Low-Precision Multi-GPU Detection Approach for Massive MIMO Technology

A. Dabah [1], H. Ltaief [1], Z. Rezki [2], M.-S. Alouini [1], and D. Keyes [1]

[1]Computer, Electrical and Mathematical Science and Engineering,
King Abdullah University of Science and Technology
{Adel.Dabah.1, Hatem.Ltaief, Slim.Alouini, David.Keyes}@kaust.edu.sa

[2] University of California Santa Cruz, USA,
zrezki@ucsc.edu

***Abstract*—Massive Multiple-Input-Multiple-Output is a crucial technology for Next-Generation networks (Next-G). It uses hundreds of antennas at transceivers to exchange data. However, its accurate signal detection relies on solving an NP-hard optimization problem in real-time latency. In this poster, we propose a new GPU-based detection algorithm that demonstrates the positive impact of low-precision arithmetic with multiple GPUs to achieve next-G latency/scalability/accuracy requirements. Our approach iteratively extends a solution with several symbols representing the best combination out of the aggregated levels. The computation at each iteration is formulated as a matrix multiplication operation to leverage GPU architectures. The obtained results using A100 GPU show a 1.7× improvement by exploiting half-precision arithmetic without loss in accuracy. Furthermore, our low-precision multi-GPU version with four A100 GPUs is 4× faster than the single-precision single GPU version and 40× faster than a similar parallel CPU implementation executed on a two-socket 28-core IceLake CPU with 56 threads.**

***Index Terms*—Half-precision arithmetic, Multi-GPU, Multi-Level approach, Massive MIMO system.**

I. Introduction

Recent Graphics Processing Unit (GPU) accelerators are equipped with tensor cores that can deliver unprecedented compute power by reducing the floating-point precision. Their impact on the wireless communication field in general and Massive Multiple-Input Multiple-Output (M-MIMO), in particular, has not been studied yet. M-MIMO technology is and will continue to be an essential part of Next-Generation (Next-G) mobile communication networks. It uses hundreds of antennas to send and receive data, which allows to support an enormous amount of data exchanged in the fifth-generation mobile communication [2], [4]. However, the more antennas we use, the more difficult it becomes to achieve a low error rate under a real-time requirement. In the literature, we find two main categories of detection algorithms: linear and non-linear detection algorithms. On the one hand, linear detection algorithms like Zero-Forcing (ZF) and Minimum-Mean-Square Error (MMSE) [5] operate under real-time requirements but fail to achieve a good error performance for a dense constellation. On the other hand, non-linear detection algorithms such as Sphere Decoder (SD) and maximum likelihood (ML) [1], [3] give an excellent estimation of the transmitted data but fail to operate under a practical latency requirement due to the exponential complexity. In this poster, we demonstrate the positive impact of using low-precision computation along with multiple GPUs to improve the latency/scalability/accuracy of M-MIMO detection to meet Next-G requirements. The proposed approach iteratively extends a single solution (empty initially) with several symbols representing the best combination of the combined tree symbols. The computation at each iteration is formulated as matrix algebra operations. The more levels we put together, the larger the matrices we manipulate, and the more accurate our approach becomes. However, the complexity also increases considerably. To satisfy a practical real-time requirement, we rely on half-precision floating-point arithmetic and multi-GPUs. Our proposed three-step technique leverages recent GPU architectures to perform in successive phases: (1) matrix multiplication, (2) square-norm calculation, and (3) sorting based on a reduction process.

The obtained results using A100 GPU show a 1.7× improvement factor by going from single (32-bit floating point operations (FP32)) to half-precision (16-bit floating point operations (FP16)) mode without any loss in accuracy (error rate performance). The reported speedup comes mostly from the overhead mitigation of data movement thanks to FP16 representation. Indeed, our computations operate on short and wide matrices, i.e., with the number of columns (all possible paths) much higher than the number of rows (the level count). Such matrix shape does not engender enough data reuse for such operation to be in the compute-bound regime of execution, as usually noticed for traditional square matrix-matrix multiplication.

Moreover, scaling to multiple GPUs further reduces the time-to-solution of the main kernel, i.e., the matrix-matrix multiplication. This latter represents more than 80% of the elapsed time for dense constellations. The idea is to simultaneously execute matrix-matrix multiplications from subsequent

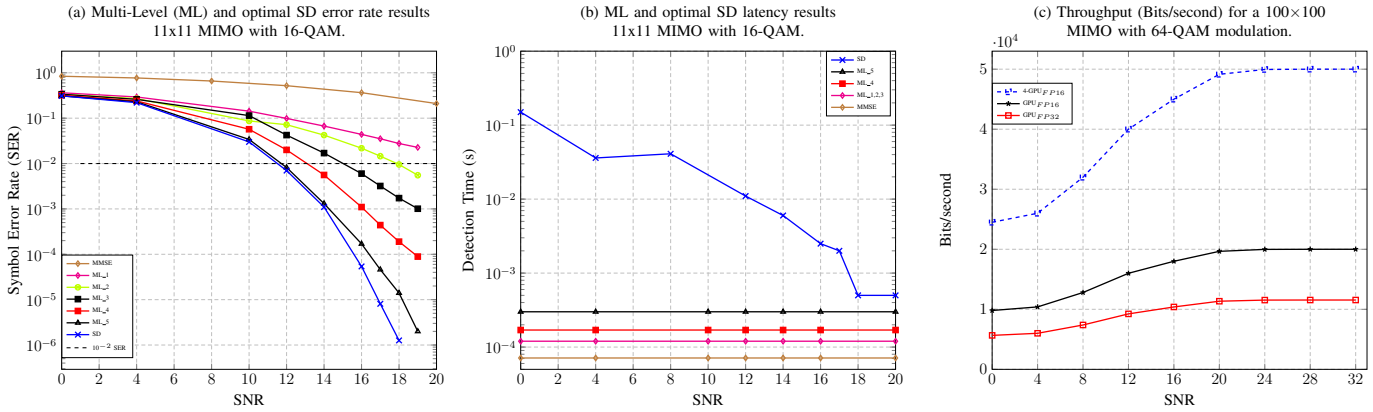

Fig. 1: Accuracy, latency and throughput of our GPU ML approach.

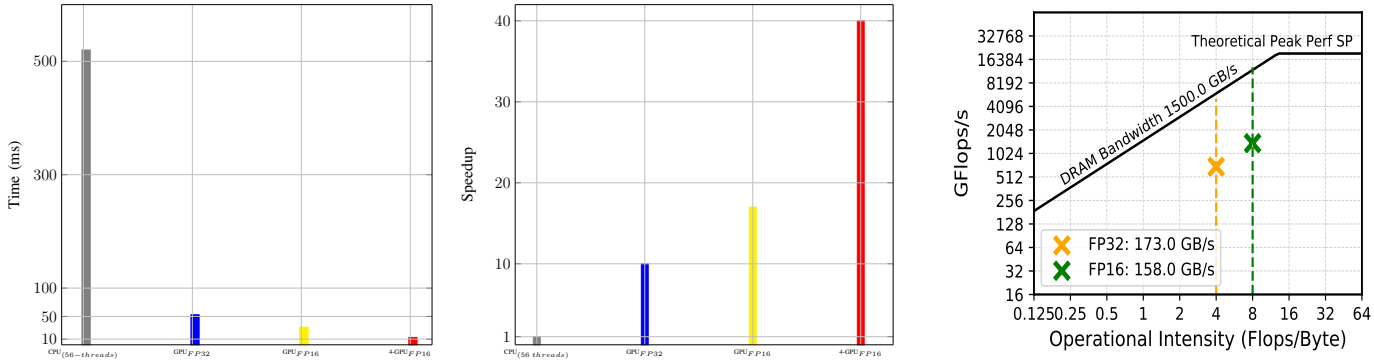

Fig. 2: Complexity and speedup for a 100×100 MIMO with 64-QAM modulation.

iterations using multiple GPUs. However, the remaining 20% of the code must be executed sequentially, which may impede strong scaling performance. The obtained results meet the Next-G requirements in terms of real-time processing. Indeed, our multi-GPU version with four A100 GPUs is up to 2.3× faster than the single GPU version with half-precision mode and achieves up to 4× improvement factor compared to the single-precision single GPU approach. Considering Amdahl's law, these reported results are near ideal speedup. In addition, our multi-GPU version is 40× faster than a similar parallel CPU implementation executed on a two-socket 28-core IceLake CPU with a total of 56 threads. As a result, we eventually achieve decent time complexity, performance scalability, and error rate. We believe this work will promote our low-precision multi-GPU approach within the wireless communication community to tackle challenging M-MIMO configurations with the advent of Internet of Things (IoTs).

II. Problem Formulation

In this poster, we consider a $M \times N$ MIMO system consisting of M transmit antennas and N receive antennas. It can be described by the following input-output relation :

$$\boldsymbol{y} = \boldsymbol{H}\boldsymbol{s} + \boldsymbol{n},$$

where the vector $\boldsymbol{y} = [y_0, ..., y_{N-1}]^T$ represents the received signal. $\boldsymbol{H}$ is an $N \times M$ channel matrix, where each element h_{ij} is a complex Gaussian random variable, with mean 0 and variance 1, modeling the fading gain between the j-th transmitter and i-th receiver. The vector $\boldsymbol{s} = [s_0, ..., s_{M-1}]$ represents the transmitted vector, where $\boldsymbol{s_i}$ belongs to a finite alphabet set denoted by Ω. Finally, $\boldsymbol{n} = [n_0, ..., n_{N-1}]^T$ represents the additive white Gaussian noise. For convenience, let us consider $\mathcal{S}$ as the set of all possible combinations of the transmitted vector $\boldsymbol{s}$. The possible number of combinations corresponds to the complexity of the MIMO system, and it is calculated as follows: $|\mathcal{S}|=|\Omega|^M$.

References

[1] Agrell, E., Eriksson, T., Vardy, A., and Zeger, K. Closest point search in lattices. *IEEE Transactions on Information Theory 48*, 8 (2002), 2201–2214.

[2] Foschini, G. J. Layered space-time architecture for wireless communication in a fading environment when using multi-element antennas. *Bell Labs Technical Journal 1*, 2 (1996), 41–59.

[3] Hassibi, B., and Vikalo, H. On the sphere-decoding algorithm part I. Expected complexity. *IEEE Transactions on Signal Processing 53*, 8 (2005), 2806–2818.

[4] Paulraj, A. J., and Kailath, T. Increasing capacity in wireless broadcast systems using distributed transmission/directional reception (DTDR), Sept. 6 1994. US Patent 5,345,599.

[5] Xie, Z., Short, R. T., and Rushforth, C. K. A family of suboptimum detectors for coherent multiuser communications. *IEEE Journal on Selected Areas in Communications 8*, 4 (1990), 683–690.