cdot)$ is the Dirac delta function. The average symbol energy is $E_s = \mathbb{E}\big[|s_{0\ell}|^2\big]$.

As in [3], [8], we partition the B BS antennas into $C \geq 1$ independent *antenna clusters*. The cth cluster is associated with $B_c = w_c B$ BS antennas so that $w_c \in (0, 1]$ and $\sum_{c=1}^{C} w_c = 1$. Each cluster contains local radio-frequency (RF) components and only requires access to local channel state information (CSI). Hence, the receive vector for cluster c can be written as $\mathbf{y}_c = \mathbf{H}_c\mathbf{s}_0 + \mathbf{n}_c$ with $\mathbf{y}_c \in \mathbb{C}^{B_c}$, $\mathbf{H}_c \in \mathbb{C}^{B_c \times U}$, and $\mathbf{n}_c \in \mathbb{C}^{B_c}$. Without loss of generality, we assume the following partitioning: $\mathbf{y} = [\mathbf{y}_1^{\mathrm{T}} \cdots \mathbf{y}_C^{\mathrm{T}}]^{\mathrm{T}}$ and $\mathbf{H} = [\mathbf{H}_1^{\mathrm{T}} \cdots \mathbf{H}_C^{\mathrm{T}}]^{\mathrm{T}}$.

B. Partially decentralized (PD) equalization architecture

The partially decentralized (PD) equalization architecture is illustrated in Fig. 2(a) for C clusters. Each cluster c independently preprocesses the partial received vector $\mathbf{y}_c$ and channel matrix $\mathbf{H}_c$ by computing the partial MRC vector $\mathbf{y}_c^{\mathrm{MRC}} = \mathbf{H}_c^{\mathrm{H}}\mathbf{y}_c$ and the partial Gram matrix $\mathbf{G}_c = \mathbf{H}_c^{\mathrm{H}}\mathbf{H}_c$. A feedforward adder tree is used to compute the complete MRC vector and Gram matrix, i.e., computes $\mathbf{y}^{\mathrm{MRC}} = \sum_{c=1}^{C} \mathbf{y}_c^{\mathrm{MRC}}$ and $\mathbf{G} = \sum_{c=1}^{C} \mathbf{G}_c$. Since the MRC output is a sufficient statistic for the transmit signal [15], we perform (linear or non-linear) equalization in a centralized manner and compute a soft symbol $\mathbf{z} \in \mathbb{C}^U$ and variance $\boldsymbol{\sigma}^2 \in \mathbb{C}^U$ vector, which can be used to either compute hard-output estimates or soft information (e.g., in the form of log-likelihood ratios) of the transmitted bits [16]. In Section III, we will analyze the performance of PD equalization for ZF, L-MMSE, and a novel LAMA-based equalization algorithm all of which directly operate on the combined MRC vector $\mathbf{y}^{\mathrm{MRC}}$ and Gram matrix $\mathbf{G}$.

C. Fully decentralized (FD) equalization

The PD architecture requires a summation of both the partial MRC vector and the partial Gram matrices, which involves the transfer and processing of large amounts of data in the adder tree. The fully decentralized (FD) equalization architecture illustrated in Fig. 2(b) often significantly reduces the overhead of data fusion at the cost of lower performance. Specifically, each cluster c independently performs CHEST, preprocessing, *and* equalization, and directly computes a soft symbol $\mathbf{z}_c \in \mathbb{C}^U$ and variance $\boldsymbol{\sigma}_c^2 \in \mathbb{C}^U$ vector. The fusion tree optimally combines the resulting soft symbols $\mathbf{z}_c$ and variance $\boldsymbol{\sigma}_c^2$ vectors in order to generate the final output tuple $\{\mathbf{z}, \boldsymbol{\sigma}^2\}$ used for hard- or soft-output detection; see Section IV for the details.

III. Partially Decentralized (PD) Equalization

We start by presenting a decentralized LAMA algorithm suitable for the PD architecture and the associated SE framework. We then adapt the well-known Tse-Hanly equations [17] to characterize the performance of PD equalization with linear equalization algorithms, such as MRC, ZF, and L-MMSE.

A. LAMA for PD equalization

The LAMA algorithm [7] operates on the conventional input-output relation $\mathbf{y} = \mathbf{H}\mathbf{s}_0 + \mathbf{n}$. We next propose a novel variant that directly operates on the MRC output $\mathbf{y}^{\text{MRC}}$ and the Gram matrix $\mathbf{G}$, i.e., the outputs of the fusion tree of the PD architecture shown in Fig. 2(a). Note that since the antenna configuration in massive MU-MIMO systems typically satisfies $U \ll B$, the LAMA-PD algorithm summarized next operates on a lower-dimensional problem while delivering exactly the same results as the original algorithm in [7].

Algorithm 1. *Initialize $s_\ell = \mathbb{E}_S[S]$ for $\ell = 1, \ldots, U$, $\phi^1 = \mathbb{V}\text{ar}_S[S]$, and $\mathbf{v}^1 = \mathbf{0}$. Then, for every algorithm iteration $t = 1, 2, \ldots$, compute the following steps:*

$$\begin{aligned} \mathbf{z}^t &= \mathbf{y}^{\text{MRC}} + (\mathbf{I} - \mathbf{G})\mathbf{s}^t + \mathbf{v}^t \\ \mathbf{s}^{t+1} &= \mathsf{F}(\mathbf{z}^t, N_0 + \beta\phi^t) \\ \phi^{t+1} &= \langle \mathsf{G}(\mathbf{z}^t, N_0 + \beta\phi^t) \rangle \\ \mathbf{v}^{t+1} &= \tfrac{\beta\phi^{t+1}}{N_0 + \beta\phi^t}(\mathbf{z}^t - \mathbf{s}^t). \end{aligned} \tag{2}$$

The functions $\mathsf{F}(s_\ell, \tau)$ and $\mathsf{G}(s_\ell, \tau)$ correspond to the message mean and variance, respectively, and are computed as follows:

$$\begin{aligned} \mathsf{F}(z_\ell, \tau) &= \textstyle\int_{s_\ell} s_\ell f(s_\ell | \hat{z}_\ell) \mathrm{d}s_\ell \\ \mathsf{G}(z_\ell, \tau) &= \textstyle\int_{s_\ell} |s_\ell|^2 f(s_\ell | \hat{z}_\ell) \mathrm{d}s_\ell - |\mathsf{F}(z_\ell, \tau)|^2. \end{aligned} \tag{3}$$

Here, $f(s|z)$ is the posterior pdf $f(s|z) = \frac{1}{Z} p(z|s)p(s)$ with $p(z|s) \sim \mathcal{CN}(s, \tau)$, $p(s)$ is given in (1), and Z is a normalization constant.

To analyze the performance of LAMA-PD using the SE framework, we need the following definition.

Definition 1 (Large-system limit)**.** *For a MIMO system with U user antennas and B BS antennas, we define the large-system limit by fixing the system ratio $\beta = U/B$ and letting $U \to \infty$.*

We also need the following decoupling property of LAMA. In the large-system limit and for every iteration t, (2) is distributed according to $\mathcal{CN}(\mathbf{s}_0, \sigma_t^2 \mathbf{I}_U)$ [7]; this shows that LAMA decouples the MIMO system into U parallel and independent AWGN channels each with noise variance σ_t^2. The SE framework in Theorem 1 with proof in [12] allows us to track the decoupled noise variance σ_t^2 in each iteration t.

Theorem 1. *Fix the system ratio $\beta = U/B$ and the signal prior in (1). In the large-system limit, the decoupled noise variance σ_t^2 of LAMA at iteration t is given by the recursion:*

$$\sigma_t^2 = N_0 + \beta\Psi(\sigma_{t-1}^2). \tag{4}$$

Here, the mean-squared (MSE) function is defined by

$$\Psi(\sigma_{t-1}^2) = \mathbb{E}_{S,Z}\Big[\big|\mathsf{F}(S + \sigma_{t-1}Z, \sigma_{t-1}^2) - S\big|^2\Big], \tag{5}$$

where F is given in (3), $S \sim p(s)$ as in (1), $Z \sim \mathcal{CN}(0,1)$, and σ_1^2 is initialized by $\sigma_1^2 = N_0 + \beta\,\mathbb{V}\text{ar}_S[S]$.

For $t \to \infty$, the recursion (4) converges to the fixed-point equation $\sigma_{\text{PD}}^2 = N_0 + \beta\Psi(\sigma_{\text{PD}}^2)$. If there are multiple fixed points, then we select the largest σ_{PD}^2, which is, in general, a sub-optimal solution [18]. Since in the large-system limit the MIMO system is decoupled into AWGN channels with noise variance σ_{PD}^2 for each user, we will use this fixed-point equation to analyze the achievable rates and error-rate performance of decentralized equalization in Section V.

B. Linear algorithms for PD equalization

As for the LAMA-PD algorithm, linear data detectors are also able to operate directly with the MRC output and the Gram matrix. For MRC equalization, we can directly use the MRC output $\mathbf{y}^{\text{MRC}}$. For ZF and L-MMSE equalization, we first compute a $U \times U$ filter matrix $\mathbf{W} = (\mathbf{G} + \alpha\mathbf{I}_U)^{-1}$, where α is set to 0 and N_0/E_s for ZF and L-MMSE, respectively. The final linear estimate $\mathbf{z}$ is then computed by $\mathbf{z} = \mathbf{W}\mathbf{y}^{\text{MRC}}$.

In the large-system limit, the output of MRC, ZF, and L-MMSE-based equalization is also decoupled into AWGN channels [17] with noise variance σ_{PD}^2 for each user U. Closed-form expression for the noise variance have been developed by Tse and Hanly in [17], and are as follows.

Theorem 2. *Fix the system ratio $\beta = U/B$. In the large-system limit, the decoupled noise variance σ_{PD}^2 for MRC, ZF, and L-MMSE is a fixed-point solution to $\sigma_{PD}^2 = N_0 + \beta\Psi(\sigma_{PD}^2)$, where $\Psi(\sigma^2)$ equals to $\mathbb{V}\text{ar}_S[S]$, σ^2, and $\frac{\mathbb{V}\text{ar}_S[S]}{\mathbb{V}\text{ar}_S[S]+\sigma^2}\sigma^2$ for MRC, ZF, and L-MMSE equalization, respectively.*

IV. Fully Decentralized (FD) Equalization

We now discuss the algorithm aspects of the FD architecture shown in Fig. 2(b) and then analyze its performance.

A. Algorithm procedure for FD architecture

Recall from Section II-C that each cluster c in the FD architecture independently computes the vectors $\mathbf{z}_c$ and $\boldsymbol{\sigma}_c^2$. Once equalization for all C clusters is completed, then the vectors $\mathbf{z}_c$ and $\boldsymbol{\sigma}_c^2$ must be fused to compute the output $\{\mathbf{z}, \boldsymbol{\sigma}^2\}$. Since the input-output relation from each cluster is decoupled into an AWGN system with i.i.d. noise in the large-system limit, optimal fusion corresponds to computing a weighted sum of $\sum_{c=1}^{C} \nu_c \mathbf{z}_c$ that minimizes the output noise variance σ_{FD}^2.

B. Equalization performance in FD architecture

The following result characterizes the performance of FD for each cluster c; the proof is given in Appendix A-A.

Theorem 3. *Let cluster c have a system ratio of $\beta_c = U/B_c = \beta/w_c$. In the large-system limit, the input-output relation is decoupled into AWGN channels with noise variance $\bar{\sigma}_c^2$ given by a solution to the fixed-point equation $\bar{\sigma}_c^2 = \frac{1}{w_c} N_0 + \beta_c \Psi(\bar{\sigma}_c^2)$. Here, $\Psi(\bar{\sigma}_c^2)$ is the MSE function of the equalizer in cluster c.*

Due to the decoupling property in the large-system limit (cf. Section III), cluster c is decoupled into AWGN channels with noise variance $\bar{\sigma}_c^2$ for each user. Hence, fusion relies on $\mathbf{z}_c$ and $\bar{\sigma}_c^2$ for all C clusters and computes a weighted sum that minimizes the final noise variance σ_{FD}^2. Lemma 4 summarizes the optimal fusion rule; the proof is given in Appendix A-B.

Lemma 4. *Assume the large-system limit. Let $\bar{\sigma}_c^2$ be the noise variance for each cluster $c = 1, \ldots, C$. Then, the input-output relation of the FD architecture is decoupled into AWGN channels, where the optimal fusion rule is given by*

$$\sigma_{FD}^2 = \big(\textstyle\sum_{c=1}^{C} \frac{1}{\bar{\sigma}_c^2} \big)^{-1} = N_0 + \beta \sum_{c=1}^{C} \nu_c \Psi(\bar{\sigma}_c^2) \qquad (6)$$

with $\nu_c = \frac{1}{\bar{\sigma}_c^2} \big(\sum_{c'=1}^{C} 1/\bar{\sigma}_{c'}^2 \big)^{-1}$ for each $c = 1, \ldots, C$.

We also have the following intuitive result that FD cannot outperform PD equalization; the proof is given in Appendix A-C.

Lemma 5. *Let $N_0 > 0$. In the large-system limit, the decoupled noise variances for the FD and PD architectures satisfy $\sigma_{FD}^2 \geq \sigma_{PD}^2$. Equality holds for $\beta \to 0$ or if MRC is used.*

V. Numerical Results

We now investigate the performance of decentralized equalization. We will use an interference-free AWGN channel as the baseline and compare the performance loss (in terms of achievable rates) of FD and PD equalization with MRC, ZF, L-MMSE, and LAMA compared to this baseline. We define the signal-to-noise ratio (SNR) as $SNR = \beta E_s/N_0$ and the SNR loss as the additional E_s/N_0 that is required by these equalizers to achieve the same performance as that given by the interference-free AWGN system. We will assume QPSK modulation and $C = 3$ clusters with $w_c = 1/C$, $c = 1, \ldots, C$.

A. Achievable rate

Fig. 1 has shown the achievable rates for $\beta^{-1} = 6$ for LAMA and L-MMSE for both PD and FD architectures. LAMA-FD and L-MMSE-FD suffer only a minimal rate loss compared to the PD counterparts, whereas MRC suffers significant loss in the high-rate regime. Also, LAMA-FD for $\beta^{-1} = 6$ achieves significantly higher rates than that given by LAMA-PD for $\beta^{-1} = 2$, which reflects the benefits of data fusion.

We now compute the minimum SNR required to achieve a target rate of 99.5% of full rate (2 bits/user/channel use for QPSK) for an AWGN channel. We compare the SNR loss for the other equalizers to achieve the same achievable rate for different values of β. Fig. 3(a) shows the results for the different equalization algorithms and decentralized architectures. As expected, LAMA-PD significantly outperforms linear algorithms that use PD equalization. For the FD architecture with an SNR loss of 2 dB, LAMA-FD achieves the target rate for any $\beta < 0.34$. LAMA-FD outperforms L-MMSE-FD and suffers only a minimal loss compared to L-MMSE-PD, which requires $\beta < 0.16$ and $\beta < 0.41$, respectively. This observation implies that LAMA-FD achieves a similar performance as linear equalizers that use the PD architecture while requiring reduced interconnect and chip I/O bandwidth.

B. Minimum required BS-to-user ratio β^{-1}

We fix the SNR loss to 2 dB and plot the minimum BS-to-user ratio β^{-1} for varying achievable rates. In the low-rate regime, MRC performs similar to all other equalizers, which confirms the well-known fact that MRC is sufficient for massive MU-MIMO in the large-antenna limit [2]. In the high-rate regime, however, MRC requires significantly higher BS-to-user ratios compared to L-MMSE or LAMA-based equalization. It is interesting to see that the minimum β^{-1} decreases for LAMA-FD and LAMA-PD at high rates; this is due to the fact that LAMA in overloaded systems is particularly robust at low and high values of SNR (see [7] for more details).

C. Symbol error-rate (SER): analysis vs. simulations

We simulate an uncoded 96×16 massive MIMO system and plot the symbol error-rate performance (SER) of LAMA and L-MMSE for the PD and FD architectures with $C = 3$ clusters. We observe that the numerical error-rate simulations (shown with solid lines) closely match the asymptotic predictions (shown with dashed lines for the corresponding color). We also simulate LAMA-PD for an uncoded 32×16 MIMO system as a baseline for comparison with the proposed architectures.

We see that LAMA-FD outperforms L-MMSE-FD and performs close to MMSE-PD. Furthermore, LAMA-FD performs within 1 dB of LAMA-PD, which achieves individually-optimal performance in the large-system limit [7]. In addition, LAMA-FD with $C = 3$ in the 96×16 system exhibits significant performance improvements over LAMA-PD in the 32×16 system, which showcases the benefit of the fusion operation in finite-dimensional systems. In summary, LAMA-FD delivers near-optimal performance while reducing the interconnect and chip I/O bottlenecks, which demonstrates the efficacy of LAMA and the proposed feedforward equalization architectures.

Appendix A
Proofs

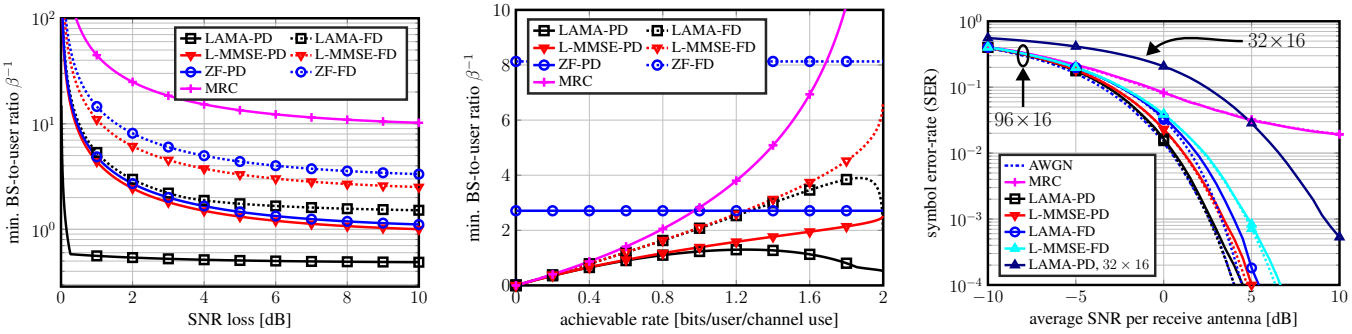
A. Proof of Theorem 3

As each entry of the partial matrix channel $\mathbf{H}_c$ is distributed as $\mathcal{CN}(0, 1/B)$, we first normalize the system by $1/\sqrt{w_c}$, which amplifies the noise variance by $1/w_c$. The result follows from Theorems 1 and 2 for LAMA and linear equalization.

B. Proof of Lemma 4

Since in the large-system limit, each input-output relation for cth cluster is statistically equivalent to AWGN with noise variance of $\bar{\sigma}_c^2$, the output of fusion will also be AWGN. If we define the fusion stage to perform $\mathbf{z} = \sum_{c=1}^{C} \nu_c \mathbf{z}_c$, we have

$$\mathbf{z} = \textstyle\sum_{c=1}^{C} \nu_c \mathbf{z}_c = \sum_{c=1}^{C} \nu_c \mathbf{s}_0 + \sum_{c=1}^{C} \nu_c \bar{\sigma}_c \mathbf{n}_c \overset{(a)}{=} \mathbf{s}_0 + \sigma_{\text{FD}} \bar{\mathbf{n}},$$



(a) Fixed rate of $R = 1.99$ [bits/user/channel use]. (b) Fixed 2 dB SNR loss. (c) SER plot for a 96×16 and 32×16 system.

Fig. 3. Minimum BS-to-user ratio β^{-1} with $C = 3$ clusters to achieve (a) a fixed achievable rate of $R = 1.99$ and (b) an SNR loss of 2 dB. Any β^{-1} above the curves are achievable for each system. (c) Simulated 96×16 MIMO system. Our numerical simulations agree with the asymptotic, analytical predictions which are shown with dashed lines. FD equalizers for 96×16 significantly outperform the error-rate performance of LAMA-PD for a 32×16 MIMO system.

where (a) follows from $\sum_{c=1}^{C} \nu_c = 1$ and $\mathbf{n}_c \sim \mathcal{CN}(0, \mathbf{I}_U)$ are independent for all $c = 1, \ldots, C$. Here, we have that $\sigma_{\text{FD}}^2 = \sum_{c=1}^{C} \nu_c^2 \bar{\sigma}_c^2$ and $\bar{\mathbf{n}} \sim \mathcal{CN}(0, \mathbf{I}_U)$.

We minimize the noise variance σ_{FD}^2 subject to the constraint $\sum_{c=1}^{C} \nu_c = 1$, which gives $\nu_c = \frac{1}{\bar{\sigma}_c^2} \big(\sum_{c=1}^{C} 1/\bar{\sigma}_c^2 \big)^{-1}$ for all $c = 1, \ldots, C$. Obtaining the first expression in (6) is straightforward; the second expression is obtained as follows:

$$\beta \textstyle\sum_{c=1}^{C} \nu_c \Psi(\bar{\sigma}_c^2) = \left(\sum_{c=1}^{C} \frac{1}{\bar{\sigma}_c^2} \right)^{-1} \sum_{c=1}^{C} \frac{\beta \Psi(\bar{\sigma}_c^2)}{\bar{\sigma}_c^2}$$
$$= \left(\textstyle\sum_{c=1}^{C} \frac{1}{\bar{\sigma}_c^2} \right)^{-1} \sum_{c=1}^{C} \left(w_c - \frac{N_0}{\bar{\sigma}_c^2} \right) = \left(\textstyle\sum_{c=1}^{C} \frac{1}{\bar{\sigma}_c^2} \right)^{-1} - N_0.$$

C. Proof of Lemma 5

We first show when equality holds. The case for $\beta \to 0$ is trivial because $\bar{\sigma}_c^2 = \sigma_{\text{FD}}^2 = \sigma_{\text{PD}}^2 = N_0$. For MRC, we have $\sigma_{\text{FD}}^2 = N_0 + \beta \sum_{c=1}^{C} \nu_c \, \mathbb{V}\text{ar}_S[S] = N_0 + \beta \, \mathbb{V}\text{ar}_S[S] = \sigma_{\text{PD}}^2$.

Let us now assume that $\beta > 0$. We show $\bar{\sigma}_c^2 > \sigma_{\text{PD}}^2$ by re-writing the fixed-point solutions as [19]: $\bar{\sigma}_c^2 = \sup\{\sigma^2 : N_0 + \beta \Psi(\sigma^2) \geq w_c \sigma^2\}$ and $\sigma_{\text{PD}}^2 = \sup\{\sigma^2 : N_0 + \beta \Psi(\sigma^2) \geq \sigma^2\}$. Note that $N_0 > 0$, so both $\bar{\sigma}_c^2$ and σ_{PD}^2 are strictly positive. It is easy to see that $\sigma_{\text{PD}}^2 \neq \bar{\sigma}_c^2$ because $\sigma_{\text{PD}}^2 = N_0 + \beta \Psi(\sigma_{\text{PD}}^2) > w_c \sigma_{\text{PD}}^2$. Since $\Psi(\sigma^2) \to \mathbb{V}\text{ar}_S[S]$ as $\sigma^2 \to \infty$ and $\Psi(\sigma^2)$ is continuous [20], there exists a $\bar{\sigma}_c^2 > \sigma_{\text{PD}}^2$ that satisfies $N_0 + \beta \Psi(\bar{\sigma}_c^2) = w_c \bar{\sigma}_c^2$ by the intermediate value theorem.

Finally, we use [20, Prop. 9] to see that $\Psi(\sigma^2)$ is strictly increasing for $\sigma^2 > 0$ for LAMA. For ZF and MMSE, this also holds by computing $\mathrm{d}\Psi(\sigma^2)/\mathrm{d}\sigma^2 > 0$. Thus, the result $\sigma_{\text{FD}}^2 > \sigma_{\text{PD}}^2$ follows directly from Lemma 4 since

$$\sigma_{\text{FD}}^2 = N_0 + \beta \sum_{c=1}^{C} \nu_c \Psi(\bar{\sigma}_c^2) > N_0 + \beta \sum_{c=1}^{C} \nu_c \Psi(\sigma_{\text{PD}}^2) = \sigma_{\text{PD}}^2.$$

REFERENCES

[1] J. Andrews, S. Buzzi, W. Choi, S. Hanly, A. Lozano, A. Soong, and J. Zhang, "What will 5G be?" *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1065–1082, Jun. 2014.

[2] E. Larsson, O. Edfors, F. Tufvesson, and T. Marzetta, "Massive MIMO for next generation wireless systems," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 186–195, Feb. 2014.

[3] K. Li, R. Sharan, Y. Chen, J. R. Cavallaro, T. Goldstein, and C. Studer, "Decentralized beamforming for massive MU-MIMO on a GPU cluster," in *Global Conf. Sig. Inform. Process. (GlobalSIP)*, Dec. 2016, pp. 590–594.

[4] S. Malkowsky, J. Vieira, K. Nieman, N. Kundargi, I. Wong, V. Öwall, O. Edfors, F. Tufvesson, and L. Liu, "Implementation of low-latency signal processing and data shuffling for TDD massive MIMO systems," in *IEEE Intl. Workshop Signal Process. Syst.*, Oct. 2016, pp. 260–265.

[5] C. Shepard, H. Yu, N. Anand, E. Li, T. Marzetta, R. Yang, and L. Zhong, "Argos: Practical many-antenna base stations," in *Proc. Ann. Intl. Conf. Mobile Comput. Netw. (MobiCom)*, 2012, pp. 53–64.

[6] J. Hoydis, S. ten Brink, and M. Debbah, "Massive MIMO: How many antennas do we need?" in *Proc. Allerton Conf. Commun., Contr., Comput.*, Sept. 2011, pp. 545–550.

[7] C. Jeon, R. Ghods, A. Maleki, and C. Studer, "Optimality of large MIMO detection via approximate message passing," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2015, pp. 1227–1231.

[8] K. Li, Y. Chen, R. Sharan, T. Goldstein, J. R. Cavallaro, and C. Studer, "Decentralized data detection for massive MU-MIMO on a Xeon Phi cluster," in *Proc. Asilomar Conf. Signals, Syst., Comput.*, Nov. 2016, pp. 468–472.

[9] R. Irmer, H. Droste, P. Marsch, M. Grieger, G. Fettweis, S. Brueck, H. P. Mayer, L. Thiele, and V. Jungnickel, "Coordinated multipoint: Concepts, performance, and field trial results," *IEEE Commun. Mag.*, vol. 49, no. 2, pp. 102–111, Feb. 2011.

[10] M. Peng, Y. Li, Z. Zhao, and C. Wang, "System architecture and key technologies for 5G heterogeneous cloud radio access networks," *IEEE Netw.*, vol. 29, no. 2, pp. 6–14, Mar. 2015.

[11] D. L. Donoho, A. Maleki, and A. Montanari, "Message-passing algorithms for compressed sensing," *Proc. Natl. Acad. Sci. USA*, vol. 106, no. 45, pp. 18 914–18 919, Nov. 2009.

[12] M. Bayati and A. Montanari, "The dynamics of message passing on dense graphs, with applications to compressed sensing," *IEEE Trans. Inf. Theory*, vol. 57, no. 2, pp. 764–785, Feb. 2011.

[13] C. Jeon, A. Maleki, and C. Studer, "On the performance of mismatched data detection in large MIMO systems," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2016, pp. 180–184.

[14] J. Zhu, A. Beirami, and D. Baron, "Performance trade-offs in multi-processor approximate message passing," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2016, pp. 680–684.

[15] A. Paulraj, R. Nabar, and D. Gore, *Introduction to Space-Time Wireless Communications*. Cambridge Univ. Press, 2003.

[16] C. Studer, S. Fateh, and D. Seethaler, "ASIC implementation of soft-input soft-output MIMO detection using MMSE parallel interference cancellation," *IEEE J. Solid-State Circuits*, vol. 46, no. 7, pp. 1754–1765, 2011.

[17] D. Tse and S. Hanly, "Linear multiuser receivers: effective interference, effective bandwidth and user capacity," *IEEE Trans. Inf. Theory*, vol. 45, no. 2, pp. 641–657, Mar. 1999.

[18] D. Guo and S. Verdú, "Randomly spread CDMA: Asymptotics via statistical physics," *IEEE Trans. Inf. Theory*, vol. 51, no. 6, pp. 1983–2010, Jun. 2005.

[19] A. Maleki, "Approximate message passing algorithms for compressed sensing," Ph.D. dissertation, Stanford University, Jan. 2011.

[20] D. Guo, Y. Wu, S. Shamai, and S. Verdú, "Estimation in Gaussian noise: Properties of the minimum mean-square error," *IEEE Trans. Inf. Theory*, vol. 57, no. 4, pp. 2371–2385, Apr. 2011.