Decentralized Equalization with Feedforward Architectures for Massive MU-MIMO

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

***Abstract*—Linear data-detection algorithms that build on zero forcing (ZF) or linear minimum mean-square error (L-MMSE) equalization achieve near-optimal spectral efficiency in massive multi-user multiple-input multiple-output (MU-MIMO) systems. Such algorithms, however, typically rely on centralized processing at the base-station (BS) which results in (i) excessive interconnect and chip input/output (I/O) data rates and (ii) high computational complexity. Decentralized baseband processing (DBP) partitions the BS antenna array into independent clusters that are associated with separate radio-frequency circuits and computing fabrics in order to overcome the limitations of centralized processing. In this paper, we investigate decentralized equalization with feedforward architectures that minimize the latency bottlenecks of existing DBP solutions. We propose two distinct architectures with different interconnect and I/O bandwidth requirements that fuse the local equalization results of each cluster in a feedforward network. For both architectures, we consider maximum ratio combining, ZF, L-MMSE, and a nonlinear equalization algorithm that relies on approximate message passing. For these algorithms and architectures, we analyze the associated post-equalization signal-to-noise-and-interference-ratio (SINR). We provide reference implementation results on a multi graphics processing unit (GPU) system which demonstrate that decentralized equalization with feedforward architectures enables throughputs in the Gb/s regime and incurs no or only a small performance loss compared to centralized solutions.**

***Index Terms*—Data detection, decentralized baseband processing, linear and nonlinear equalization, general-purpose computing on graphics processing units (GPGPU), massive MU-MIMO.**

I. Introduction

MASSIVE multi-user (MU) multiple-input multiple-output (MIMO) will be a key technology for next-generation wireless systems [2]–[4]. By equipping the infrastructure base-stations (BSs) with hundreds or thousands of active antenna elements and serving tens or hundreds of user equipments (UEs) simultaneously and in the same frequency band, massive MU-MIMO promises orders-of-magnitude improvements in spectral efficiency and energy efficiency compared to traditional, small-scale MIMO [5], [6]. However, the large number of antennas at the BS causes significant challenges when implementing this technology in practice. One of the most prominent challenges is the excessively high amount of fronthaul data that must be transferred from the radio-frequency (RF) antenna units at the BS antenna array to the baseband processing unit (BBU) [7]–[10]. For example, the fronthaul data rates (from RF chains to the BBU) exceed 200 Gbit/s for a massive MU-MIMO system with 128 BS antennas, each using two 10 bit analog-to-digital converters (for in-phase and quadrature components) operating at 80 MS/s sampling rate. Such high data rates not only exceed the bandwidth of existing high-speed interconnect standards, such as the common public radio interface (CPRI) [11], but will also approach the limits of existing chip input/output (I/O) interfaces in terms of bandwidth and power dissipation [12]. Furthermore, traditional equalization-based data-detection algorithms that achieve near-optimal spectral efficiency in the MU-MIMO uplink [5], such as zero-forcing (ZF) and linear minimum mean-square error (L-MMSE)-based equalization, rely on centralized processing in a single computing fabric, which results in excessively high complexity and power consumption for systems with large antenna arrays [8], [13].

A. Decentralized Baseband Processing

In order to mitigate the bandwidth and computing bottlenecks of centralized massive MU-MIMO architectures, existing testbeds either distribute the most critical baseband processing tasks in the frequency domain or use maximum ratio combining (MRC). Concretely, the testbeds described in [14]–[17] parallelize the key baseband processing tasks across the subcarriers of orthogonal frequency-division multiplexing (OFDM)-based systems. While this approach enables high parallelism, it requires that each frequency cluster obtains data from *all* BS antennas, which alone does not enable one to scale such systems to thousands of antenna elements [8]. In contrast to frequency parallelization, MRC enables antenna parallelization that divides array into independent clusters [2], [18]; this approach significantly reduces the interconnect bandwidth between the RF chains and the BBUs. MRC, however, suffers from low spectral efficiency for realistic antenna configurations and high-rate modulation and coding schemes [5]. Consequently, realizing massive MU-MIMO in practice requires solutions that reduce the interconnect and chip I/O bandwidth as well as the baseband processing complexity per computing fabric, without sacrificing spectral efficiency.

Decentralized baseband processing (DBP) has been proposed in [8] to alleviate the fronthaul and I/O bandwidth bottlenecks, and enables parallel baseband processing across BS antennas on multiple computing fabrics, such as application-specific

C. Jeon was with the School of Electrical and Computer Engineering (ECE) at Cornell University, Ithaca, NY, and is now with Intel Labs, Hillsboro, OR; email: cj339@cornell.edu.

C. Studer was with the School of ECE at Cornell University, Ithaca, NY, and is now with Cornell Tech, New York City, NY; email: studer@cornell.edu; website: vip.ece.cornell.edu.

K. Li and J. R. Cavallaro are with the Department of ECE at Rice University, Houston, TX; email: kl33@rice.edu, cavallar@rice.edu.

Parts of the theoretical analysis in this paper have been presented at the 2017 IEEE International Symposium on Information Theory (ISIT) [1].

integrated circuits (ASICs), field-programmable gate arrays (FPGAs), or graphics processing units (GPUs) [19], [20], while achieving high spectral efficiency. The idea of DBP is to partition the BS antenna array into C independent clusters, each associated with local computing fabrics that carry out the necessary RF and baseband processing tasks in a decentralized and parallel fashion. The algorithms proposed in [8] perform linear equalization and precoding in an iterative manner by exchanging consensus information among the clusters. However, implementation results on a GPU cluster revealed that the transfer latency of such consensus-sharing methods are limiting the achievable throughput. To avoid this drawback, references [1], [7], [12], [21] recently proposed *feedforward* architectures that minimize the transfer latency.

B. Contributions

We propose two distinct feedforward architectures for partially decentralized (PD) and fully decentralized (FD) equalization, which mitigate the interconnect, I/O, latency, and computation bottlenecks. For both of these architectures, we investigate the efficacy of MRC, ZF, L-MMSE, and a nonlinear equalization method that builds upon the large-MIMO approximate message passing (LAMA) algorithm [22]. Our main contributions can be summarized as follows:

- We develop a framework that enables a precise analysis of the post-equalization signal-to-noise-and-interference-ratio (SINR) of decentralized equalization with feedforward architectures in the large-system limit.
- We show that the PD feedforward architecture achieves the same SINR performance as centralized solutions for equalization with MRC, ZF, L-MMSE, and PD-LAMA.
- We show that the FD feedforward architecture is able to provide near-optimal SINR performance, but further reduces the interconnect and I/O bandwidths.
- We analyze optimal antenna partitioning strategies that maximize the SINR for the FD architecture.
- We conduct error-rate simulations for a realistic 3GPP long-term evolution (LTE)-like massive MU-MIMO system that support our theoretical findings.
- We provide reference throughput and latency results for linear and nonlinear equalization in centralized, PD, and FD architectures on a multi-GPU system.

Our results demonstrate that feedforward equalization enables throughputs in the Gb/s regime for massive MU-MIMO systems with hundreds of antenna elements, and incurs no or only a small loss in post-equalization SINR and error-rate performance compared to that of centralized solutions.

C. Relevant Prior Art

DBP for massive MI-MIMO systems has been proposed in [8] together with consensus-sharing equalization and precoding algorithms. Distributed processing across antenna elements is also a critical component of coordinated multipoint (CoMP) [23] and cloud radio access networks (CRANs) [24] for multi-cell transmission. While all these architectures and algorithms are able to reduce the raw baseband data rates and mitigate the computation bottlenecks, their performance has not been analyzed and the achievable throughput suffers from high interconnect latency caused by iterative exchange of consensus information. To avoid iterative consensus sharing among antenna clusters, we focus on decentralized *feedforward* architectures that minimize the transfer latency and enable a precise theoretical performance analysis.

Feedforward architectures for decentralized massive MU-MIMO equalization have been proposed in [1], [7], [12], [21]. The present paper extends our theoretical results from [1] and, in contrast to [7], [12], [21], provides two distinct architectures and a corresponding SINR analysis for a range of linear and nonlinear equalization algorithms. In addition, we provide reference implementation results on a GPU cluster to assess the throughput and latency of our architectures and algorithms.

The post-equalization SINR performance of *centralized* linear equalization algorithms, such as MRC, ZF, and L-MMSE, has been analyzed in [25]–[28] in the large-system limit. We will investigate the SINR performance of these algorithms for the two proposed decentralized feedforward architectures, and also investigate the efficacy of nonlinear equalization for decentralized massive MU-MIMO architectures.

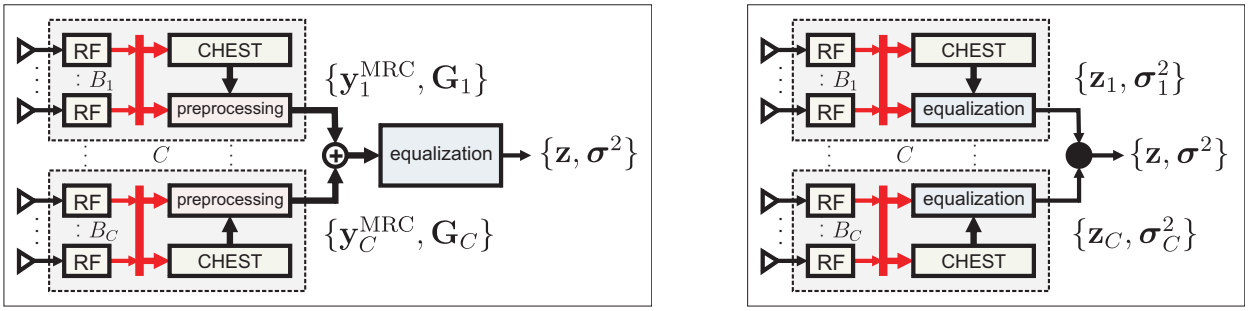
Nonlinear equalization for massive MU-MIMO systems via approximate message passing (AMP) has been studied in [22], [29], [30]. Message passing has also been used recently for data detection in non-orthogonal multiple access (NOMA) systems [31]. A distributed version of AMP has been proposed in [32] for compressive sensing applications [33], [34]. The key differences of our nonlinear equalization algorithm to these results are as follows: (i) We consider decentralized feedforward architectures; (ii) the methods in [22], [29]–[31] are centralized; (iii) the distributed AMP-based method in [32] requires iterative consensus sharing; (iv) we analyze the post-equalization SINR and error-rate performance in massive MU-MIMO systems.

D. Notation

Lowercase and uppercase boldface letters designate vectors and matrices, respectively; uppercase calligraphic letters denote sets. The transpose and Hermitian of the matrix $\mathbf{A}$ are represented by $\mathbf{A}^{\mathrm{T}}$ and $\mathbf{A}^{\mathrm{H}}$, respectively. The $M \times N$ all-zeros matrix is $\mathbf{0}_{M\times N}$ and the N-dimensional identity matrix is $\mathbf{I}_N$. The kth entry of a vector $\mathbf{a}$ is a_k. We define $\langle \mathbf{x} \rangle = \frac{1}{N}\sum_{k=1}^{N} x_k$. The circularly-symmetric multivariate complex-valued Gaussian probability density function (pdf) with covariance $\mathbf{K}$ is denoted by $\mathcal{CN}(\mathbf{0}, \mathbf{K})$. $\mathbb{E}_X[\cdot]$ and $\mathbb{V}\mathrm{ar}_X[\cdot]$ represent the mean and variance with respect to the random variable X, respectively.

E. Paper Outline

The rest of the paper is organized as follows. Section II introduces the system model and the two feedforward architectures. Section III and Section IV investigate equalization algorithms for the PD and FD architecture, respectively. Section V provides theoretical and simulative results for the proposed methods. Section VI details the multi-GPU implementation and provides throughput and latency results. Section VII concludes the paper. All derivations and proofs are relegated to the appendices.



(a) Partially decentralized (PD) equalization architecture.

(b) Fully decentralized (FD) equalization architecture.

Fig. 1. Partially decentralized (PD) and fully decentralized (FD) feedforward equalization architectures for the massive MU-MIMO uplink. The antenna array is divided in C clusters, each associated with local radio-frequency (RF) processing and channel estimation (CHEST). (a) PD performs decentralized CHEST and preprocessing; equalization is performed in a centralized fashion and operates on a low-dimensional data (dimension is the number of UEs). (b) FD performs CHEST, preprocessing, and equalization in a decentralized manner; the final equalization result is formed by a weighted average of local estimates. The $\oplus$ operator in (a) denotes matrix/vector-additions and $\bullet$ in (b) denotes a weighted vector addition; see Eq. 19 in Section IV for the details.

II. Decentralized Equalization Architectures

We start by introducing the considered massive MU-MIMO system model and the basics of equalization-based data detection. We then discuss the two feedforward equalization architectures for DBP depicted in Fig. 1, and detail the SINR analysis framework that we will use throughout the paper.

A. Uplink System Model and Equalization

We consider a narrowband massive MU-MIMO uplink system in which U single-antenna UEs transmit data to a BS with B antenna elements. To model this scenario, we use the standard input-output relation [2]

$$\mathbf{y} = \mathbf{H}\mathbf{s}_0 + \mathbf{n}. \tag{1}$$

Here, $\mathbf{y} \in \mathbb{C}^B$ is the receive vector at the BS, $\mathbf{H} \in \mathbb{C}^{B\times U}$ represents the MIMO system matrix, which we assume is perfectly known at the BS, $\mathbf{s}_0 \in \mathcal{O}^U$ contains the transmit symbols for each UE, $\mathcal{O}$ is the constellation set (e.g., QPSK or 16-QAM), 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 $p(\mathbf{s}_0) = \prod_{u=1}^U p(s_{0u})$ for the transmit vector and the following distribution for each transmit symbol:

$$p(s_{0u}) = \frac{1}{|\mathcal{O}|}\sum_{a\in\mathcal{O}} \delta(s_{0u} - a), \tag{2}$$

where $|\mathcal{O}|$ is the cardinality of the constellation $\mathcal{O}$ and $\delta(\cdot)$ is the Dirac delta function. In what follows, we assume zero-mean constellations and define the average energy per transmit symbol as $E_s = \mathbb{E}\big[|s_{0u}|^2\big]$, $u = 1, \ldots, U$.

Equalization is concerned with forming an estimate $\mathbf{z}$ of the transmit signal vector $\mathbf{s}_0$ along with reliability estimates for each entry in $\mathbf{z}$. These two quantities are then used by the data detector to compute hard-output estimates for the transmit symbols or bit-wise soft information in the form of log-likelihood ratios [35], [36]. Consider a general *centralized* equalizer $\{\mathbf{z}, \boldsymbol{\sigma}^2\} = \mathcal{E}(\mathbf{y}, \mathbf{H})$ that takes the received vector $\mathbf{y}$ and the MIMO channel matrix $\mathbf{H}$ in order to compute (i) an estimate $\mathbf{z}$ for the true transmit vector $\mathbf{s}_0$ and (ii) the associated error variance vector $\boldsymbol{\sigma}^2$. The error variance vector characterizes the post-equalization residual interference and noise variance on each entry of the estimate $\mathbf{z}$. Mathematically, this quantity corresponds to the variances of each entry in the residual interference and noise vector defined as $\mathbf{e} = \mathbf{z} - \mathbf{s}_0$, i.e., $\boldsymbol{\sigma}^2 = \mathbb{E}\big[|\mathbf{e}|^2\big]$ where $|\cdot|^2$ operates element-wise on vectors.

The literature describes a range of linear and nonlinear equalization algorithms for small-scale and massive MU-MIMO data detection [13], [22], [37], [38]. Linear methods, such as MRC, ZF, and L-MMSE are among the most common algorithms, mainly due to their simplicity and low computational complexity [13]. Nevertheless, nonlinear equalizers, such as the LAMA algorithm put forward in [22], have been shown to (often significantly) outperform linear equalizers at the cost of higher computational complexity [22], [30], [39].

B. Basics of Decentralized Equalization

As in [8], we partition the B BS antenna elements into $C \in \{1, 2, \ldots, B\}$ independent *antenna clusters*. The cth antenna 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 RF components and only requires access to local channel state information (CSI) acquired in a local channel estimation (CHEST) unit. Without loss of generality, we partition the receive vector $\mathbf{y} = [\mathbf{y}_1^{\mathrm{T}}, \ldots, \mathbf{y}_C^{\mathrm{T}}]^{\mathrm{T}}$, the channel matrix $\mathbf{H} = [\mathbf{H}_1^{\mathrm{T}}, \ldots, \mathbf{H}_C^{\mathrm{T}}]^{\mathrm{T}}$, and the noise vector $\mathbf{n} = [\mathbf{n}_1^{\mathrm{T}}, \ldots, \mathbf{n}_C^{\mathrm{T}}]^{\mathrm{T}}$ in (1). For this antenna partitioning scheme, the input-output relation corresponding to the local receive vector $\mathbf{y}_c$ associated with the cth cluster can be written as

$$\mathbf{y}_c = \mathbf{H}_c\mathbf{s}_0 + \mathbf{n}_c, \quad c = 1, \ldots, C, \tag{3}$$

with $\mathbf{y}_c \in \mathbb{C}^{B_c}$, $\mathbf{H}_c \in \mathbb{C}^{B_c\times U}$, and $\mathbf{n}_c \in \mathbb{C}^{B_c}$. The following subsections describe two decentralized equalization architectures that compute estimates for the transmit vector $\mathbf{s}_0$ by performing local computations in each antenna cluster using only information of the local receive vector $\mathbf{y}_c$ and channel matrix $\mathbf{H}_c$ followed by fusion of the results from all clusters.

C. Partially Decentralized (PD) Equalization Architecture

The partially decentralized (PD) equalization architecture is illustrated in Fig. 1(a). First, each cluster $c = 1, \ldots, C$

independently (and in parallel) preprocesses the local receive vector $\mathbf{y}_c$ and channel matrix $\mathbf{H}_c$ by computing the U-dimensional local MRC vector $\mathbf{y}_c^{\text{MRC}} = \mathbf{H}_c^{\text{H}}\mathbf{y}_c$ and the $U \times U$ local Gram matrix $\mathbf{G}_c = \mathbf{H}_c^{\text{H}}\mathbf{H}_c$. Second, a *feedforward* adder tree, indicated with the symbol $\oplus$ in Fig. 1(a), is used to compute the complete MRC vector and Gram matrix as follows:

$$\mathbf{y}^{\text{MRC}} = \sum_{c=1}^{C} \mathbf{y}_c^{\text{MRC}} \quad \text{and} \quad \mathbf{G} = \sum_{c=1}^{C} \mathbf{G}_c. \tag{4}$$

Third, we perform linear or nonlinear equalization in a centralized unit that computes the estimate $\mathbf{z} \in \mathbb{C}^U$ and the post-equalization error variance vector $\boldsymbol{\sigma}^2 \in \mathbb{C}^U$. The tuple $\{\mathbf{z}, \boldsymbol{\sigma}^2\}$ is then used to compute hard- or soft-output estimates.

In Section III, we will detail MRC, ZF, L-MMSE equalization, and a new LAMA-based equalization algorithm [22] for the PD architecture, all of which directly operate on the U-dimensional fused MRC vector $\mathbf{y}^{\text{MRC}}$ and Gram matrix $\mathbf{G}$. Since the MRC vector is a sufficient statistic for the transmit signal $\mathbf{s}_0$ [35], we will show that the PD equalization does not incur an SINR performance loss compared to centralized MRC, ZF, L-MMSE, and LAMA equalizers.

D. Fully Decentralized (FD) Equalization Architecture

The PD architecture requires a summation of both the local MRC vectors and the local Gram matrices, which involves potentially large amounts of data to be transmitted to the central equalization unit, especially in channels with short coherence time. The fully decentralized (FD) equalization architecture illustrated in Fig. 1(b) avoids the transmission of the local Gram matrices altogether at the cost of a (typically small) performance loss. First, each cluster $c = 1, \ldots, C$ independently (and in parallel) performs CHEST, preprocessing, *and* equalization, i.e., directly forms a local estimate $\mathbf{z}_c \in \mathbb{C}^U$ and local post-equalization error variance vector $\boldsymbol{\sigma}_c^2 \in \mathbb{C}^U$. Second, a feedforward fusion tree, indicated with the symbol $\bullet$ in Fig. 1(b), optimally combines the local estimates $\mathbf{z}_c$ using information from the error variance vectors $\boldsymbol{\sigma}_c^2$ in order to generate the final output tuple $\{\mathbf{z}, \boldsymbol{\sigma}^2\}$.

In Section IV, we will detail the optimal fusion rule as well as MRC, ZF, L-MMSE, and LAMA-based equalization for the FD architecture. We will also provide an SINR performance analysis in the large-system limit.

E. Signal-to-Interference-and-Noise-Ratio (SINR) Analysis

To analyze the performance of linear and nonlinear equalization algorithms for the PD and FD architectures, we will focus on the large-system limit and Rayleigh-fading channels. Hence, we will make frequent use of the following two definitions.

Definition 1 (Large-system limit)**.** *The* large-system limit *is defined by fixing the* system ratio $\beta = U/B$ *and* $U, B \to \infty$.

Definition 2 (Rayleigh fading)**.** *A MIMO channel is* Rayleigh fading *if the channel matrix* $\mathbf{H}$ *has i.i.d. circularly symmetric complex Gaussian entries with variance* $1/B$ *per entry.*

By considering the large-system limit and Rayleigh-fading channels, Tse and Hanly have shown in [26] that linear equalizers, such as MRC, ZF, and L-MMSE, *decouple* MIMO systems into parallel and independent additive white Gaussian noise (AWGN) channels. This means that the estimate $\mathbf{z}$ of such linear equalizers can be modeled on a per-UE basis in a statistically equivalent manner as follows:

$$z_u = s_{0u} + e_u, \quad u = 1, \ldots, U, \tag{5}$$

where $e_u \in \mathbb{C}$ represents residual interference and noise. Furthermore, the quantity e_u turns out to be (i) statistically independent of s_{0u} and (ii) circularly symmetric complex Gaussian with *decoupled noise variance* σ^2, which does not depend on the UE index u. This result also implies that all entries of the error variance vector $\boldsymbol{\sigma}^2$ correspond to σ^2. In Section III and Section IV for the PD and FD architecture, respectively, we will build upon this asymptotic analysis framework in order to theoretically characterize the per-UE post-equalization SINR of the decoupled system (5):

$$\mathsf{sinr} \triangleq \frac{E_s}{\sigma^2}. \tag{6}$$

Numerical results that validate our asymptotic analysis in finite-dimensional systems will be presented in Section V.

III. Partially Decentralized (PD) Equalization

We start by reviewing linear equalization algorithms for the PD architecture depicted in Fig. 1(a), and adapt the well-known Tse-Hanly equations [26] to analyze the associated post-equalization SINR performance in the large-system limit. We then present a new, nonlinear equalization algorithm that builds upon LAMA proposed in [22], and we develop a corresponding SINR performance analysis for the PD architecture.

A. Linear Equalization Algorithms for the PD Architecture

Since the MRC output $\mathbf{y}^{\text{MRC}}$ in (4) is a sufficient statistic for the transmit signal vector $\mathbf{s}_0$, a variety of optimal and suboptimal equalization-based data detection algorithms can be derived from this quantity [35]. For MRC-based equalization in the PD architecture, we use (4) to form the estimate

$$\mathbf{z}^{\text{MRC}} = \operatorname{diag}(\mathbf{G})^{-1}\mathbf{y}^{\text{MRC}},$$

where the diagonal matrix $\operatorname{diag}(\mathbf{G})^{-1}$ is computed in the centralized equalization unit; see Fig. 1(a). The MRC estimate $\mathbf{z}^{\text{MRC}}$ can then be used to perform either hard- or soft-output data detection. For soft-output data detection, one requires the error variance vector given by

$$\begin{aligned}\boldsymbol{\sigma}^2_{\text{MRC}} = \operatorname{diag}\big(&\operatorname{diag}(\mathbf{G})^{-1}\mathbf{G}\operatorname{diag}(\mathbf{G})^{-\text{H}}N_0 \\ &+ \big(\operatorname{diag}(\mathbf{G})^{-1}\mathbf{G} - \mathbf{I}_U\big)\big(\operatorname{diag}(\mathbf{G})^{-1}\mathbf{G} - \mathbf{I}_U\big)^{\text{H}}E_s\big)\end{aligned}$$

that contains the post-equalization SINR for each entry of $\mathbf{z}^{\text{MRC}}$. Note that MRC-based equalization was shown to be optimal (i) for a fixed number of UEs and infinitely many BS antennas [2], which is equivalent to $\beta \to 0$ in the large-system limit, or (ii) in the low-SNR regime [26]. The estimate of the ZF equalizer for the PD architecture is given by

$$\mathbf{z}^{\text{ZF}} = \mathbf{G}^{-1}\mathbf{y}^{\text{MRC}},$$

where the matrix $\mathbf{G}^{-1}$ is computed in the centralized equalization unit. The associated error variance vector is given by

$$\boldsymbol{\sigma}^2_{\text{ZF}} = \text{diag}(\mathbf{G}^{-1})N_0.$$

For L-MMSE equalization, we have

$$\mathbf{z}^{\text{L-MMSE}} = (\mathbf{G} + \rho\mathbf{I}_U)^{-1}\mathbf{y}^{\text{MRC}}.$$

where the matrix $(\mathbf{G} + \rho\mathbf{I}_U)^{-1}$ is computed in the centralized equalization unit. The L-MMSE regularization parameter is set to $\rho = N_0/E_s$ for complex-valued constellations (e.g., QPSK or 16-QAM). The associated error variance vector is given by

$$\begin{aligned}\boldsymbol{\sigma}^2_{\text{L-MMSE}} = \text{diag}\big((\mathbf{G} + \rho\mathbf{I}_U)^{-1}\mathbf{G}(\mathbf{G} + \rho\mathbf{I}_U)^{-\text{H}}N_0 \\ + \big((\mathbf{G} + \rho\mathbf{I}_U)^{-1}\mathbf{G} - \mathbf{I}_U\big)\big((\mathbf{G} + \rho\mathbf{I}_U)^{-1}\mathbf{G} - \mathbf{I}_U\big)^{\text{H}}E_s\big).\end{aligned}$$

We reiterate that the MRC, ZF, and L-MMSE equalizers for the PD architecture deliver exactly the *same estimates* as their centralized counterparts—the only difference is the way the involved quantities are computed.

As shown in [26] and outlined in Section II-E, centralized MRC, ZF, and L-MMSE equalizers decouple MIMO systems in the large-system limit and for Rayleigh fading channels; this implies that the entries of the error variance vectors $\boldsymbol{\sigma}^2_{\text{MRC}}$, $\boldsymbol{\sigma}^2_{\text{ZF}}$, and $\boldsymbol{\sigma}^2_{\text{L-MMSE}}$ converge to the decoupled noise variance σ^2 of the MRC, ZF, and L-MMSE equalizer, respectively. Since linear equalizers in the PD architecture yield exactly the same estimates as in a centralized architecture, we can directly characterize the associated decoupled noise variance σ^2_{PD} in the PD architecture using the following result.

Theorem 1 ([26, Thm. 3.1])**.** *Fix the system ratio $\beta = U/B$, and assume the large-system limit and Rayleigh fading channels. Then, the decoupled noise variance σ^2_{PD} for MRC, ZF, and L-MMSE equalization in a centralized or PD architecture, is the solution to the following fixed-point equation*

$$\sigma^2_{PD} = N_0 + \beta\Psi(\sigma^2_{PD}), \tag{7}$$

where the MSE function $\Psi(\sigma^2)$ is given by

$$\Psi(\sigma^2) = E_s, \tag{MRC}$$
$$\Psi(\sigma^2) = \sigma^2, \text{ for } \beta < 1, \tag{ZF}$$
$$\Psi(\sigma^2) = \frac{E_s}{E_s + \sigma^2}\sigma^2, \tag{L-MMSE}$$

for MRC, ZF, and L-MMSE equalization, respectively.

We note that the expression for the ZF equalizer only holds for $\beta < 1$, whereas the expressions for MRC and L-MMSE hold for general system ratios $\beta \geq 0$.[1] From Theorem 1, we obtain closed-form expressions for the post-equalization sinr in (6) for MRC, ZF, and L-MMSE equalization in the PD architecture.

[1]The asymptotic SINR performance of ZF equalization via the Moore-Penrose pseudo inverse when $\beta \geq 1$ was analyzed in [40].

Corollary 2. *Assume that the conditions of Theorem 1 hold. Then, the post-equalization* sinr *for MRC, ZF, and L-MMSE equalization in the PD architecture are given by*

$$\mathsf{sinr}^{\text{MRC}}_{\text{PD}} = \frac{E_s/N_0}{1 + \beta E_s/N_0}, \tag{MRC}$$
$$\mathsf{sinr}^{\text{ZF}}_{\text{PD}} = \frac{E_s}{N_0}(1 - \beta), \text{ for } \beta < 1, \tag{ZF}$$
$$\begin{aligned}\mathsf{sinr}^{\text{L-MMSE}}_{\text{PD}} = \frac{1}{2}\Bigg(\sqrt{\bigg(1 - \frac{E_s}{N_0}(1-\beta)\bigg)^2 + 4\frac{E_s}{N_0}} \\ - \bigg(1 - \frac{E_s}{N_0}(1-\beta)\bigg)\Bigg).\end{aligned} \tag{L-MMSE}$$

We note that in the massive MU-MIMO regime, which corresponds to $\beta \to 0$, all post-equalization SINR expressions converge to E_s/N_0, which confirms the well-known fact that MRC is optimal in this regime [2]. It can also be shown that $\mathsf{sinr}^{\text{L-MMSE}}_{\text{PD}}$ bounds $\mathsf{sinr}^{\text{MRC}}_{\text{PD}}$ and $\mathsf{sinr}^{\text{ZF}}_{\text{PD}}$ from above for all system ratios and in all SNR regimes. Hence, L-MMSE equalization is often the preferred choice in realistic massive MU-MIMO systems [5], [13]. We reiterate that the SINR expressions listed in Corollary 2 are also valid for centralized architectures.

B. LAMA for the PD Architecture

The LAMA algorithm developed in [22] is a nonlinear equalizer is able to achieve individually-optimal performance in the large-system limit given certain conditions on the antenna ratio β and the noise variance N_0 are satisfied. LAMA operates directly on the input-output relation in (1) and is, hence, designed for centralized processing. We now develop a novel variant of LAMA that directly operates on the complete MRC output $\mathbf{y}^{\text{MRC}}$ and the Gram matrix $\mathbf{G}$ in (4) to enables its use in the PD architecture. Since the antenna configuration in massive MU-MIMO systems typically satisfies $U \ll B$, the LAMA-PD algorithm operates on a lower dimension which reduces complexity while delivering exactly the same estimates as the original LAMA algorithm. We note that LAMA was derived in the large-system limit and for Rayleigh fading channels [41], but these assumptions are not required in practice. We next summarize the LAMA-PD algorithm; the derivation can be found in Appendix A-A.

Algorithm 1 (LAMA-PD)**.** *Initialize $s_\ell = \mathbb{E}_S[S]$ for $\ell = 1, \ldots, U$, $\phi^{(1)} = \mathbb{V}\text{ar}_S[S]$, and $\mathbf{v}^{(1)} = \mathbf{0}_{B\times 0}$. Then, for every iteration $t = 1, 2, \ldots, T_{max}$, compute the following steps:*

$$\begin{aligned}\mathbf{z}^t &= \mathbf{y}^{\text{MRC}} + (\mathbf{I} - \mathbf{G})\mathbf{s}^t + \mathbf{v}^t \\ \mathbf{s}^{t+1} &= \mathsf{F}(\mathbf{z}^t, N_0 + \beta\phi^t) \\ \phi^{t+1} &= \langle \mathsf{G}(\mathbf{z}^t, N_0 + \beta\phi^t)\rangle \\ \mathbf{v}^{t+1} &= \frac{\beta\phi^{t+1}}{N_0 + \beta\phi^t}(\mathbf{z}^t - \mathbf{s}^t).\end{aligned} \tag{8}$$

The functions $\mathsf{F}(s_\ell, \tau)$ and $\mathsf{G}(s_\ell, \tau)$ are the message mean and variance, respectively, operate element-wise on vectors, and

are computed as follows:

$$\mathsf{F}(z_\ell,\tau)=\int_{s_\ell} s_\ell f(s_\ell|\hat{z}_\ell)\mathrm{d}s_\ell \tag{9}$$
$$\mathsf{G}(z_\ell,\tau)=\int_{s_\ell} |s_\ell|^2 f(s_\ell|\hat{z}_\ell)\mathrm{d}s_\ell - |\mathsf{F}(z_\ell,\tau)|^2.$$

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 (2), and Z is a normalization constant. The estimates and error variances of LAMA are $\mathbf{z}^t$ and $\sigma^2_{t,\text{LAMA}} = N_0 + \beta\phi^t$, respectively.

In order to analyze the post-equalization SINR of LAMA-PD, it is key to realize that the equalization output $\mathbf{z}^t$ is equivalent to that of the original centralized LAMA algorithm in [22]. As shown in [22], LAMA (and hence LAMA-PD) decouples the MIMO system into parallel AWGN channels. More specifically, the equalizer output (8) of LAMA can be modeled as in (5), where the decoupled noise variance σ_t^2 at iteration t can be tracked using the state evolution (SE) framework in the large-system limit and for Rayleigh fading channels. The following result, with proof in [42], summarizes this SE framework.

Theorem 3 ([22, Thm. 1])**.** *Fix the system ratio $\beta = U/B$ and the signal prior (2). Assume the large-system limit and Rayleigh fading channels. Then, the decoupled noise variance σ_t^2 of LAMA and LAMA-PD at iteration t is given by the recursion:*

$$\sigma_t^2 = N_0 + \beta\Psi(\sigma_{t-1}^2). \tag{10}$$

Here, the MSE function is given by

$$\Psi(\sigma_{t-1}^2) = \mathbb{E}_{S,Z}\Big[\big|\mathsf{F}(S+\sigma_{t-1}Z,\sigma_{t-1}^2)-S\big|^2\Big], \tag{11}$$

where the function F is given in (9), $S \sim p(s)$ as in (2), $Z \sim \mathcal{CN}(0,1)$, and σ_1^2 is initialized by $\sigma_1^2 = N_0 + \beta E_s$.

For $t \to \infty$, the SE recursion in (10) converges to the same fixed-point equation of linear equalizers in (7), where the only difference is the MSE function (11). If there are multiple fixed points, then we select the largest σ^2_{PD}, which is, in general, a sub-optimal solution.[2] As for linear equalizers, we can use the fixed-point equation in (7) to analyze the post-equalization SINR performance of LAMA and LAMA-PD. Unfortunately, there are no closed-form expressions known for the decoupled noise variance or the SINR for LAMA and LAMA-PD with discrete constellations, due to the specific form of the MSE function (11). Nevertheless, we can numerically compute (11) and, hence, analyze the SINR. A corresponding SINR comparison with linear equalizers is given in Section V.

IV. Fully Decentralized (FD) Equalization

We next discuss optimal fusion for linear and nonlinear equalization in the PD architecture as depicted in Fig. 1(b). We then analyze the post-equalization SINR of the proposed equalizers depending on the antenna cluster allocation strategy.

[2]See [39] for more details on the existence of multiple fixed-points and on conditions for which LAMA achieves individually optimal performance.

A. Optimal Fusion for the FD Architecture

As detailed in Section II-D, each cluster $c = 1,\ldots,C$ in the FD architecture independently computes a local estimate $\mathbf{z}_c$ and associated error variance vector $\boldsymbol{\sigma}_c^2$. Then, the vectors $\mathbf{z}_c$ and $\boldsymbol{\sigma}_c^2$ for $c = 1,\ldots,C$ are fused to compute the final output tuple $\{\mathbf{z},\boldsymbol{\sigma}^2\}$. Since in the large-system limit and for Rayleigh fading channels, the considered equalizers decouple MIMO systems into parallel and independent AWGN channels (see Section II-E), we focus on *linear* fusion of the local estimates, indicated with $\bullet$ in Fig. 1(b). Specifically, the proposed FD architecture computes the fused estimate $\mathbf{z}$ by combining the local estimates for each UE as follows:

$$z_u = \sum_{c=1}^{C} \nu_{c,u} z_{c,u}, \quad u = 1,\ldots,U. \tag{12}$$

Here, $z_{c,u}$ is the local estimate for UE u at cluster c and the weights $\nu_{c,u}$, $u = 1,\ldots,U$, depend on the per-cluster error variance vector $\boldsymbol{\sigma}_c^2$. In what follows, we are interested in the optimal set of weights for the following criterion.

Definition 3 (Optimal fusion)**.** *Optimal fusion for the FD architecture maximizes the per-UE post-equalization SINR of the final estimate z_u obtained from (12) while $\sum_{c=1}^{C}\nu_{c,u} = 1$.*

In other words, optimal fusion defines a set of weights $\nu_{c,u}$, $c = 1,\ldots,C$, $u = 1,\ldots,U$, so that the decoupled noise variances contained in $\boldsymbol{\sigma}^2$ associated with the fused estimate $\mathbf{z}$ are minimized. The following result summarizes the optimal fusion rule; a short proof is given in Appendix A-B.

Lemma 4. *Let $\sigma^2_{c,u}$, $c = 1,\ldots,C$, $u = 1,\ldots,U$, be a set of given error variances for UE u and cluster c. Assume that the residual interference and noise terms are zero mean and uncorrelated among the clusters. Then, the weights that yield optimal fusion according to Definition 3 are given by*

$$\nu_{c,u} = \frac{1}{\sigma^2_{c,u}}\left(\sum_{c'=1}^{C}\frac{1}{\sigma^2_{c',u}}\right)^{-1}, \tag{13}$$

for $c = 1,\ldots,C$ and $u = 1,\ldots,U$.

B. SINR Analysis of Optimal Fusion in the FD Architecture

We are now interested in analyzing the post-fusion SINR for the FD architecture in the large-system limit. The following theorem analyzes the decoupled noise variance σ_c^2 for each cluster $c = 1,\ldots,C$; the proof is given in Appendix A-C.

Lemma 5. *Assume MRC, ZF, L-MMSE, or LAMA equalization in each cluster $c = 1,\ldots,C$. Consider the large-system limit and Rayleigh fading channels. Then, the input-output relation of each cluster is decoupled into parallel channels of the form (5) with decoupled noise variance σ_c^2 given by a solution to the following fixed-point equation:*

$$w_c\sigma_c^2 = N_0 + \beta\Psi(\sigma_c^2).$$

Here, $\Psi(\sigma_c^2)$ is the MSE function of the equalizer in cluster c.

This result shows that the per-UE error variances $\sigma^2_{c,u}$ will become independent of the UE index u in the large-antenna limit and for Rayleigh-fading channels. Furthermore,

the decoupled noise variances σ_c^2 depend on the fraction w_c of BS antennas associated with cluster c.

The following result establishes the post-fusion SINR in the FD architecture; a short proof is given in Appendix A-D.

Theorem 6. *Let the assumptions of Lemma 5 hold and σ_c^2, $c = 1, \ldots, C$, be the per-cluster decoupled noise variances. Then, the decoupled noise variance σ_{FD}^2 of the fused estimate in (12) of the FD architecture is given by*

$$\sigma_{FD}^2 = \left(\sum_{c=1}^{C} \frac{1}{\sigma_c^2} \right)^{-1} = N_0 + \beta \sum_{c=1}^{C} \nu_c \Psi(\sigma_c^2). \tag{14}$$

We note that this result implies that the post-fusion SINR with optimal fusion according to Definition 3, denoted by $\mathsf{sinr}_{\text{FD}}$, corresponds to the sum of the per-cluster SINR values.

Finally, we have the following intuitive result which implies that for a given equalizer, the FD architecture cannot outperform the PD architecture; the proof is given in Appendix A-E.

Lemma 7. *Let $N_0 > 0$ and assume the large-system limit and Rayleigh-fading channels. Then, the output SINR for the FD and PD architectures satisfy $\mathsf{sinr}_{\text{FD}} \leq \mathsf{sinr}_{\text{PD}}$. Equality holds for $\beta \to 0$, $C = 1$, or if MRC-based equalization is used.*

C. Antenna Partitioning Strategies for Linear Equalizers

We now analyze the post-fusion SINR performance of linear algorithms for the FD architecture, depending on the antenna allocation strategy, i.e., on the fraction of antennas w_c used per cluster c. For the following analysis, we assume the large-system limit and Rayleigh fading channels.

For MRC with the FD architecture, the post-fusion SINR is equivalent to that of the PD architecture (and that of centralized processing), as shown in Lemma 7. Hence, the antenna partitioning strategy does not affect the SINR performance.

For ZF equalization in the FD architecture, we have the following result; the proof is given in Appendix A-F.

Lemma 8. *Assume that the B BS antennas are divided into C clusters so that $w_c \geq \beta$ holds for $c = 1, \ldots, C$. Then, the post-fusion SINR for ZF equalization is given by*

$$\mathsf{sinr}_{\text{FD}}^{\text{ZF}} = \frac{E_s}{N_0}(1 - C\beta). \tag{15}$$

Interestingly, we observe that the post-fusion SINR $\mathsf{sinr}_{\text{FD}}^{\text{ZF}}$ does not depend on the antenna allocation strategy; this implies that the per-cluster antenna fraction w_c can be chosen arbitrarily as long as $w_c \geq \beta$ for $c = 1, \ldots, C$.[3] Note, however, that equally-sized clusters are desirable in practice as they may minimize the interconnect or chip I/O bandwidth as well as the computational complexity per computing fabric.

For L-MMSE equalization in the FD architecture, we have the following result; the proof is given in Appendix A-G.

Lemma 9. *Assume that the B BS antennas are divided into C clusters so that $\sum_{c=1}^{C} w_c = 1$ holds with $w_c \geq 0$ for $c = 1, \ldots, C$. Then, the post-fusion SINR of the L-MMSE equalizer is given by*

$$\mathsf{sinr}_{\text{FD}}^{\text{L-MMSE}} = \frac{1}{2} \sum_{c=1}^{C} \sqrt{\left(1 - \frac{E_s}{N_0}(w_c - \beta)\right)^2 + 4\frac{E_s}{N_0} w_c} - \frac{1}{2}\left(C - \frac{E_s}{N_0}(1 - C\beta)\right) \tag{16}$$

We see from Lemma 9 that the post-fusion SINR expression for L-MMSE equalization depends on the antenna allocation strategy, i.e., on the weights w_c, which is in contrast to ZF equalization (cf. Lemma 8). In addition, L-MMSE equalization does not require the restriction $w_c \geq \beta$ for ZF-equalization as the post-fusion SINR expression holds even for underdetermined systems [26]. Hence, it is natural to ask what the optimal cluster allocation strategy is. The following result is rather disappointing; the proof is given in Appendix A-H.

Lemma 10. *The cluster allocation strategy that maximizes the post-fusion SINR for the L-MMSE equalizer $\mathsf{sinr}_{\text{FD}}^{\text{L-MMSE}}$ is $w_c = 1$ for some c and $w_{c'} = 0$ for $c' \neq c$.*

Clearly, without any systematic requirements on the cluster ratios w_c, $c = 1, \ldots, C$, besides $w_c \geq 0$, maximizing $\mathsf{sinr}_{\text{FD}}^{\text{L-MMSE}}$ so that $\sum_{c=1}^{C} w_c = 1$ corresponds to a centralized architecture, i.e., all antennas should be allocated to a single cluster. Since the key idea of DBP was to mitigate interconnect and I/O bandwidth as well as computation bottlenecks, such an optimal allocation strategy is undesirable in practice. We next show that the most desirable (from a practical viewpoint) cluster allocation strategy, i.e., one for which all clusters have an equal number of antennas, yields the worst post-fusion SINR; the proof is given in Appendix A-I.

Lemma 11. *Assume that the B BS antennas are divided into C clusters so that $\sum_{c=1}^{C} w_c = 1$ with $w_c \geq 0$, $c = 1, \ldots, C$. Then, we have the following lower bound on the post-fusion SINR:*

$$\mathsf{sinr}_{\text{FD}}^{\text{L-MMSE}} \geq \frac{1}{2} \sqrt{\left(1 - \frac{E_s}{N_0}(1 - C\beta)\right)^2 + 4\frac{E_s}{N_0} C} - \frac{1}{2}\left(C - \frac{E_s}{N_0}(1 - C\beta)\right). \tag{17}$$

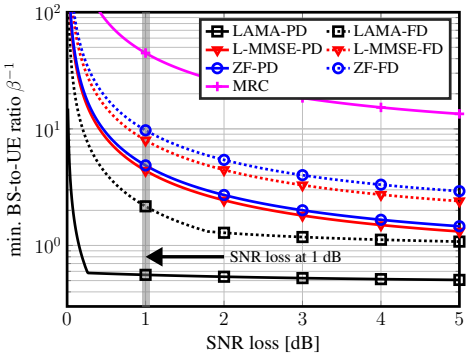
Furthermore, the lower bound is achieved with equality if the antennas are distributed uniformly across all clusters, i.e., where $w_c = 1/C$ for $c = 1, \ldots, C$.

We conclude by noting that even though uniform cluster sizes are the worst for L-MMSE equalization in the FD architecture, L-MMSE equalization outperforms ZF equalization for all possible partitioning schemes in terms of the post-fusion SINR. Hence, L-MMSE equalization is desirable in practice.

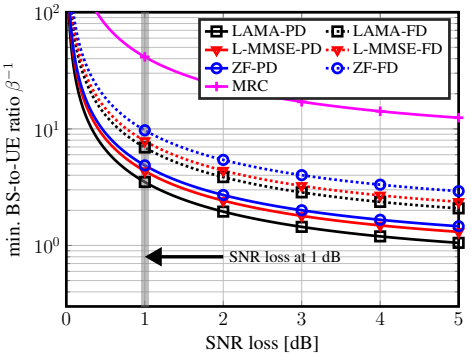
V. Numerical Results

We now investigate the performance of decentralized equalization in the large-system limit and for Rayleigh-fading channels using the SINR expressions from Sections III and IV. We show error-rate simulation results to validate our asymptotic results in finite-dimensional systems. We also provide results for an LTE-like massive MU-MIMO system to demonstrate the efficacy of our solutions in a more realistic scenario.

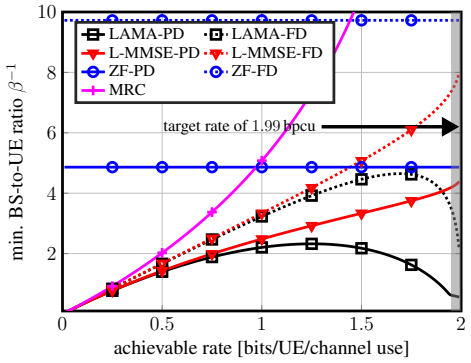
[3]The condition $w_c \geq \beta$ implies that $\sum_{c=1}^{C} w_c = 1 \geq C\beta$ so the SINR expression in Lemma 8 is well-defined.



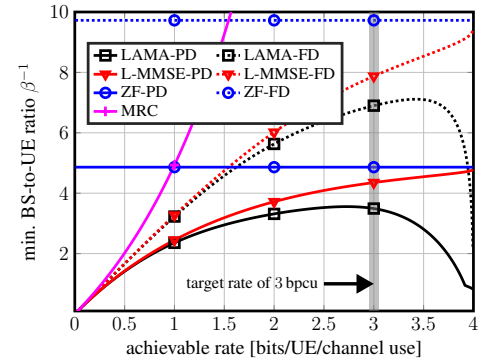
(a) QPSK and target rate of $R = 1.99$ [bits/UE/channel use].



(b) 16-QAM with target rate of $R = 3$ [bits/UE/channel use].



(c) QPSK with fixed 1 dB SNR loss.



(d) 16-QAM with fixed 1 dB SNR loss.

Fig. 2. Achievable rate analysis of DBP with feedforward architectures. Minimum BS-to-UE antenna ratio β^{-1} versus SNR loss for QPSK (a) and 16-QAM (b) for fixed target rates R. Minimum β^{-1} versus achievable rate for QPSK (a) and 16-QAM (b) for a given SNR loss of 1 dB. LAMA outperforms MRC, ZF, and L-MMSE in the FD and PD architectures, especially at high target rates. At low target rates, all equalizers and architectures perform equally well.

A. Achievable Rate Analysis

We first investigate the achievable rates of our feedforward architectures with focus on the large-system limit and Rayleigh fading channels. We consider $C = 2$ clusters with uniform antenna partitioning, i.e., $w_c = 1/C$ for $c = 1, \ldots, C$. We define the average receive signal-to-noise ratio (SNR) as $SNR = \beta E_s / N_0$ and use an interference-free AWGN channel with variance N_0 as the baseline, which coincides to the large-antenna limit of massive MU-MIMO systems with $\beta \to 0$. Concretely, we will use the following performance metric.

Definition 4 (SNR loss)**.** *We define the SNR loss of an equalizer as the excess SNR required to achieve the same target rate R of an interference-free AWGN channel with variance N_0.*

In Fig. 2(a), we use QPSK and a target rate of $R = 1.99$ bits/UE/channel use; in Fig. 2(b) we use 16-QAM and a target rate of $R = 3$ bits/UE/channel use. Both figures investigate the minimum required BS-to-UE ratio β^{-1} for a given SNR loss, which characterizes how many more BS antennas are required by a given equalizer and feedforward architecture to be able to approach AWGN performance up to a given SNR gap. We observe that for a small SNR loss (i.e., when achieving similar performance as that of an interference-free AWGN channel), we require significantly more BS antennas than UEs, irrespective of the algorithm and architecture. For an SNR loss of 1 dB (shown by a thick vertical line in Figures 2(a) and 2(b)), we see that the PD architecture outperforms the FD architecture; this fact is more pronounced in the QPSK scenario as we are trying to achieve 99.5% of the maximum possible rate of 2 bits/UE/channel use for QPSK, whereas for 16-QAM, we are only trying to achieve 75% of the maximum rate. We also see that LAMA-PD significantly outperforms linear equalizers in the PD and FD architectures; MRC requires significantly higher BS-to-UE antenna ratios. Interestingly, for QPSK, LAMA-FD significantly outperforms linear equalization algorithms for the PD architecture in Fig. 2(a) but performs strictly worse for 16-QAM in Fig. 2(b); this is due to the fact that the system-ratio threshold for LAMA to achieve individually-optimal performance is higher for QPSK than for 16-QAM [22]. In summary, LAMA-FD achieves similar performance as linear equalizers with the PD architecture while reducing interconnect and chip I/O bandwidths.

In Fig. 2(c) and Fig. 2(d), we fix the SNR loss to 1 dB and plot the minimum BS-to-UE ratio β^{-1} and varying achievable rates for QPSK and 16-QAM, respectively. For both constellations, MRC performs equally well than all other methods in the low-rate regime; note that the low-rate regime translates to the low-SNR regime for which MRC is known to be optimal. For higher rates, however, MRC requires significantly higher BS-to-UE antenna ratios compared to L-MMSE or LAMA-based equalization. The PD architecture significantly outperforms the FD architecture for all equalizers,

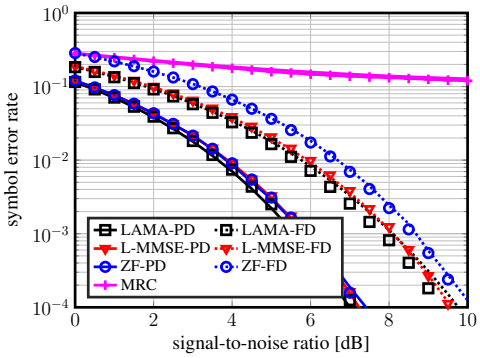


Fig. 3. Symbol error-rate (SER) comparison of equalization in the large-antenna limit (indicated with lines) vs. the simulated performance (indicated by the markers) in a $B = 256$ BS antenna $U = 16$ UE massive MU-MIMO system with Rayleigh-fading channels. Evaluating the SER using our analytical SINR expressions for the large-antenna limit closely matches numerical simulations in finite-dimensional systems for all considered equalizers and architectures.

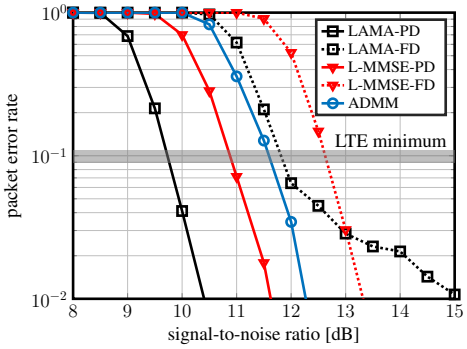


Fig. 4. Packet error-rate (PER) of an LTE-like massive MU-MIMO-OFDM system with $B = 64$, $U = 16$, and 56 k bits per OFDM packet. LAMA for the PD and FD architectures clearly outperforms L-MMSE while meeting the 10% LTE minimum PER requirement. LAMA-PD or L-MMSE clearly outperform the consensus-based ADMM method from [8] (which suffers from transfer latency), whereas LAMA-FD closely approaches the performance at minimal lower latency overheads.

which implies that for high-rates the PD architecture is the preferred choice. Interestingly, the minimum BS-to-UE antenna ratio remains constant for ZF; this implies that as long as one operates below a certain antenna ratio β^*, ZF is able to support all transmission rates; see [28] for additional details on this behavior. Finally, we see that the minimum BS-to-UE ratio β^{-1} decreases for LAMA-FD and LAMA-PD at high rates; this behavior is due to the fact that LAMA in overloaded systems is particularly robust at low and high values of SNR (see [22] for a detailed discussion).

B. Asymptotic vs. Finite-Dimensional Systems

In Fig. 3, we compare our analytical SINR expressions in the large-antenna limit to those in a finite-dimensional massive MU-MIMO scenario. Specifically, we use the SINR from (5) in a decoupled AWGN channel to analytically compute the symbol error-rate (SER) as well as the simulated SER in an uncoded $B = 256$ BS antenna, $U = 16$ UE massive MU-MIMO system with Rayleigh fading channels. We consider $C = 8$ clusters with uniform antenna partitioning, i.e., $w_c = 1/8$ for all $C = 8$ clusters. First, we observe that the simulated results (indicated with markers) closely match our analytical expressions (indicated with lines). Second, we see that the PD architecture significantly outperforms the FD architecture for $C = 8$ clusters and LAMA outperforms ZF and L-MMSE for both architectures. Third, we see that MRC yields poor SER performance, which is due to the fact that MRC requires extremely high BS-to-UE antenna ratios to support 4 bits/UE/channel use for 16-QAM; see also Fig. 2(d).

C. Coded Error-rate Performance in Realistic Systems

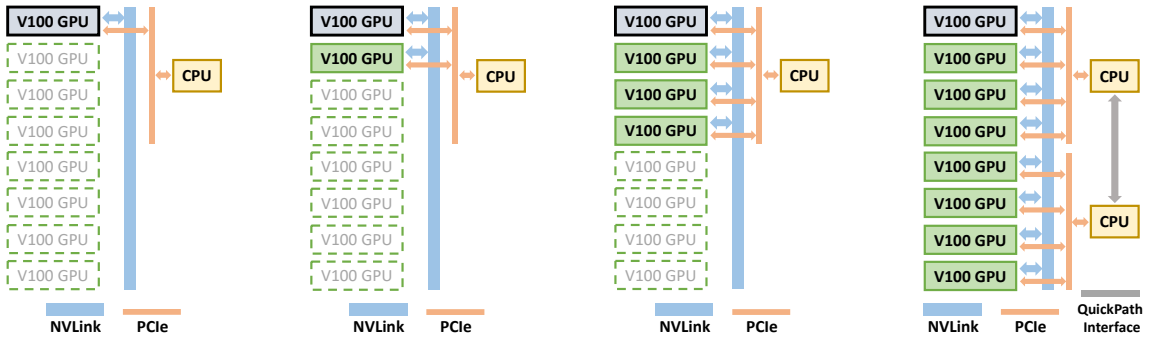
Since our theoretical analysis relies on the large system limit and Rayleigh fading channels, we now investigate the efficacy of our work in a more realistic scenario. Specifically, Fig. 4 shows the coded packet error-rate (PER) in a realistic LTE-like massive MU-MIMO system with $B = 64$ BS antennas and $U = 16$ UEs. We consider $C = 2$ clusters with uniform antenna partitioning. We use OFDM with 2048 subcarriers (1200 used for data transmission) with 16-QAM, 14 OFDM symbols per packet, and use a weak rate-$5/6$ convolutional code with soft-input Viterbi decoding. We consider a WINNER II channel model in an outdoor-to-indoor scenario. For LAMA-PD and FD, we use 10 iterations and perform message damping to mitigate performance loss for finite-dimensional systems with correlated MIMO channel matrices [43]. We observe that LAMA-FD exhibits an error floor near a PER of 10^{-2} due to the small effective system dimension compared to that given by LAMA-PD. Nevertheless, LAMA-FD still outperforms L-MMSE-FD for the LTE minimum PER specification of 10%.

We also compare LAMA and L-MMSE to the consensus-based ADMM method for DBP proposed in [8], where we use 10 iterations. First, we see that LAMA-PD outperforms all other equalization algorithms by a significant margin, when considering the LTE minimum PER specification of 10%. Second, we observe that the consensus-sharing ADMM method performs slightly better than that of LAMA-FD. The ADMM-based method, however, requires iterative consensus exchange among the clusters which results in low throughput; see our GPU implementation results in Section VI-C.

VI. Multi-GPU System Implementation

We now present implementation results of our algorithms and architectures on a multi-GPU system to validate the scalability, throughput, and latency in a realistic scenario. We furthermore provide a comparison to existing consensus-based DBP algorithms and centralized equalizers. In order to fully utilize the available GPU resources, we consider an OFDM-based system as in Section V-C, which enables us to not only parallelize across antennas but also across subcarriers.

Remark 1. *As in [8], the provided multi-GPU implementations serve as a proof-of-concept to assess the efficacy and scalability of our solutions when implemented on real hardware. In practice, we expect that our solutions will be implemented on clusters consisting of field-programmable gate arrays (FPGAs) or application-specific integrated circuits (ASICs), which offer higher computing efficiency and lower transfer latency.*



(a) $C = 1$ (b) $C = 2$ (c) $C = 4$ (d) $C = 8$

Fig. 5. System architecture overview of the multi-GPU experimental platform. The system consists of eight Tesla Volta GPUs with high-bandwidth NvLink GPU-to-GPU interconnect. The master GPU (highlighted) gathers local equalization results and performs all centralized computations. For (a) $C = 1$, (b) $C = 2$, and (c) $C = 4$ clusters, we use a single CPU that controls the GPUs via PCIe; (d) for $C = 8$ clusters, we use two CPUs for control purposes.

A. Experimental Platform

In Fig. 5, we show the used Nvidia DGX-1 multi-GPU system [44] consisting of eight Tesla V100 Volta GPUs with 300 GB/s bi-directional NvLink interconnect and two 20-core Intel Xeon E5-2698 v4 CPUs. Each GPU provides 5120 CUDA cores and 16 GB high bandwidth memory (HBM). In what follows, we use C GPUs when processing C antenna clusters; other partitioning schemes are possible. Our equalization algorithms and architectures are implemented with the CUDA 9.2 toolkit [45] and the Nvidia collective communications library (NCCL) v2.2 [46] software stack. The NCCL uses the message passing interface (MPI) library, which improves the efficiency of inter-GPU collective communication over NvLink [47] by leveraging the CUDA-aware MPI technology [48] for direct GPU-to-GPU memory copy.

B. Implementation Details

All decentralized feedforward equalizers proposed in Sections III and IV are processed in the following three steps: (i) calculate local intermediate results at each antenna cluster, (ii) fuse local results at the centralized processor, and (iii) perform necessary centralized computations to obtain the final equalization outputs. We use the available CPUs to schedule the workload and initialize C MPI processes, each process supervising an individual GPU for design decentralization. The proposed decentralized algorithms were implemented on the multi-GPU system using the following procedure: (i) We accelerate local computations at the C antenna clusters, each using a dedicated GPU with multi-threaded CUDA kernel functions. (ii) We utilize collective inter-process communication among the C GPUs via NCCL over NvLink to realize low-latency gathering of the local results at the master GPU. (iii) We complete the centralized computations at the master GPU for high efficiency. We note that equalization with the PD architecture generally requires fewer local computations, but more data for fusion and computations at the centralized master GPU, compared to equalization with the FD architecture. We now describe the PD and FD architectures in detail. To keep the discussion of the FD architecture short, we only discuss the procedure for LAMA, as the same methodology is used for the linear equalization algorithms. In order to minimize the complexity of computing the post-equalization variance and SINR [36], our implementations resort to the low-complexity approximation put forward in [49].

1) PD Architecture: The local computations at each GPU include the partial MRC output $\mathbf{y}_c^{\text{MRC}} = \mathbf{H}_c\mathbf{y}_c$ and the partial Gram matrix $\mathbf{G}_c = \mathbf{H}_c^{\text{H}}\mathbf{H}_c$. To maintain high GPU occupancy, we aggregate these workloads across a batch of subcarriers and process them in parallel. To process the batched matrix-matrix or matrix-vector multiplications required for partial MRC and Gram computations efficiently, we take advantage of the *cuBLAS* library, a collection of CUDA-accelerated basic linear algebra subprograms (BLAS), and specifically, the `cublasCgemmBatched` library function for fast matrix multiplications with complex-valued floating-point entries. In a single invocation of the batched function call, we calculate $\mathbf{G}_c$ for N_{sc} active subcarriers ($\texttt{batchsize} = N_{\text{sc}}$) associated with a certain OFDM symbol, and reuse them across N_{sym} OFDM symbols within the channel coherence time to reduce complexity. In contrast to the Gram matrix calculation, we compute $\mathbf{y}_c^{\text{MRC}}$ for a total of $N_{\text{sc}} \times N_{\text{sym}}$ subcarriers ($\texttt{batchsize} = N_{\text{sc}} \times N_{\text{sym}}$) in a function call. This is necessary because $\mathbf{y}_c^{\text{MRC}}$ also depends on the receive vector $\mathbf{y}_c$, which varies for each OFDM symbol. We finally fuse the local $\mathbf{G}_c$ and $\mathbf{y}_c^{\text{MRC}}$ that results in a total message size $m_{\text{PD}} = U^2 \times N_{\text{sc}} \times C + U \times N_{\text{sc}} \times N_{\text{sym}} \times C$ at the master GPU with the `ncclReduce` NCCL function.

For LAMA-PD shown in Algorithm 1, each iteration mainly involves matrix-vector multiplication and vector operations, which include vector addition, subtraction, and dot-products. Although matrix-vector multiplications can be efficiently computed by the batched cuBLAS function with $\texttt{batchsize} = N_{\text{sc}} \times N_{\text{sym}}$, as done for preprocessing, we also designed customized multi-threaded kernel functions for the vector operations to fully exploit the on-chip memories of GPU. Specifically, for certain kernels, we combine multiple vector operations which can share intermediate results using fast local registers. Since the intermediate vectors within each LAMA iteration, such as $\mathbf{s}$, $\mathbf{v}$, and $\mathbf{z}$, scale with the UE number U, multi-threaded computations for vector additions,

TABLE I
LATENCY (L) AND THROUGHPUT (TP) PERFORMANCE OF DECENTRALIZED FEEDFORWARD EQUALIZERS ($U = 16$, $B_c = 32$, L IN MS, TP IN GB/S)

	Partially Decentralized (PD) Feedforward Equalizers			Fully Decentralized (FD) Feedforward Equalizers		
B	64	128	256	64	128	256
C	2	4	8	2	4	8
Performance	L / TP	L / TP	L / TP	L / TP	L / TP	L / TP
LAMA, $T_{\text{max}} = 1$	0.651 / 1.76	0.677 / 1.69	0.785 / 1.46	0.591 / 1.94	0.599 / 1.91	0.672 / 1.71
LAMA, $T_{\text{max}} = 2$	0.777 / 1.48	0.803 / 1.43	0.912 / 1.26	0.718 / 1.60	0.727 / 1.58	0.799 / 1.44
LAMA, $T_{\text{max}} = 3$	0.903 / 1.27	0.929 / 1.23	1.039 / 1.10	0.846 / 1.36	0.854 / 1.34	0.925 / 1.24
LAMA, $T_{\text{max}} = 4$	1.030 / 1.11	1.056 / 1.09	1.165 / 0.98	0.973 / 1.18	0.980 / 1.17	1.052 / 1.09
LAMA, $T_{\text{max}} = 5$	1.156 / 0.99	1.183 / 0.97	1.291 / 0.89	1.099 / 1.04	1.106 / 1.03	1.180 / 0.97
L-MMSE	0.719 / 1.60	0.745 / 1.54	0.856 / 1.34	0.666 / 1.72	0.676 / 1.70	0.751 / 1.53
ZF	0.690 / 1.66	0.717 / 1.61	0.827 / 1.39	0.635 / 1.81	0.643 / 1.79	0.715 / 1.61
MRC	0.411 / 2.79	0.421 / 2.72	0.499 / 2.30	0.411 / 2.79	0.421 / 2.72	0.499 / 2.30

TABLE II
PERFORMANCE OF DECENTRALIZED CONSENSUS-SHARING EQUALIZERS DEVELOPED IN [8] ($U = 16$, $B_c = 32$, L IN MS, TP IN GB/S)

B	64	128	256
C	2	4	8
Performance	L / TP	L / TP	L / TP
ADMM, $T_{\text{max}} = 2$	0.783 / 1.47	0.789 / 1.45	0.858 / 1.34
ADMM, $T_{\text{max}} = 3$	0.982 / 1.17	0.995 / 1.15	1.075 / 1.07
CG, $T_{\text{max}} = 2$	0.808 / 1.42	0.815 / 1.41	0.880 / 1.30
CG, $T_{\text{max}} = 3$	0.997 / 1.15	1.010 / 1.14	1.098 / 1.04

TABLE III
PERFORMANCE OF CENTRALIZED EQUALIZERS ($U = 16$, $C = 1$, $B_c = B$, L IN MS, TP IN GB/S)

B	64	128	256
Performance	L / TP	L / TP	L / TP
LAMA, $T_{\text{max}} = 2$	0.766 / 1.50	1.182 / 0.97	2.004 / 0.57
L-MMSE [49]	0.710 / 1.62	1.125 / 1.02	1.947 / 0.59
ZF [49]	0.682 / 1.68	1.095 / 1.05	1.917 / 0.60
MRC	0.457 / 2.51	0.859 / 1.34	1.650 / 0.70

subtractions, and scaling are straightforward; we launch a kernel with $N_{\text{sc}} \times N_{\text{sym}} \times U$ threads to process each of U entries in a vector for a total of $N_{\text{sc}} \times N_{\text{sym}}$ subcarriers in parallel. For the vector dot-product operations, as an example, when we update ϕ in Algorithm 1, we resort to the *warp shuffle* technique, where a thread can directly retrieve register values from another thread efficiently, to obtain a sum reduction of U vector entries across U threads in the same *warp*. If $U >$ `warpsize` $= 32$, then we can also use the on-chip shared memory for the necessary inter-thread communication. After the last LAMA iteration T_{max}, we finally calculate the global LAMA-PD output at the master GPU.

2) FD Architecture: Preprocessing and equalization is computed in a decentralized manner at each of the C GPUs. We reuse the *cuBLAS* library and customized kernel functions for those local computations in LAMA-FD, and fuse the local result $\mathbf{z}_c$ at the master GPU using the NCCL `ncclReduce` function with a smaller total message size $m_{\text{FD}} = U \times N_{\text{sc}} \times N_{\text{sym}} \times C$ than that of LAMA-PD. We implement optimal fusion at the master GPU.

C. Implementation Results

We now present measured latency and throughput of our decentralized feedforward equalizers. In Table I, we list results of different equalizers with PD and FD architectures. These results are obtained via CPU wall clock timing that is synchronized with the start and end of all GPU computations. In what follows, we consider 16-QAM, and we fix the number of UEs to $U = 16$ and BS antennas per cluster to $B_c = 32$ We simulate an LTE-like system as in Section V-C with $N_{\text{sym}} = 14$ OFDM symbols per packet and $N_{\text{sc}} = 1200$ active subcarriers (out of 2048 subcarriers), which corresponds to a 20 MHz LTE subframe. To compare the scalability of our proposed feedforward equalization architectures, we scale the total number of BS antennas $B = CB_c$ by increasing the number of clusters from $C = 2$, $C = 4$, to $C = 8$.

1) Throughput and Latency: We see from Table I that all of our proposed decentralized feedforward equalizers achieve throughput in the Gb/s regime with latencies of 1 ms or less. Specifically, the non-linear LAMA-PD and LAMA-FD equalizers are able to reach higher throughput for a small number of iterations ($T_{\text{max}} = 1$ or $T_{\text{max}} = 2$). We note that for more LAMA iterations, our decentralized linear equalizers are able to outperform LAMA in terms of the throughput. This is due to the fact that linear equalizers can reuse both local Gram matrix multiplication and matrix inversion results across N_{sym} OFDM symbols, which effectively reduces complexity. Among all those feedforward equalizers, MRC (which is equivalent for MRC-PD and MRC-FD) has the highest throughput and lowest latency, but entails a significant error-rate performance loss; see the simulation results in Section V. Furthermore, we see that equalization with the FD architecture generally achieves higher throughput than that of the PD architecture, mainly caused by smaller message sizes and lower data-transfer latency during the data fusion process. This advantage, however, comes at the cost of reduced error-rate performance for the FD architecture. Put simply, there exists a trade-off between throughput and error-rate performance for the FD and PD architectures.

Interestingly, the latency and throughput of our decentralized feedforward equalization implementations degrades only slightly when we increase the number of BS antennas (with a larger number of clusters C) as our designs benefit from direct GPU-to-GPU communication with efficient NCCL over NvLink.

This observation demonstrates that our proposed decentralized feedforward equalizers have excellent scalability to support hundreds to even thousands of BS antennas while maintaining high throughput.

We now compare our proposed decentralized feedforward architectures to the decentralized consensus-based methods proposed in [8]. In Table II, we show reference latency and throughput results measured on the same multi-GPU platform. As discussed in [8], consensus-sharing methods, such as ADMM-based and decentralized CG-based equalizers, rely on iterative update of local variables and global consensus, which requires the data gathering for consensus calculation and consensus broadcasting *in each iteration*. In contrast, the decentralized feedforward equalizers proposed in this paper significantly reduce the data transfer latency with only *one-shot* (feedforward) message passing, which leads to (often significantly) higher throughput. More specifically, both LAMA-PD and LAMA-FD achieve higher throughput than D-ADMM or D-CG for the same number of iterations; the same trend continues to hold for a larger number of iterations.

In Table III, we show the latency and throughput performance of several *centralized* equalizers. We see that the throughput of centralized equalizers decreases quickly when increasing the number of BS antennas; this demonstrates that centralized solutions exhibit poor scalability to large BS antenna arrays. Also note that centralized solutions suffer from high interconnect and chip I/O data rates, which is not visible in this comparison.

2) Data Transfer: The key advantages of the proposed decentralized feedforward architectures are as follows: (i) They reduce the raw front-end baseband transfer rate by a factor of C at each antenna cluster compared to that of centralized architectures and (ii) they minimize the back-end message passing data among clusters. For example, given $N_{\text{sc}} = 1200$, $N_{\text{sym}} = 14$, $U = 16$ and $C = 4$, the total message passing data across C GPUs with the PD architecture is $m_{\text{PD}} = 17.58\,\text{MB}$ while the FD architecture requires only $m_{\text{FD}} = 8.20\,\text{MB}$. In contrast, the decentralized consensus-sharing (CS) equalizers in [8] require multiple iterations of message passing. Furthermore, in each each iteration the message passing data of both data gathering and consensus broadcasting doubles compared to that of the FD architecture, i.e., we have $m_{\text{CS}} = 2m_{\text{FD}} = 16.40\,\text{MB}$ per iteration.

3) Practical Considerations: While our multi-GPU implementations exceed 1 Gb/s throughput, the estimated total power dissipation of C Tesla V100 GPUs with 300 W thermal design power (TDP) each can be as high as $C \times 300\,\text{W}$. To reduce the power consumption in practice, one can resort to FPGA or ASIC implementations. In fact, some of the latest massive MU-MIMO testbeds, such as the LuMaMi [15], [16] testbed, are built with FPGA-driven software-defined radios and PXIe switches, and demonstrate 0.8 Gb/s throughput for a $B = 100$ antenna $U = 12$ user system. However, the LuMaMi testbed processes the entire workload for different sub-bands (a small portion of the entire frequency band) on different FPGA co-processors, while each sub-band still connects to all BS antennas. Thus the PXIe switches can be saturated when further scaling up the antenna number. In contrast, the proposed feedforward architectures divide the processing workload *across antennas*, which yields improved scalability and modularity. In practice, one can combine antenna domain and frequency domain decentralization to combine the best of both approaches.

VII. Conclusions

We have presented two feedforward architectures for decentralized equalization in massive MU-MIMO systems that mitigate the interconnect and I/O bandwidth bottlenecks and enable parallel processing on multiple computing fabrics. For the two proposed architectures, we have presented linear and nonlinear equalization algorithms, and we have analyzed their post-equalization SINR performance in the large-antenna limit. We have also performed numerical simulations that confirm our analysis. Our results indicate that nonlinear equalizers are able to achieve near-optimal SINR performance while enabling decentralized computations and low communication overhead among the antenna clusters. Linear equalizers perform equally well for scenarios in which the number of BS antennas is significantly larger than the number of UEs or for systems that use strong coding or low data rates. Our reference implementations on a multi-GPU system have shown that our feedforward architectures achieve throughputs in the Gb/s regime, even for massive MU-MIMO systems with hundreds of antenna elements. Specifically, our measurement results show that feedforward architectures are able to overcome the latency limits of existing decentralized baseband processing schemes, such as the consensus-sharing methods proposed in [8].

There are many avenues for future work. First, an implementation of our algorithms and architectures on multi-FPGA or multi-ASIC platforms would demonstrate the full capabilities of our solutions. Second, an investigation of antenna partitioning schemes and beamforming-based methods for directional channels, such as those experienced in millimeter wave (mmWave) systems, is part of ongoing research. Third, a theoretical analysis of the precoding architectures and algorithms proposed recently in [50] for the massive MU-MIMO downlink is left for future work.

Acknowledgments

The work of C. Jeon and C. Studer was supported in part by Xilinx, Inc. and by the US National Science Foundation (NSF) under grants ECCS-1408006, CCF-1535897, CCF-1652065, CNS-1717559, and ECCS-1824379. The work of K. Li and J. R. Cavallaro was supported in part by Xilinx, Inc. and by the US NSF under grants ECCS-1408370, CNS-1717218, and CNS-1827940, for the "PAWR Platform POWDER-RENEW: A Platform for Open Wireless Data-driven Experimental Research with Massive MIMO Capabilities." The authors thank T. Goldstein for discussions on DBP, and we acknowledge the hardware support of the DGX-1 multi-GPU systems at the Nvidia Technology Center (the PSG Cluster).

APPENDIX A
DERIVATIONS AND PROOFS

A. Derivation of Algorithm 1

Algorithm 1 builds upon the original LAMA algorithm [22]:

$$\begin{aligned}\mathbf{z}^t &= \mathbf{s}^t + \mathbf{H}^{\mathrm{H}}\mathbf{r}^t \\ \mathbf{s}^{t+1} &= \mathsf{F}(\mathbf{z}^t, N_0(1+\tau^t)) \\ \tau^{t+1} &= \frac{\beta}{N_0}\langle \mathsf{G}(\mathbf{z}^t, N_0(1+\tau^t))\rangle \\ \mathbf{r}^{t+1} &= \mathbf{y} - \mathbf{H}\mathbf{s}^{t+1} + \frac{\tau^{t+1}}{1+\tau^t}\mathbf{r}^t.\end{aligned}$$

We start with $\mathbf{y}^{\mathrm{MRC}} = \mathbf{H}^{\mathrm{H}}\mathbf{y}$ and $\mathbf{G} = \mathbf{H}^{\mathrm{H}}\mathbf{H}$ and define $\phi^t = \frac{N_0\tau^t}{\beta}$ so that $\frac{\tau^{t+1}}{1+\tau^t} = \frac{\beta\phi^{t+1}}{N_0+\beta\phi^t}$. Then,

$$\begin{aligned}\mathbf{H}^{\mathrm{H}}\mathbf{r}^{t+1} &= \mathbf{H}^{\mathrm{H}}\mathbf{y} - \mathbf{H}^{\mathrm{H}}\mathbf{H}\mathbf{s}^{t+1} + \frac{\tau^t}{1+\tau^t}\mathbf{H}^{\mathrm{H}}\mathbf{r}^t \\ &= \mathbf{y}^{\mathrm{MRC}} - \mathbf{G}\mathbf{s}^{t+1} + \frac{\beta\phi^{t+1}}{N_0+\beta\phi^t}\mathbf{H}^{\mathrm{H}}\mathbf{r}^t.\end{aligned}$$

Algorithm 1 follows by noticing that $\mathbf{v}^{t+1} = \mathbf{H}^{\mathrm{H}}\mathbf{r}^t = \mathbf{z}^t - \mathbf{s}^t$.

B. Proof of Lemma 4

As in (5), we write the estimate $z_{c,u}$ for UE u at cluster c as $z_{c,u} = s_{0u} + e_{c,u}$, where $e_{c,u}$ represents residual interference and noise with known error variance $\mathbb{E}\big[|e_{c,u}|^2\big] = \sigma_{c,u}^2$. At UE u, optimal fusion is $z_u = \sum_{c=1}^C \nu_{c,u} z_{c,u}$ so that $\sum_{c=1}^C \nu_{c,u} = 1$. Hence, the fused estimate is $z_u = s_{0u} + \sum_{c=1}^C \nu_{c,u} e_{c,u}$, with the following post-fusion SINR:

$$\mathsf{sinr} = \frac{\mathbb{E}\big[|s_{0u}|^2\big]}{\mathbb{E}\Big[\big|\sum_{c=1}^C \nu_{c,u} e_{c,u}\big|^2\Big]} = \frac{E_s}{\sum_{c=1}^C \nu_{c,u}^2 \sigma_{c,u}^2}. \tag{18}$$

Here, we used the assumption that the residual interference and noise terms $e_{c,u}$ are zero mean and uncorrelated across clusters $c = 1, \ldots, C$. We are now interested in maximizing the post-fusion SINR in (18) subject to $\sum_{c=1}^C \nu_{c,u} = 1$. Using the method of Lagrange multipliers, it is easy to see that

$$\nu_{c,u} = \frac{1}{\sigma_{c,u}^2}\left(\sum_{c'=1}^C \frac{1}{\sigma_{c',u}^2}\right)^{-1}, \quad c = 1, \ldots, C. \tag{19}$$

C. Proof of Lemma 5

For Rayleigh-fading channels, each entry in the partial channel matrix $\mathbf{H}_c$ is distributed as $\mathcal{CN}(0, 1/B)$. To ensure that the expected column-norm of $\mathbf{H}_c$ is one, we normalize the per-cluster input-output relation in (3) by $1/\sqrt{w_c}$. This normalization amplifies the noise variance by $1/w_c$ in each cluster. In addition, since overall system dimension is $Bw_c \times U$, the resulting system ratio is given by $\beta = U/(Bw_c) = \beta/w_c$. By realizing that N_0/w_c is the per-cluster noise variance, the fixed-point equation follows immediately from Theorems 1 and 3 for linear and LAMA-based equalization, respectively.

D. Proof of Theorem 6

To simplify notation, we omit the UE index u. The proof follows from (19) in Lemma 4. The first expression in (14) is trivial whereas the second expression is obtained as follows:

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

E. Proof of Lemma 7

We first show when equality holds. The case for $C = 1$ is trivial because the PD and FD architectures are equivalent for $C = 1$. The case for $\beta \to 0$ is also straightforward because $\sigma_c^2 = \sigma_{\mathrm{FD}}^2 = \sigma_{\mathrm{PD}}^2 = N_0$. For MRC, we have $\sigma_{\mathrm{FD}}^2 = N_0 + \beta\sum_{c=1}^C \nu_c \,\mathbb{V}\mathrm{ar}_S[S] = N_0 + \beta\,\mathbb{V}\mathrm{ar}_S[S] = \sigma_{\mathrm{PD}}^2$.

Let us now assume that $\beta > 0$. We show that $\sigma_c^2 > \sigma_{\mathrm{PD}}^2$ by re-writing the fixed-point solutions as [51]: $\sigma_c^2 = \sup\{\sigma^2 : N_0 + \beta\Psi(\sigma^2) \geq w_c\sigma^2\}$ and $\sigma_{\mathrm{PD}}^2 = \sup\{\sigma^2 : N_0 + \beta\Psi(\sigma^2) \geq \sigma^2\}$. Note that $N_0 > 0$, so both σ_c^2 and σ_{PD}^2 are strictly positive. It is easy to see that $\sigma_{\mathrm{PD}}^2 \neq \sigma_c^2$ because $\sigma_{\mathrm{PD}}^2 = N_0 + \beta\Psi(\sigma_{\mathrm{PD}}^2) > w_c\sigma_{\mathrm{PD}}^2$. Since $\Psi(\sigma^2) \to \mathbb{V}\mathrm{ar}_S[S]$ as $\sigma^2 \to \infty$ and $\Psi(\sigma^2)$ is continuous [52], there exists a $\sigma_c^2 > \sigma_{\mathrm{PD}}^2$ that satisfies $N_0 + \beta\Psi(\sigma_c^2) = w_c\sigma_c^2$ by the intermediate value theorem.

Finally, we use [52, 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 inspection of $\mathrm{d}\Psi(\sigma^2)/\mathrm{d}\sigma^2 > 0$. Thus, the result $\sigma_{\mathrm{FD}}^2 > \sigma_{\mathrm{PD}}^2$ follows directly from Lemma 4 since

$$\sigma_{\mathrm{FD}}^2 = N_0 + \beta\sum_{c=1}^C \nu_c \Psi(\sigma_c^2) > N_0 + \beta\sum_{c=1}^C \nu_c \Psi(\sigma_{\mathrm{PD}}^2) = \sigma_{\mathrm{PD}}^2.$$

F. Proof of Lemma 8

The proof is straightforward and follows from Theorem 1 and Lemma 4. Given that cluster c has $Bw_c > \beta$ antennas across all clusters C, the input-output relation of cluster c in the large-system limit under ZF equalization results in a AWGN channel with decoupled noise variance: $\sigma_c^2 = \frac{N_0}{w_c - \beta}$. The proof follows from Lemma 4 noting that $\sum_{c=1}^C w_c = 1$:

$$\sigma_{\mathrm{FD}}^2 = \left(\sum_{c=1}^C \frac{1}{\sigma_c^2}\right)^{-1} = \left(\sum_{c=1}^C \frac{w_c - \beta}{N_0}\right)^{-1} = \frac{N_0}{1 - C\beta}.$$

G. Proof of Lemma 9

The proof follows Theorem 1 with the fixed-point equation

$$w_c\sigma_c^2 = N_0 + \beta\frac{E_s}{E_s + \sigma_c^2}\sigma_c^2,$$

which results in the following sinr expression for cluster C for L-MMSE with the FD architecture:

$$\begin{aligned}\mathsf{sinr}_{\mathrm{FD},c}^{\text{L-MMSE}} = \frac{1}{2}\Bigg(&\sqrt{\left(1 - \frac{E_s}{N_0}(w_c - \beta)\right)^2 + 4\frac{E_s}{N_0}w_c} \\ &- \left(1 - \frac{E_s}{N_0}(w_c - \beta)\right)\Bigg).\end{aligned}$$

Lemma 9 follows from $\mathsf{sinr}_{\mathrm{FD}}^{\text{L-MMSE}} = \sum_{c=1}^C \mathsf{sinr}_{\mathrm{FD},c}^{\text{L-MMSE}}$.

H. Proof of Lemma 10

The proof of Lemma 10 starts from (16). Let us denote $\overline{\mathsf{sinr}}_{\text{FD}}^{\text{L-MMSE}}$ as $\mathsf{sinr}_{\text{FD}}^{\text{L-MMSE}}$ when $w_1 = 0$ and $w_2 = \cdots = w_C = 0$. We also define $\overline{\beta} = 1 - \beta$. Then,

$$\begin{aligned}\max_{\mathbf{w}} \mathsf{sinr}_{\text{FD}}^{\text{L-MMSE}} &\geq \overline{\overline{\mathsf{sinr}}}_{\text{FD}}^{\text{L-MMSE}} \\ &= \frac{1}{2}\left(\sqrt{\left(1-\frac{E_s}{N_0}\overline{\beta}\right)^2 + 4\frac{E_s}{N_0}} - \left(1-\frac{E_s}{N_0}\overline{\beta}\right)\right) \\ &\quad + \frac{C-1}{2}\left(\sqrt{\left(1-\frac{E_s}{N_0}\beta\right)^2} - \left(1-\frac{E_s}{N_0}\beta\right)\right) \\ &= \frac{1}{2}\left(\sqrt{\left(1-\frac{E_s}{N_0}\overline{\beta}\right)^2 + 4\frac{E_s}{N_0}} - \left(1-\frac{E_s}{N_0}\overline{\beta}\right)\right) \\ &\overset{(a)}{=} \mathsf{sinr}_{\text{PD}}^{\text{L-MMSE}},\end{aligned}$$

where (a) follows from (L-MMSE) in Corollary 2. Since we know from Lemma 7 that $\mathsf{sinr}_{\text{PD}}^{\text{L-MMSE}} \geq \mathsf{sinr}_{\text{FD}}^{\text{L-MMSE}}$, we have that $\max_{\mathbf{w}} \mathsf{sinr}_{\text{FD}}^{\text{L-MMSE}} = \overline{\overline{\mathsf{sinr}}}_{\text{FD}}^{\text{L-MMSE}} = \mathsf{sinr}_{\text{PD}}^{\text{L-MMSE}}$.

I. Proof of Lemma 11

The proof of Lemma 11 starts from (16) with the definition $f(w,\alpha) = \sqrt{(1-\alpha(w-\beta))^2 + 4\alpha w}$. Note that $f(w,\alpha)$ is convex in w as $f''(w,\alpha) \geq 0$ for $\alpha \geq 0$. The final step follows from Jensen's inequality, which implies

$$\frac{1}{C}\sum_{c=1}^{C} f(w_c,\alpha) \geq f\left(\frac{1}{C}\sum_{c=1}^{C} w_c, \alpha\right),$$

where equality holds if $w_1 = w_2 = \cdots = w_C$.

REFERENCES

[1] C. Jeon, K. Li, J. R. Cavallaro, and C. Studer, "On the achievable rates of decentralized equalization in massive MU-MIMO systems," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2017, pp. 1102–1106.

[2] T. L. Marzetta, "Non-cooperative cellular wireless with unlimited numbers of base station antennas," *IEEE Trans. Wireless Comm.*, vol. 9, no. 11, pp. 3590–3600, Nov. 2010.

[3] L. Lu, G. Y. Li, A. L. Swindlehurst, A. Ashikhmin, and R. Zhang, "An Overview of Massive MIMO: Benefits and Challenges," *IEEE J. Sel. Topics in Sig. Proc.*, vol. 8, no. 5, pp. 742–758, Oct. 2014.

[4] 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.

[5] 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.

[6] 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.

[7] A. Puglielli, N. Narevsky, P. Lu, T. Courtade, G. Wright, B. Nikolic, and E. Alon, "A scalable massive MIMO array architecture based on common modules," in *IEEE Intl. Conf. Commun. Workshop (ICCW)*, June 2015, pp. 1310–1315.

[8] K. Li, R. R. Sharan, Y. Chen, T. Goldstein, J. R. Cavallaro, and C. Studer, "Decentralized baseband processing for massive MU-MIMO systems," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 7, no. 4, pp. 491–507, Nov. 2017.

[9] L. Van der Perre, L. Liu, and E. G. Larsson, "Efficient DSP and circuit architectures for massive MIMO: State-of-the-art and future directions," *IEEE Trans. Signal Process.*, vol. 66, no. 18, pp. 4717–4736, Sept. 2018.

[10] S. Jacobsson, Y. Ettefagh, G. Durisi, and C. Studer, "All-digital massive MIMO with a fronthaul constraint," in *Proc. IEEE Workshop Stat. Signal Process. (SSP)*, Jun. 2018.

[11] http://www.cpri.info, *Common public radio interface*.

[12] A. Puglielli, A. Townley, G. LaCaille, V. Milovanović, P. Lu, K. Trotskovsky, A. Whitcombe, N. Narevsky, G. Wright, T. Courtade *et al.*, "Design of energy-and cost-efficient massive MIMO arrays," *Proc. IEEE*, vol. 104, no. 3, pp. 586–606, Dec. 2016.

[13] M. Wu, B. Yin, G. Wang, C. Dick, J. Cavallaro, and C. Studer, "Large-scale MIMO detection for 3GPP LTE: Algorithm and FPGA implementation," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 916–929, Oct. 2014.

[14] 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.

[15] 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*, Dec. 2014, pp. 287–293.

[16] S. Malkowsky, J. Vieira, L. Liu, P. Harris, K. Nieman, N. Kundargi, I. C. Wong, F. Tufvesson, V. Öwall, and O. Edfors, "The world's first real-time testbed for massive MIMO: Design, implementation, and validation," *IEEE Access*, vol. 5, pp. 9073–9088, 2017.

[17] Q. Yang, X. Li, H. Yao, J. Fang, K. Tan, W. Hu, J. Zhang, and Y. Zhang, "BigStation: Enabling scalable real-time signal processing in large MU-MIMO systems," in *Proc. ACM SIGCOMM Conf.*, 2013, pp. 399–410.

[18] 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.

[19] 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.*, Dec. 2016, pp. 590–594.

[20] 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.

[21] E. Bertilsson, O. Gustafsson, and E. G. Larsson, "A scalable architecture for massive MIMO base stations using distributed processing," in *Proc. Asilomar Conf. Signals, Syst., Comput.*, Nov. 2016, pp. 864–868.

[22] 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.

[23] 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.

[24] 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.

[25] S. Verdú and S. Shamai, "Spectral efficiency of CDMA with random spreading," *IEEE Trans. Inf. Theory*, vol. 45, no. 2, pp. 622–640, Mar. 1999.

[26] 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.

[27] S. Shamai and S. Verdu, "The impact of frequency-flat fading on the spectral efficiency of CDMA," *IEEE Trans. Inf. Theory*, vol. 47, no. 4, pp. 1302–1327, May 2001.

[28] R. Ghods, C. Jeon, G. Mirza, A. Maleki, and C. Studer, "Optimally-tuned nonparametric linear equalization for massive MU-MIMO systems," in *IEEE Intl. Symposium on Info. Theory (ISIT)*, June 2017, pp. 2118–2122.

[29] S. Wu, L. Kuang, Z. Ni, J. Lu, D. Huang, and Q. Guo, "Low-complexity iterative detection for large-scale multiuser MIMO-OFDM systems using approximate message passing," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 902–915, Oct. 2014.

[30] 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.

[31] L. Liu, C. Yuen, Y. L. Guan, Y. Li, and C. Huang, "Gaussian message passing for overloaded massive MIMO-NOMA," *IEEE Trans. Wireless Commun.*, vol. 18, no. 1, pp. 210–226, Jan. 2019.

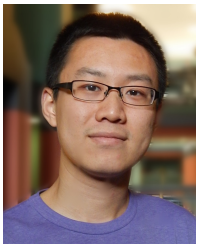
[32] J. Zhu, A. Beirami, and D. Baron, "Performance trade-offs in multi-processor approximate message passing," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2016, pp. 680–684.

[33] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[34] E. J. Candès, J. Romberg, and T.Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inf. Theory*, vol. 52, pp. 489–509, Jan. 2006.

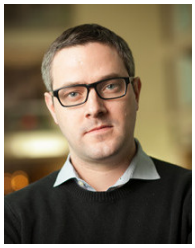
[35] A. Paulraj, R. Nabar, and D. Gore, *Introduction to Space-Time Wireless Communications*. Cambridge Univ. Press, 2003.
[36] C. Studer, S. Fateh, and D. Seethaler, "ASIC implementation of soft-input soft-output MIMO detection using MMSE parallel interference cancellation," *IEEE J. Solid-State Circuits*, vol. 46, no. 7, pp. 1754–1765, 2011.
[37] Z. Wu, C. Zhang, Y. Xue, S. Xu, and X. You, "Efficient architecture for soft-output massive MIMO detection with Gauss-Seidel method," in *Proc. IEEE Int. Symp. Circuits and Syst. (ISCAS)*, May 2016, pp. 1886–1889.
[38] J. Pan, W.-K. Ma, and J. Jalden, "MIMO detection by Lagrangian dual maximum-likelihood relaxation: Reinterpreting regularized lattice decoding," *IEEE Trans. Signal Process.*, vol. 62, no. 2, pp. 511–524, Nov. 2014.
[39] 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.
[40] Y. Eldar and A. Chan, "On the asymptotic performance of the decorrelator," *IEEE Trans. Inf. Theory*, vol. 49, no. 9, pp. 2309–2313, Sep. 2003.
[41] D. Donoho, A. Maleki, and A. Montanari, "Message passing algorithms for compressed sensing: I. Motivation and construction," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Jan. 2010, pp. 1–5.
[42] 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.
[43] J. Vila, P. Schniter, S. Rangan, F. Krzakala, and L. Zdeborova, "Adaptive damping and mean removal for the generalized approximate message passing algorithm," in *IEEE Int'l Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2015, pp. 2021–2025.
[44] *Nvidia DGX-1 system*. [Online]. Available: https://www.nvidia.com/en-us/data-center/dgx-1
[45] *Nvidia CUDA programming guide*. [Online]. Available: http://docs.nvidia.com/cuda
[46] *Nvidia collective communications library*. [Online]. Available: https://developer.nvidia.com/nccl
[47] *Nvidia NvLink fabric*. [Online]. Available: https://www.nvidia.com/en-us/data-center/nvlink
[48] *CUDA-aware MPI*. [Online]. Available: https://devblogs.nvidia.com/parallelforall/introduction-cuda-aware-mpi
[49] K. Li, B. Yin, M. Wu, J. R. Cavallaro, and C. Studer, "Accelerating massive MIMO uplink detection on GPU for SDR systems," in *Proc. IEEE Dallas Circuits Syst. Conf. (DCAS)*, Oct 2015, pp. 1–4.
[50] K. Li, C. Jeon, J. R. Cavallaro, and C. Studer, "Feedforward architectures for decentralized precoding in massive MU-MIMO systems," *Proc. Asilomar Conf. Signals, Syst., Comput.*, pp. 1659–1665, Oct. 2018.
[51] A. Maleki, "Approximate message passing algorithms for compressed sensing," Ph.D. dissertation, Stanford University, Jan. 2011.
[52] 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.

Charles Jeon (S'12) received his M.S. degree in electrical engineering and B.S. degree in electrical engineering and mathematics both at the University of Pennsylvania in 2013. During the same year, he received his B.S. degree in economics from The Wharton School. Dr. Jeon received his Ph.D. degree in electrical and computer engineering at Cornell University, Ithaca, NY, in 2019. He is currently a research scientist with Intel Labs, Hillsboro, OR. His research interests include wireless communications, signal processing, and digital VLSI design.



Kaipeng Li (S'15) received his B.S. degree in physics from Nanjing University, Nanjing, China, in 2013, and his M.S. and Ph.D. degrees in electrical and computer engineering from Rice University in 2015 and 2019, respectively. His research interests include statistical signal processing, parallel and distributed computing on GPGPU and multicore CPU platforms, VLSI design, software-defined radios, and massive MIMO systems.

Joseph R. Cavallaro (S'78, M'82, SM'05, F'15) received the B.S. degree from the University of Pennsylvania, Philadelphia, PA, in 1981, the M.S. degree from Princeton University, Princeton, NJ, in 1982, and the Ph.D. degree from Cornell University, Ithaca, NY, in 1988, all in electrical engineering. He is an IEEE Fellow. From 1981 to 1983, he was with AT&T Bell Laboratories, Holmdel, NJ. In 1988, he joined the faculty of Rice University, Houston, TX, where he is currently a Professor of electrical and computer engineering and Associate Chair. His research interests include computer arithmetic, and DSP, GPU, FPGA, and VLSI architectures for applications in wireless communications. During the 1996–1997 academic year, he served at the US National Science Foundation as Director of the Prototyping Tools and Methodology Program. He was a Nokia Foundation Fellow and a Visiting Professor at the University of Oulu, Finland in 2005 and continues his affiliation there as an Adjunct Professor. He is currently the Director of the Center for Multimedia Communication at Rice University. He is an advisory board member of the IEEE SPS TC on Design and Implementation of Signal Processing Systems and the Chair of the IEEE CAS TC on Circuits and Systems for Communications. He was an Associate Editor of the IEEE Transactions on Signal Processing and the IEEE Signal Processing Letters, and currently serves as an AE for the Journal of Signal Processing Systems. He was General/Program Co-chair of the 2003, 2004, and 2011 IEEE International Conference on Application-Specific Systems, Architectures and Processors (ASAP), and General/Program Co-chair for the 2012, 2014 ACM/IEEE GLSVLSI conferences. At the IEEE SiPS workshop, he was TPC Co-Chair in 2016 and General Co-Chair in 2020. At the IEEE Asilomar Conference on Signals, Systems, and Computers, he was TPC Chair in 2017 and General Chair in 2020. He served on the IEEE CAS Society Board of Governors during 2014.



Christoph Studer (S'06, M'10, SM'14) received his Ph.D. degree in Information Technology and Electrical Engineering from ETH Zurich in 2009. In 2005, he was a Visiting Researcher with the Smart Antennas Research Group at Stanford University. From 2006 to 2009, he was a Research Assistant in both the Integrated Systems Laboratory and the Communication Technology Laboratory (CTL) at ETH Zurich. From 2009 to 2012, Dr. Studer was a Postdoctoral Researcher at CTL, ETH Zurich, and the Digital Signal Processing Group at Rice University. In 2013, he has held the position of Research Scientist at Rice University. From 2014 to 2019, Dr. Studer was an Assistant Professor at Cornell University. Since 2019, he is an Assistant Professor at Cornell Tech in New York City. Since 2014, he is also an Adjunct Assistant Professor at Rice University. Dr. Studer's research interests include the design of very large-scale integration (VLSI) circuits, as well as wireless communications, signal and image processing, optimization, and machine learning.

Dr. Studer received ETH Medals for his M.S. and Ph.D. theses in 2006 and 2009, respectively. He received a Swiss National Science Foundation fellowship for Advanced Researchers in 2011 and a US National Science Foundation CAREER Award in 2017. Dr. Studer won a Michael Tien '72 Excellence in Teaching Award from the College of Engineering, Cornell University, in 2016. He shared the Swisscom/ICTnet Innovations Award in both 2010 and 2013. Dr. Studer was the winner of the Student Paper Contest of the 2007 Asilomar Conf. on Signals, Systems, and Computers, received a Best Student Paper Award of the 2008 IEEE Int. Symp. on Circuits and Systems (ISCAS), and shared the best Live Demonstration Award at the IEEE ISCAS in 2013.