King Abdullah University of Science and Technology
جامعة الملك عبدالله للعلوم والتقنية

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

Adel Dabah[1] Hatem Ltaief [1] Zouheir Rezki [2] Mohamed-Slim Alouini [1] David Keyes[1]

[1]Computer, Electrical and Mathematical Science and Engineering, King Abdullah University of Science and Technology, KSA
[2]Department of Electrical and Computer Engineering, University of California Santa Cruz, USA.

King Abdullah University of Science and Technology
جامعة الملك عبدالله للعلوم والتقنية

Introduction

- Massive Multiple-Input Multiple-Output (M-MIMO) technology uses hundreds of antennas at transceivers to exchange data.
- It represents one of the key enabling technologies for next-generation wireless communication networks.
- Signal detection in M-MIMO represents the most critical task since the network's performance depends on it in terms of error rate and latency.

$$y = Hs + n.$$
$$\hat{s}_{ML} = arg\min_{s \in S} ||y - Hs||^2.$$

We assume baseband $M \times N$ MIMO where signal s is transmitted by M transmit antennas over a channel matrix H subject to a noise vector n. Vector y, represents a collection of N receiver antennas observations.

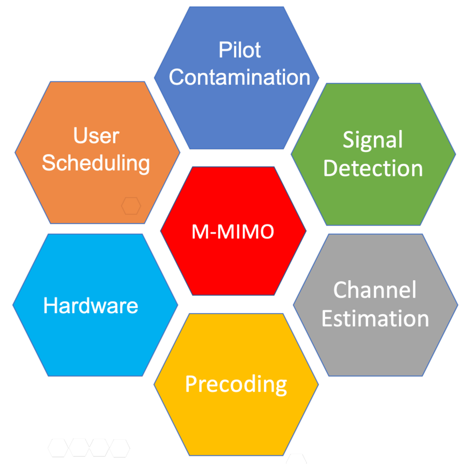

Figure 1. Massive MIMO challenges.

- The goal is to find the composition of symbols $\hat{s}_{ML}$ that minimizes the distance from the received observations. Each position in $\hat{s}_i$ belongs to a finite alphabet set Ω.
- This is an NP-hard problem where the number of solutions $|S|$ increases exponentially with the number of antennas (Ω^M).

Why Do We Need New Signal Detection Algorithms?

- **Linear-detection algorithms**, such as Zero Forcing (ZF) and Minimum Mean Square Error (MMSE), have acceptable latency but a poor error rate performance, especially for dense constellations.
- **Non-linear detection algorithms**, such as Sphere Decoder (SD), have excellent error rate performance but are challenging to use for M-MIMO in practice due to their exponential complexity.
- **Approximate non-linear detection algorithms**, such as K-best, constitute a trade-off between complexity and performance. However, they are not scalable and sensitive to dense constellations.

New GPU-Based Non-Linear Signal Detection Algorithms: Take-Home Messages

To answer the challenges of signal detection in M-MIMO, we developed a new algorithm, named *GPU Multi-Level (ML)*, to benefit from the high throughput of emerging massively parallel architectures. Our goals are:

- **Low latency** by exploiting the high-density computing power of Graphic Processing Unit (GPU) architectures.
- **Near-optimal error rate** by targeting ML solution.
- **High data rate** by relaying on dense constellations and a massive number of antennas.
- **Reduction in energy consumption** by operating in a practical SNR regime and relying on energy-efficient hardware.

Our approach reports good error rate performance for up to 100 antennas under real-time requirements.

Proposed GPU Multi-Level Approach (ML)

Our proposed approach operates on the search tree that models all possible combinations of the transmitted signal.

- **Combines coefficients from multiple levels to target ML solution.**
- **Casts this process into matrix algebra operations $A * B + C$.**
- **Relies on GPU hardware accelerators to keep practical time complexity.**

The algorithm performs two main steps:

- A matrix-matrix multiplication with a short and wide matrix **B (8 $\times$ 16M)**.
- Norm calculation and sorting using a reduction process.

Our approach avoids thread divergence and enables data reuse to efficiently exploit GPU capability and operate within real-time requirements.

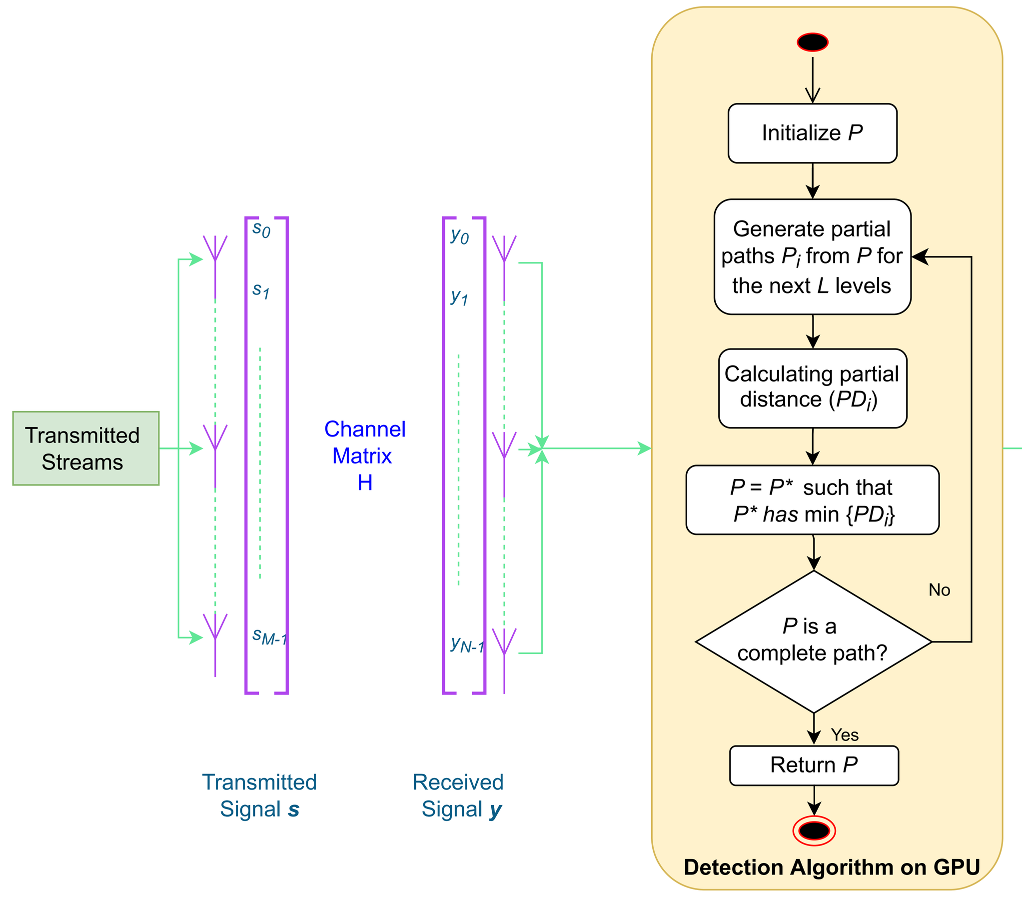

Figure 2. Proposed signal detection approach for M-MIMO.

Half-Precision and Multi-GPU Versions

Exploiting 16-bit Floating-Point (FP16) operation:

- FP16 mode is mainly driven by artificial intelligence.
- We exploit it in our case to reduce the latency requirement of the MIMO detection process.
- We rely on with FP16 representation of *A*, *B*, and *C* matrices.
- Using A100 GPU 40GB with FP16 precision, we achieve **1.7** $\times$ latency improvement (Speedup) compared to FP32 precision, without altering the detection accuracy.
- The shape of matrix ***B*** significantly impacts the obtained performance.

Exploiting multiple GPUs:

- Matrix multiplications represent 80% of the global execution time.
- Matrix multiplications on all iterations are independent and can be executed simultaneously.
- This version performs subsequent matrix multiplications in parallel using multiple GPUs.
- We achieve **2.3** $\times$ improvement factor with four A100 GPUs 40GB (FP16) compared to a single GPU with FP16 representation.

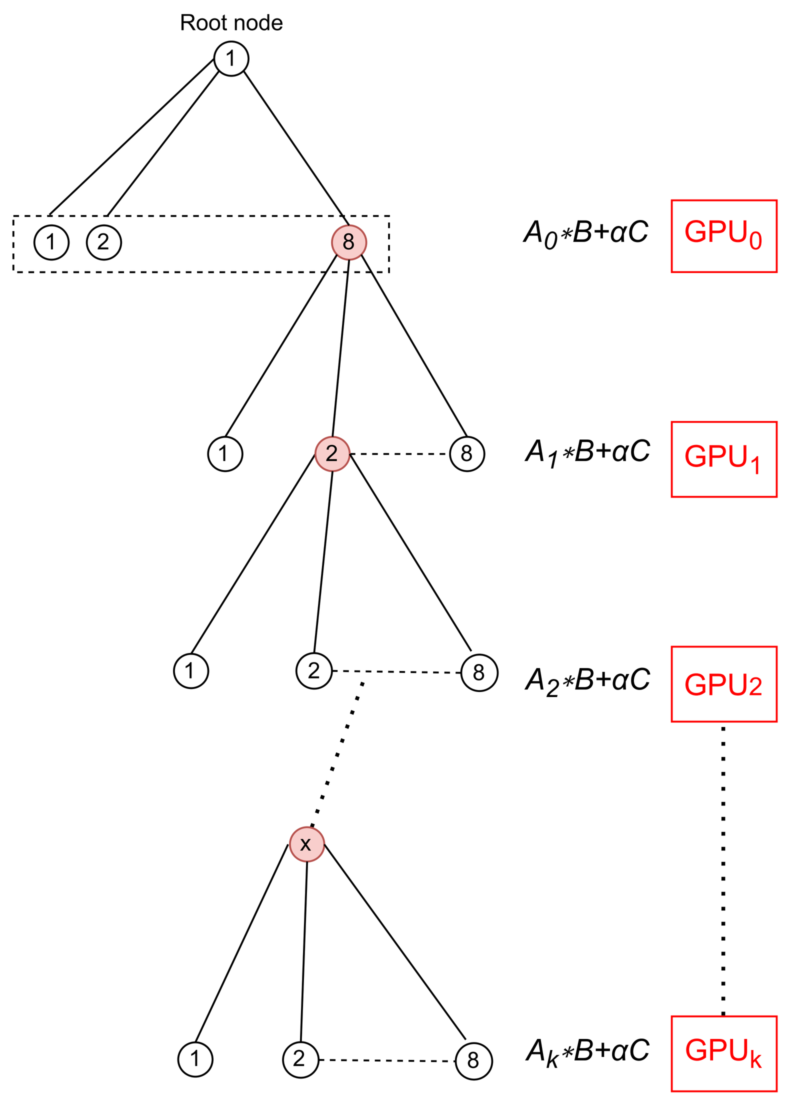

Figure 3. Multi-GPU version in which all GEMM operations during the detection process can be overlapped and performed in parallel using multiple GPUs.

Performance and Complexity Results

- Achieving near optimal Sphere Decoder (SD) results with low fixed complexity.
- Achieving 4$\times$ throughput improvement compared to single A100 GPU FP32 version for 100 $\times$ 100 MIMO with 64-QAM modulation.

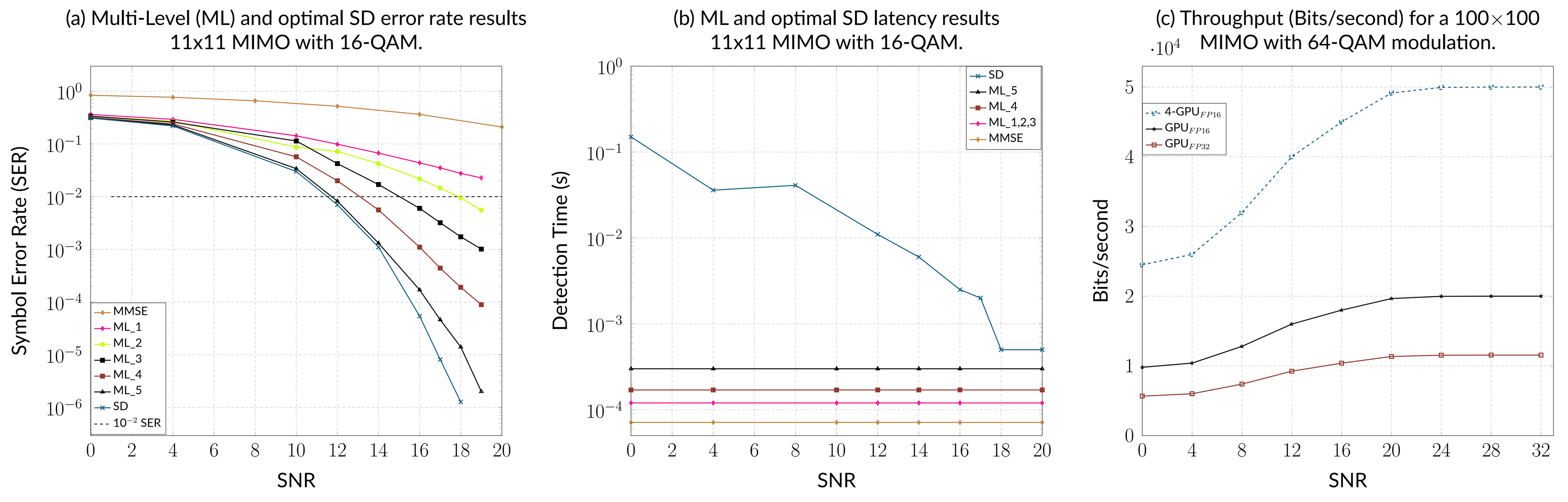

Figure 4. Accuracy, latency and throughput of our GPU ML approach.

- Achieving up to 40$\times$ compared to a similar parallel implementation on a two-socket 28-core Intel IceLake (56 threads total).

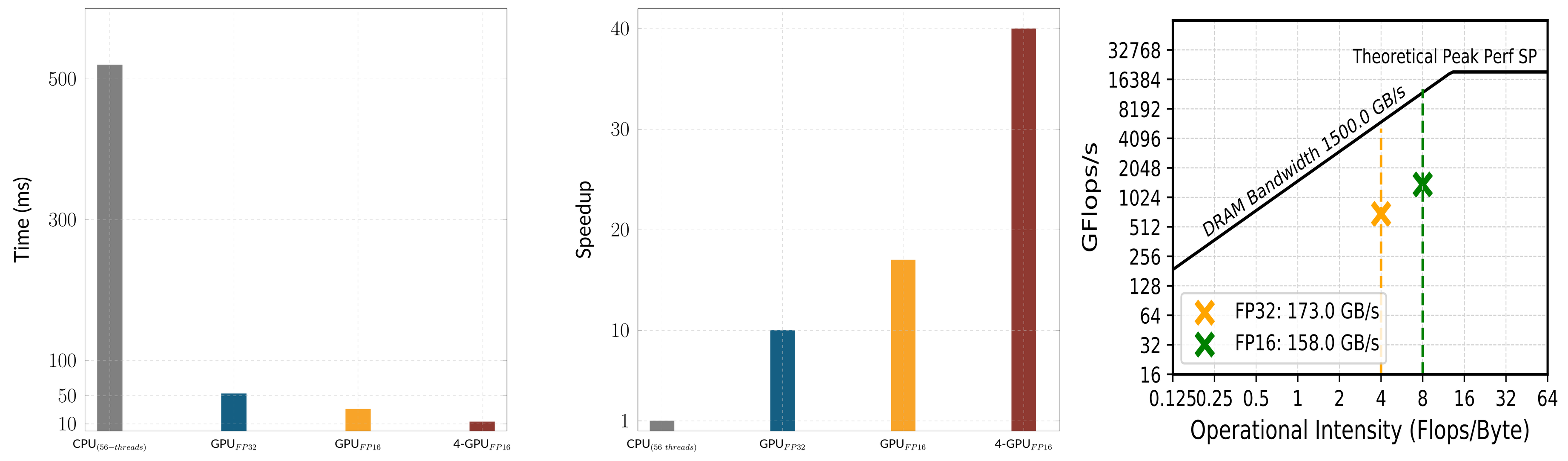

Figure 5. Complexity results using A100 GPUs for 100×100 MIMO system with 64-QAM modulation and 4 levels (ML_4).

Conclusion and Future Directions

- Designing new algorithms to exploit new hardware features is critical to meet the requirements of next-generation wireless communication networks.
- Exploiting half-precision in the MIMO detection process improves the latency without compromising accuracy thanks to our ML technique which provides more resilience for low precision arithmetic.

What is next?

- Increase the number of levels to reshape the matrix sizes and better exploit tensor cores capabilities.
- Reduce further the precision (i.e., FP8) without altering the accuracy of the detection process.

References

[1] Mohamed-Amine Arfaoui, Hatem Ltaief, Zouheir Rezki, Mohamed-Slim Alouini, and David Keyes. Efficient sphere detector algorithm for Massive MIMO using GPU hardware accelerator. *Procedia Computer Science*, 80:2169 – 2180, 2016.

[2] Adel Dabah, Hatem Ltaief, Zouheir Rezki, M-S Alouini, and David Keyes. Massive Multiple-Input Multiple-Output System and Method, December 14 2021. US Patent 11,201,645.

 adel.dabah.1@kaust.edu.sa