Towards 6G MIMO Systems, with Massively-Parallel Non-Linear Processing

J.C. De Luna Ducoing, Chathura Jayawardena, Marcin Filo and Konstantinos Nikitopoulos
Wireless Systems Lab, 5G and 6G Innovation Centre, Institute for Communication Systems,
University of Surrey, Guildford GU2 7XH, UK

Abstract—Next-generation 6G networks are expected to feature an extremely high density of network and user devices. MU-MIMO non-linear processing can provide substantially improved performance over linear processing in dense conditions, but suffers from a high complexity and processing latency. The use of the massively parallel non-linear (MPNL) processing framework can overcome such limitations. This work discusses three potential 6G transmission scenarios and evaluates their detection and precoding performance using link-level simulations and a system-level, over-the-air, 3GPP standards–based testbed. The results validate that MPNL processing has the potential to transform the way 6G MU-MIMO systems are designed.

I. INTRODUCTION

With the deployment of 5G mobile networks underway, research is gradually shifting to the next-generation wireless networks. New research visions for 6G mobile include use cases for remote presence and extended reality (XR), in a fusion of virtual and physical worlds [1]. New technologies may enable these new cases, which include terahertz (THz) communication [2], large intelligent surfaces, ubiquitous artificial intelligence (AI) across all communication layers, among others. Key performance metrics, such as throughput, reliability, and latency, will also continue to be improved.

Nevertheless, this work addresses another aspect of future 6G networks: they are expected to become extremely dense. Network equipment will be placed closer together as cell sizes will shrink, and the number of connected devices will increase, driven by the Internet of Things (IoT) [3] and machine-type communications (MTC) [4], along with traditional human-centric communications. To deal with such high connectivity requirements, future generations of mobile networks will need to use the time, frequency, and spatial dimensions in a highly efficient manner.

Multiple-user, multiple-input, multiple-output (MU-MIMO) is a spatial-domain technology that can be key for meeting such challenges in connectivity. A successful example is massive MIMO [5], which has been widely deployed in 5G mobile networks. Its working principle is that “extra base station antennas are always beneficial” [6]. By deploying a much larger number of base station (BS) antennas than the number of data streams, the propagation channel of massive MIMO is usually well-conditioned, which is because the effective channel matrix is highly over-determined. Consequently, when the number of BS antennas is much higher than the number of users, massive MIMO can achieve near-optimal signal detection (in the uplink) and precoding (in the downlink) while employing low-complexity linear processing, such as zero forcing (ZF) and linear minimum mean square error (MMSE).

However, where there is a high number of simultaneous spatial streams, comparable to the number of BS antennas, as envisioned in a 6G user-dense network, the achievable throughput may be far below the MIMO capacity, since linear processing may not provide near-optimal performance. Non-linear (NL) processing could close the capacity gap and achieve a higher sum throughput. Nevertheless, it is widely known that NL processing achieves higher performance at the cost of a high computational complexity, which becomes impractical for a large number of streams. As an example, even to provide hard-output detection for 8 streams, a sphere decoder must meet a rate on the order of $10^3$ GFLOPS [7], which is beyond the capabilities of current processors. Fortunately, it has been recently shown that massively parallel processing is a promising method to alleviate the processing complexity and latency issues to help make NL processing practical.

In modern computer chips, the clock rate has practically reached a ceiling several years ago [8]. Nevertheless, it is considered that Moore’s Law is still in effect as of this writing. This is possible, among other things, because additional CPU cores are being added to computer chips [9]. This suggests that rather than making sequential algorithms faster, it may be more efficient to design parallel signal processing algorithms to take advantage of the trends in computer chip development.

In this direction, an MU-MIMO massively parallel processing framework has been recently proposed in the literature [10]–[13]. In the uplink, some examples include MultiSphere [10] and Flexcore [7]. These serve as a framework that enables massively parallel non-linear (MPNL) processing through the use of a metric-of-promise (MoP), which allows the algorithm to search the most likely paths to a solution in parallel. In the downlink, the ViPer technique [14] uses the MPNL paradigm to enable efficient non-linear precoding.

In this work, we advocate the use of MPNL processing as an enabler of user-dense 6G networks. We showcase, discuss and evaluate three scenarios where MPNL provides substantial advantages over linear processing.

- MU-MIMO detection and precoding in a spatially-correlated scenario (Section III-A)
- Overloaded uplink MU-MIMO scenario, where there are more spatial streams than BS antennas (Section III-B)
• Decentralized uplink MU-MIMO case, where signal detection is performed within independent groups of BS antennas (Section III-C)

In each of these scenarios, we present evaluations in the form of link-level computer simulations using a 3rd Generation Partnership Project (3GPP)-compliant clustered delay line (CDL) wireless channel [15]. Furthermore, to further validate our results and to show that the simulated gains can be reproduced in a real-world environment, we also perform over-the-air, proof-of-concept system-level evaluations using the Soft and Open Radio Design for Rapid Development, Profiling, Validation, and Testing (SWORD) testbed [16], which is a full-stack, 3GPP standards-based platform. SWORD has been proven to be a useful tool to validate complex signal processing algorithms without the need for time-consuming and expensive hardware optimizations [16], [17].

The results substantiate that MPNL processing is capable of providing high throughput and connectivity and is a feasible candidate for deployment in future wireless networks.

II. PRIMER ON PARALLEL AND MASSIVELY PARALLEL NON-LINEAR PROCESSING

A. Overview of Parallel and Non-Linear Processing

Efficient parallel algorithms are highly useful for signal processing. By sharing tasks among several processing elements (PEs), the processing latency can be reduced considerably compared to sequential procedures. Additionally, parallel processing makes efficient use of modern computer processors, as these are typically packed with many cores. “Emarrassingly parallel” algorithms, where the tasks can be neatly separated among parallel tasks with minimal dependencies, are highly preferred. Nevertheless, this is not always possible, and inevitably some parallel algorithms incur additional overhead compared to their non-parallel counterparts.

Another desired attribute of parallel algorithms is the flexibility to operate with any number of PEs, and with the performance improving with the number of PEs. However, efficiently and consistently increasing the provided performance by increasing the number of PEs can be very challenging in practice, since it would require the ability to focus the additional processing on tasks that are directly related to performance improvement.

MU-MIMO NL detection and precoding algorithms achieve higher performance than their linear counterparts when the propagation channel is ill-conditioned, at the cost of a substantially higher complexity [18]–[20]. The objective function(s) of NL detection and precoding is one (in the case of precoding or hard-decision detection) or several (for soft-decision detection) integer least-squares function(s) of the form

\[ \hat{x} = \arg \min_{x \in \mathcal{X}} \| y - Ax \|^2, \quad (1) \]

where \( y \) is the vector of observables, \( A \) is the interference matrix, which depends on the channel, and \( x \) are the unknown integer vectors with dimension \( K \), which are the number of spatial layers. It is assumed that each element of \( x \) is taken from the set \( \mathcal{X} \), which is the transmit constellation or a subset of it (for soft-decision detection). The optimization needs to be performed over all the possible \( x \), which account to \( |\mathcal{X}|^K \).

From (1), it can be deduced that a brute-force optimization exhibits a polynomial complexity that is polynomial with the modulation order and exponential with the number of spatial layers.

Through the use of the sphere decoder algorithms [18], [21], [22], the optimization procedure is converted into a tree search; this offers a reduced complexity, but it is still significantly high. However, the conventional sphere decoder algorithm is highly sequential. Suboptimal sphere decoding-based techniques with parallel processing have been proposed [23], but at the expense of a substantial loss of efficiency. Other approaches attempt to solve (1) in parallel using graphics processing units (GPUs) [24], [25]. However, the methods are based on heuristics, and a high degree of communication is needed between the parallel tasks as they traverse the branches of the tree search. As a result, they incur a high overhead and are not scalable to large numbers of streams. Additionally, since they are based on heuristics, they lack a theoretical context.

B. Massively Parallel Non-Linear Processing Framework

Compared to the non-linear parallel processing approaches described in the previous section, the MultiSphere MPNL framework [10]–[13] evades most of their limitations. In the MPNL framework, heuristics are replaced with a mathematical model that predicts the branches where the optimal solution will be found with a high probability. In this way, each PE is preassigned a task to search the candidate solutions that are more likely to be the ones that solve (1) and do not require communication with other PEs until its task is completed. In this way, redundant operations are reduced, which allows the framework to work with a very large number of PEs without reaching a performance ceiling. Furthermore, its complexity is more than an order of magnitude compared to other parallel approaches, and its latency is comparable to that of linear methods [7], [14].

As discussed, the MultiSphere MPNL framework first identifies the most promising solutions to Eq. (1) prior to detection or precoding as relative position vectors (RPVs) by employing a new MoP based on the channel information. Then, when data transmission takes place, these RPVs are demapped to actual symbol vectors by PEs. Due to this, the available processing power is focused on the most promising solutions to Eq. (1). Further details describing the approach are found in [10], [7]. The MPNL framework has been implemented in a field-programmable gate array (FPGA) [26] and Compute Unified Device Architecture (CUDA) in [7]. Real-time operation was verified for large order MIMO of up to 12 × 12. The MPNL framework is flexible and has been extended to G-MultiSphere [11], which applies to generalized systems, where the channel matrix is under-determined, since the BS has fewer antennas than the number of spatial streams it serves.
III. Massively Parallel Non-Linear Processing Scenarios for 6G

Current generation networks (5G) have not made substantial use of NL processing. Current 3GPP standards require changes to signaling procedures and upper layers to enable NL processing [27]. The lack of NL processing in 5G can be due to various reasons, some of which perhaps are that the current use cases or the density of existing networks do not yet justify the higher complexity associated with NL methods. However, this is expected to change for 6G networks since, as discussed, the vision for 6G networks includes new demanding use cases and substantially increased network density. In this work, we advocate the use of MPNL processing for the next-generation 6G networks. We argue that MPNL processing can benefit future 6G networks, and we present and evaluate three likely scenarios where this could happen. It is worth noting that the three scenarios presented here are not an exhaustive list, as MPNL may be beneficial in additional 6G wireless communication setups. For each of the three scenarios, we validate the gains of MPNL processing in the form of computer simulations and use the SWORD testbed [16] to conduct over-the-air system-level evaluations using a standards-compliant, full-stack environment. We present the results in Section V.

A. MPNL Processing for User-Dense Spatially-Correlated Environments

This might be the most evident scenario of the use of NL processing, as it is well-known that when the number of spatial streams is equal or nearly equal to the number of BS antennas, NL processing greatly outperforms linear methods. When the network needs to serve a large amount of user equipment devices (UEs) (where a high spatial correlation would be likely), using the spatial domain to serve them would significantly increase the multiplexing gain. In this case, the use of the massive MIMO paradigm with linear processing may not be practical (or efficiently) possible, due to the great many BS antennas that would be required. In this case, NL processing could offer high throughput and reduce the burden of the number of network-side transceivers. It should be noted that NL offers gains over linear processing even when the number of BS antennas greatly exceeds the number of UEs.

B. Overloaded MPNL Processing for Extremely Dense Machine Communications

To enable extreme connectivity, recently-proposed non-orthogonal multiple access (NOMA) schemes [28], [29] rely on specifically designed sparse signals, which enable receiver processing by the message-passing algorithm (MPA). However, these sparse signals underutilize the available resources, require changes to the standard and additional control overhead which prevents spontaneous and unsupervised transmissions. Here we consider the case where there is a high number of low-rate UEs, as in IoT sensors or MTC devices connected to the network, and the number of UEs is more than the number of BS antennas (an overloaded scenario). The use of user scheduling in this scenario may be inefficient, as it introduces a high signaling overhead and offers a low multiplexing gain. Therefore, we showcase the use of G-MultiSphere [11], where we show that a BS can serve twice or more UEs than what is possible through the use of linear precoding. G-MultiSphere generalized MPNL processing can be thought of as a way to enable massive connectivity by trading off transceiver hardware to computational resources.

C. Decentralized MPNL Processing

In this scenario, we investigate the case of a decentralized BS for uplink communications. As shown in Fig. 1, we consider a BS where the MIMO signal detection is performed independently over N detectors. The output log-likelihood ratios (LLRs) of the detectors are then simply added before being processed by the decoder. Decentralized processing [30] can be beneficial in at least two cases: Firstly, for a massive MIMO BS, the communication between the BS’s remote radio head (RRH) and baseband unit (BBU) can be prohibitively high. As an example, for a massive MIMO system with 40 MHz bandwidth and 128 base station antennas, with 10-bit analog-to-digital converters (ADCs), the combined bit rate can exceed 200 Gigabits-per-second (Gbps) [30]. Furthermore, it can decrease the burden of synchronizing a large number of transceiver chains. Processing the incoming data in smaller groups can reduce the complexity (and cost) of the BS. And secondly, decentralized MPNL processing would be favorable in distributed MIMO networks connected by fronthaul links or in cell-free MIMO [31]. A decentralized scenario allows each partition to process its received data independently of others. Additionally, only the post-detection LLRs need to be transmitted over the fronthaul links (a fully-cooperating network setup would need to transmit the full in-band and quadrature (IQ) samples). Furthermore, each partition would only need to know its local (rather than global) channel state information (CSI).

IV. METHODOLOGY OF EVALUATIONS AND MEASUREMENTS

The three MPNL processing scenarios were evaluated using link-level simulations and system-level over-the-air measurements using the SWORD testbed. The objective of these evaluations was to validate the gains of MPNL processing and identify potential issues or bottlenecks.
We used link-level simulations and system-level over-the-air measurements because these approaches can be complementary to each other. Link-level simulations have the advantage that they are flexible and can be performed for a large number of BS antennas and UEs. In contrast, over-the-air system-level measurements validate real-life gains of the algorithms under test, with actual communication procedures, such as synchronization and channel estimation.

A. Computer Simulation Methodology

The channel model used for the simulations was the 3GPP CDL-B, which models a non-line-of-sight scenario. The carrier frequency was 3.5 GHz, and the RMS of the channel delay spread was 300 nS. The UEs are assumed to be at the same distance from the BS and distributed randomly within a 60-degree angle. It was assumed that the BS had perfect CSI knowledge. Low-density parity-check (LDPC) was used for forward error correction (FEC) coding. The evaluations were conducted over a minimum of 100 frames. For the MPNL processing, 48 PEs were used, since we found that this value provided a good balance between performance and complexity.

B. Over-the-Air Measurement Methodology

In order to perform over-the-air system-level measurements, we used our recently-proposed SWORD research platform [16]. SWORD is a flexible, open for research, software-driven platform that facilitates rapid evaluation of advanced algorithms in a 3GPP standards–compliant environment. It features a pseudo real-time mode of operation based on the use of pause periods which facilitates validation of high-complexity algorithms without extensive hardware optimizations. The system setup used during measurements included 4 single-antenna UEs and up to 8 BS antennas, with the number of PEs used for MPNL processing set to 64.

The measurements were conducted in an indoor location with UEs placed close together and approximately 4 meters from the BS, as depicted in Fig. 2, which resulted in a signal-to-noise ratio of approximately 15 dB.

V. RESULTS OF COMPUTER SIMULATION EVALUATIONS AND OVER-THE-AIR MEASUREMENTS

Here we present the results and observations of the link-level simulations and over-the-air measurements for each of the three MPNL processing scenarios for 6G.

A. MPNL Processing for User-Dense Spatially-Correlated Environments

For these experiments, the objective was to compare the throughput between MPNL and linear processing. For the link-level simulations, a 64-antenna BS was used, uplink detection and downlink precoding throughput evaluations were conducted for different numbers of UEs in the range from 8 (lightly-loaded massive MIMO case) to 64 (fully-loaded case). MultiSphere is used as the MPNL approach. Fig. 3 shows the sum throughput achieved by MMSE and MPNL processing. The results show that the throughput of MMSE roughly reaches a ceiling at 32 UEs. In contrast, the throughput of MPNL detection keeps increasing up to the fully-loaded 64 × 64 MIMO case. This example also demonstrates that MPNL detection exhibits low complexity and can be used for a large number of streams, something that is difficult when using conventional sphere decoders.

Furthermore, system-level over-the-air measurements were conducted for uplink detection and downlink precoding using MPNL and ZF processing for a 4- and 8-antenna BS and 4 UEs. Fig. 4 shows the throughput gains of MPNL over ZF processing. The results show that for the 4 × 4 case, MPNL offers significant gains, but these are reduced for the 8 × 4
case. The results confirm that NL offers gains in fully-loaded MU-MIMO, while linear processing offers good performance in underloaded conditions. Limitations of the testbed on the number of BS antennas and UEs prevented us from measuring higher-order MIMO systems. Nevertheless, the results are significant because they validate that MPNL detection and precoding work in an actual 3GPP standards-based system.

B. Overloaded MPNL Processing for Extremely Dense Machine Communications

Here we examine the case where the density of the network is such that there are more spatial streams than the number of BS antennas. This scenario models the case where a large number of low-rate machine-type devices are connected to the BS. The simulation results are presented in Fig. 5, which shows the uplink sum throughput of an 8-antenna BS when the number of UEs varies between 2 and 18. G-MultiSphere is used as the MPNL approach. The plot shows that the throughput of MPNL detection increases up to 14 UEs (an overloaded factor of 1.75), whereas for MMSE detection, the throughput increases only up to 6 UEs (an overloaded factor of 0.75), with zero throughput when 16 UEs are connected.

The results validate the substantial gains of MPNL over linear processing in overloaded MIMO.

Over-the-air measurements were conducted for an overloaded case for a BS with a single antenna serving \{2, 3, 4\} UEs, and for a 2-antenna BS serving \{3, 4\} UEs, with MPNL detection. The measurements were repeated for MCS 4, 7, and 10, which correspond to QPSK and code rates 0.25, 0.4, and 0.6, respectively. The results show the block pass rate (BLPR) for the single-antenna BS in Fig. 6 (left) and the two-antenna BS in Fig. 6 (right). From the figures, it can be observed that lower code rates have higher BLPR as the overload factor is increased. For the single-antenna BS, MPNL detection can achieve 85% BLPR when using MCS 4 with a \(3 \times\) overload factor. Furthermore, for the two-antenna BS, a 99% BLPR is achieved using MCS 10 with a \(2 \times\) overload factor.

The simulation and over-the-air measurement results corroborate that MPNL can deal with high connectivity conditions, which will likely be present in 6G mobile networks.

C. Decentralized MPNL Processing

Computer simulation evaluations were conducted for a decentralized 24-antenna BS with 3 independent detectors, each with 8 antennas. The evaluations were done for underloaded (4 UEs), fully-loaded (8 UEs), and overloaded (16 and 20 UEs) cases with QPSK modulation and 0.75 code rate. G-
MultiSphere is used as the MPNL approach. The results are shown in Fig. 7 and show that MPNL detection significantly outperforms MMSE detection in the fully-loaded and over-loaded cases; at a $2 \times$ overloading factor, MPNL reaches the highest throughput, while for MMSE it is zero.

Furthermore, system-level over-the-air measurements were conducted for a decentralized 8-antenna BS, with 2 independent detectors, with 4 antennas each. Fig. 8 shows the sum spectral efficiency achieved for MMSE and MPNL detection for the centralized and decentralized cases. The results confirm that MMSE loses performance in the decentralized case, but MPNL detection can achieve nearly the same throughput in the decentralized and centralized cases.

The results corroborate that for decentralized BSs, MPNL processing achieves higher performance and enables over-loaded conditions, which are not feasible by the use of linear processing.

VI. CONCLUSION AND FUTURE WORK

It is expected that next-generation wireless networks will feature a highly dense network. In such cases, massively parallel non-linear processing can offer a substantial improvement over linear processing in MU-MIMO systems, while alleviating the complexity issues of conventional non-linear processing. In this work, we discuss three MU-MIMO scenarios that could be present in 6G networks. Our results show that MPNL can significantly outperform linear processing and validate that the MPNL paradigm can enable massive connectivity and significantly increase the sum throughput of 6G networks. For future work, the MPNL concept can be extended to architectures that also include the channel decoder, which can further improve the network performance. Furthermore, current hardware platforms, such as CUDA, are not optimized for massively parallel wireless communications. Work in this area could further improve the performance and applicability of MPNL processing.

REFERENCES


