Scalable GPU-based Decoding Approach for Massive MIMO Technology

Adel Dabah

adel.dabah.1@kaust.edu.sa

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

Computer, Electrical and Mathematical Science and Engineering,
Extreme Computing Research Center (ECRC) && Communication Theory Lab (CTL)
King Abdullah University of Science and Technology

June 14, 2022

Complexity of Next-G mobile networks

By 2027,
1) 40 billion connected devices from different technologies from smart cities to Self-driving cars and UAVs.

2) 4.4X increase in data traffic, and 54% of it is in 5G.

3)Video traffic is estimated to be 79% of data.

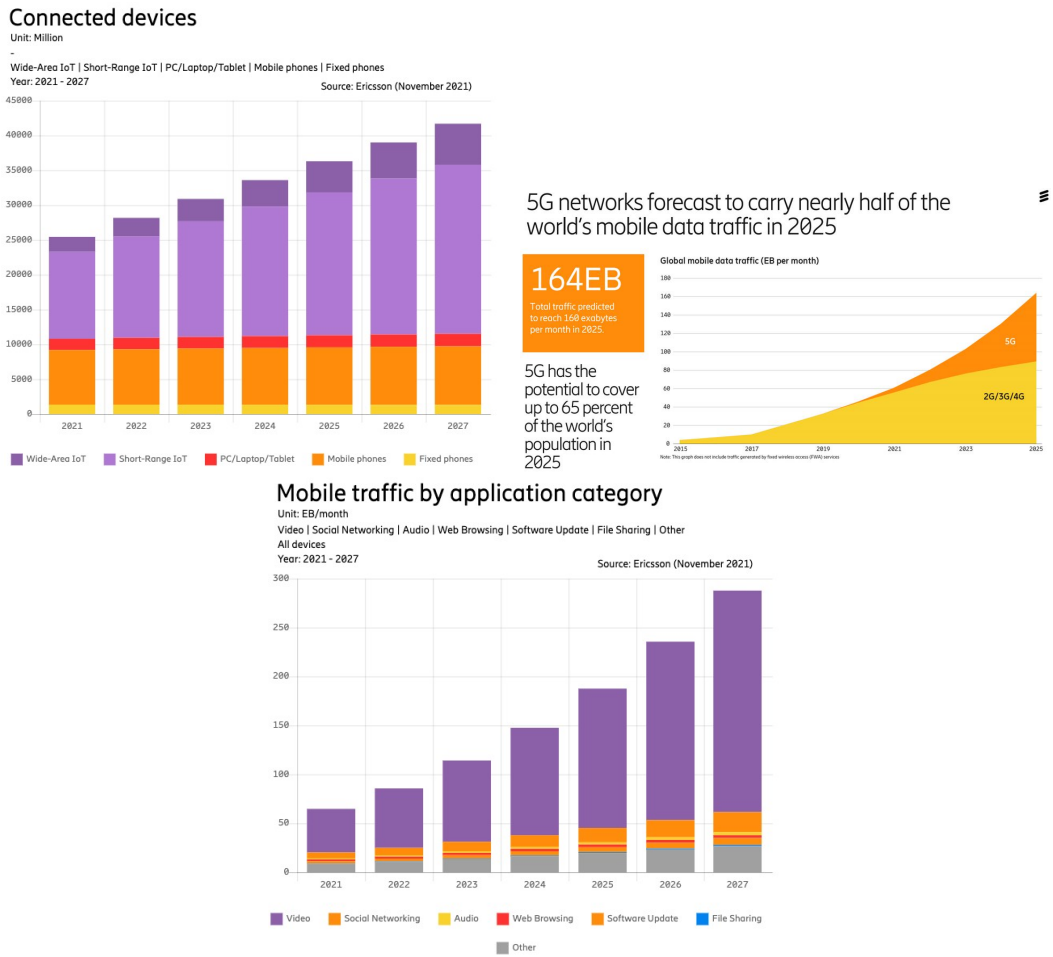
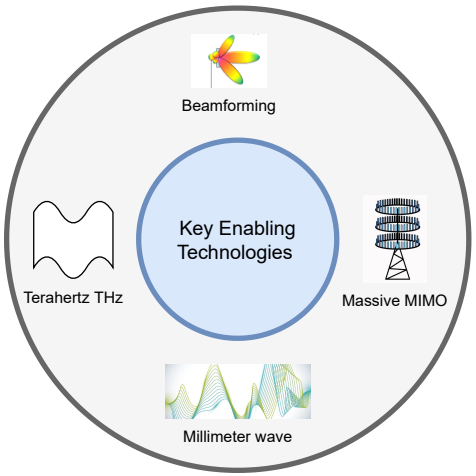


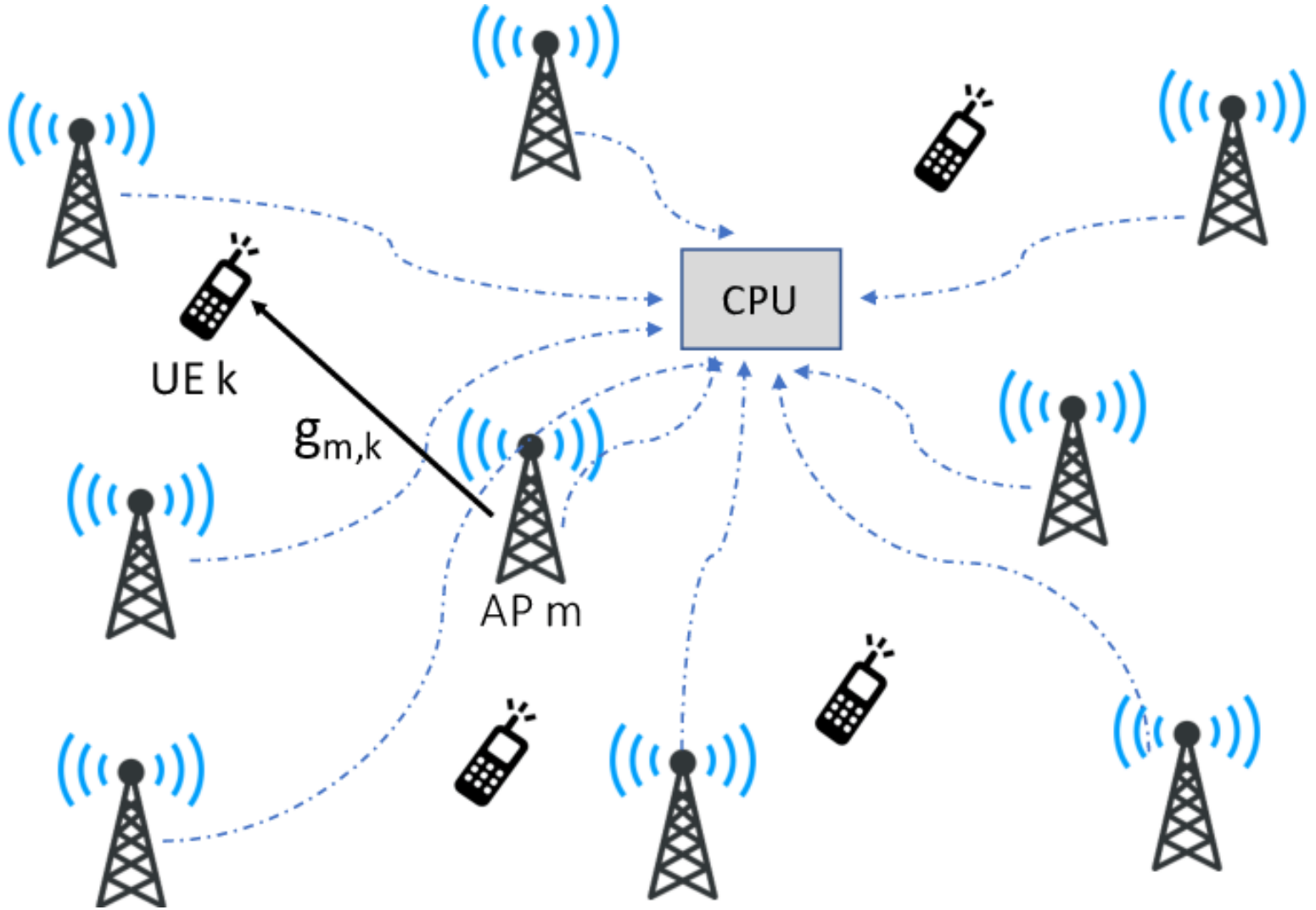
Figure 2: Ericsson data-traffic forecast.

Massive MIMO as a solution

- Massive Multiple-Input Multiple-Output (M-MIMO) is a generalization of single-input single-output technology, where we use hundreds of antennas at transverses instead of one.
- It aims to amplify all benefits of classical MIMO in terms of data rate, diversity gain, spectral efficiency, and network reliability.
- M-MIMO is one of the key enabling technology for next-generation wireless communication networks.
- It is motivated by the advent of graphic nano-antennas that allow the integration of hundreds of antennas in various terminals.



Cell-Free Massive MIMO [1]



Massive MIMO Key problems

- Main challenge about M-MIMO is to provide scalable/accurate physical layers algorithms.
- Signal detection represents the most critical task since the network's performance depends on it.

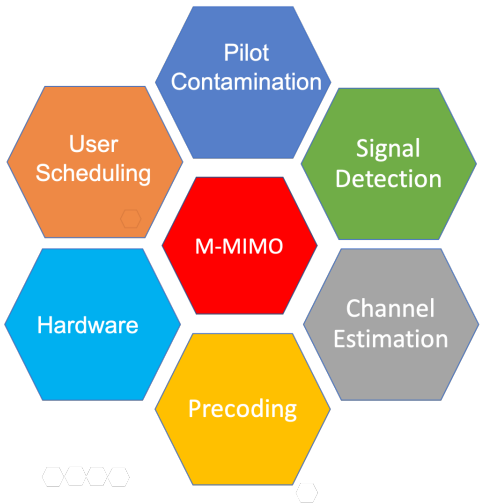


Figure 3: Massive-MIMO physical layers.

Poor performance of linear-decoders

Why Do We Need New Signal Decoding Algorithms for M-MIMO?

- Zero Forcing (ZF) and Minimum Mean Square Error (MMSE) have low latency, but they have a poor error rate performance, especially for a **large number of users** and **dense constellations**. Thus, inducing a throughput loss and low network reliability.

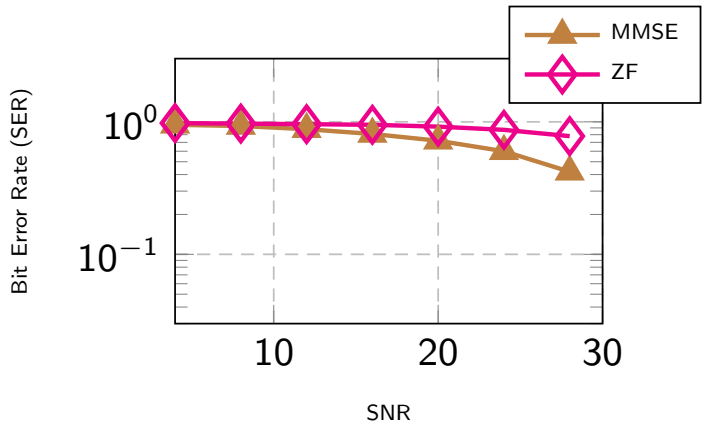


Figure 4: Error rate of MMSE and ZF for a 100 $\times$ 100 MIMO system with 64-QAM modulation.

- **Scalability issue** due to matrix-inversion operation needed by these algorithms.

M-MIMO optimization problem

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

$$\hat{\boldsymbol{s}}_{\boldsymbol{ML}} = \boldsymbol{arg}\min_{\boldsymbol{s}\in\boldsymbol{S}} ||\boldsymbol{y} - \boldsymbol{H}\boldsymbol{s}||^2. \tag{2}$$

$$\begin{aligned} ||\boldsymbol{y} - \boldsymbol{H}\boldsymbol{s}||^2 &= ||\boldsymbol{y} - \boldsymbol{Q}\boldsymbol{R}\boldsymbol{s}||^2 \\ &= ||\bar{\boldsymbol{y}} - \boldsymbol{R}\boldsymbol{s}||^2, \text{ where } \bar{\boldsymbol{y}} = \boldsymbol{Q}^{\boldsymbol{H}}\boldsymbol{y}, \end{aligned}$$

where $\boldsymbol{R} \in \boldsymbol{C}^{\boldsymbol{N}\times\boldsymbol{M}}$ is an upper triangular matrix and $\boldsymbol{Q} \in \boldsymbol{C}^{\boldsymbol{N}\times\boldsymbol{N}}$ is an orthogonal matrix.

$$\min \sum_{k=1}^{M} g_k(s_{M-1}, ..., s_{M-k}), \quad \textit{where} \tag{3}$$

$$g_k(s_{M-1}, ..., s_{M-k}) = ||\bar{y}_{M-k} - \sum_{i=M-k}^{M-1} r_{(M-k),i} s_i||^2. \tag{4}$$

Complexity of non-linear approaches

M-MIMO discrete optimization problem with Ω^M possible solutions

- **Optimal algorithms**, such as Maximum Likelihood (ML) and Sphere Decoder (SD), have excellent error rate performance but are challenging to use for M-MIMO in practice due to their exponential complexity.
- **Approximate algorithms**, such as K-best, constitute a trade-off between complexity and performance. However, they are sensitive to dense constellations and can not be used beyond a two-digit number of antennas. Thus, they are far from M-MIMO requirements.

Goals

To answer the challenges of signal decoding in M-MIMO, we develop new algorithms to match the high throughput of emerging massively parallel architectures. Our goals:

- **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 constellation and massive number of antennas.
- **Reduction in energy consumption** by operating in a practical SNR regime and relying on energy-efficient hardware.

Our proposed approach reports good error rate performance for 400×400 antennas under real-time requirements and practicable SNR.

HPC-based existing works

- GPU-based approaches perform a partial or complete tree exploration on GPU in a multi-thread way.
  - induces a high thread-divergence and low scalability.
  - overhead of managing a tree i.e. large number of data-structures.
  - not usable for M-MIMO systems.
- CPU/FPGA Flexecore, multi-sphere
  - multiple SD instances running in parallel.
  - explores many paths to guarantee decent error rate performance.
  - Relatively better success as compared to GPU-based approaches.

All existing approaches explore a large number of paths, leading to memory-bound and instruction-bound issues. This induces a high latency making these non-linear detection approaches non-suitable for massive MIMO even when using massively parallel architectures.

Proposed Multi-Level Approach (GML)

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

- Combines coefficient from multiple levels to target ML solution.
- Casts this process into matrix algebra operations.
- Relies on GPU hardware accelerators to keep practical time complexity.

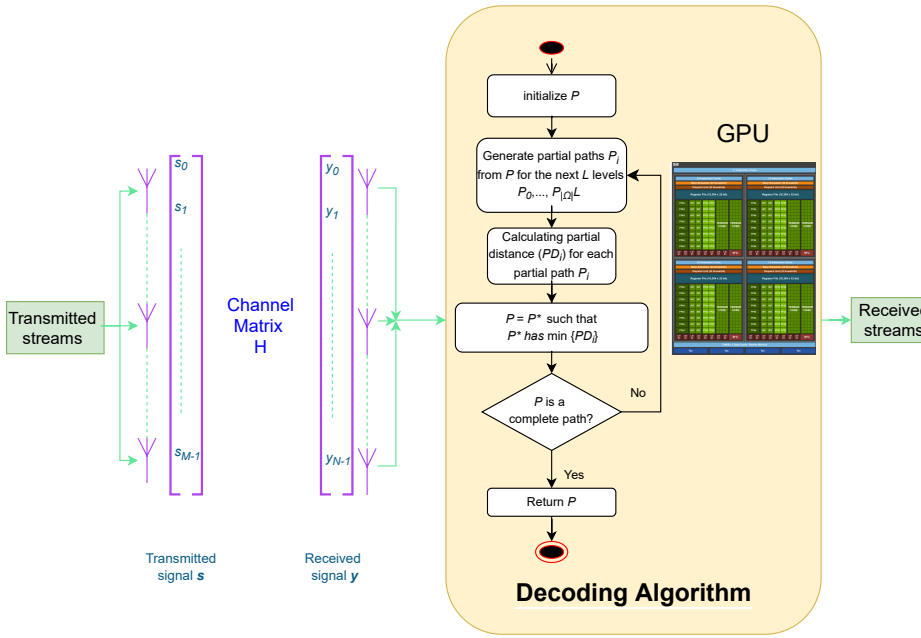
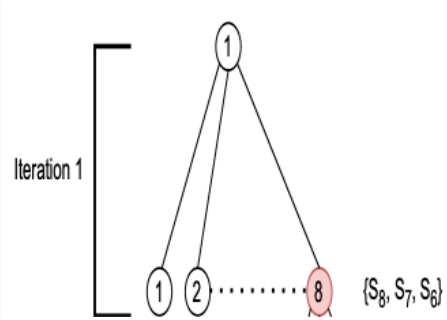
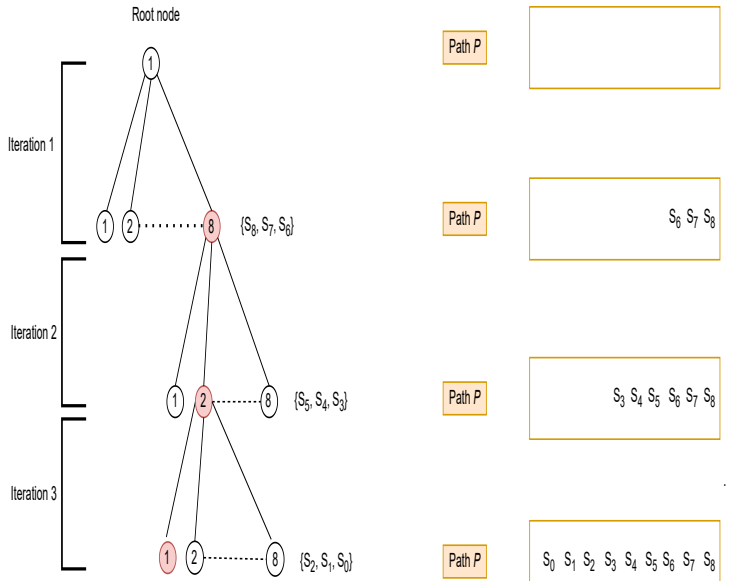


Figure 5: Proposed signal detection approach for M-MIMO.

Grouping the detection of L symbols



- Matrix-matrix multiplication $\boldsymbol{R'} \times \boldsymbol{B}$.
- High accuracy by using coefficients from different levels.
- Avoid error propagation.



Proposed GML Approach

The evaluation is incremental

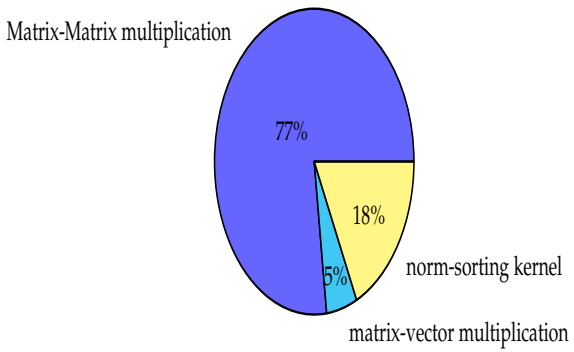
$$E(P_i) = \sum_{k=1}^{L_i} g_k(s_{M-1}, \ldots, s_{M-k})$$

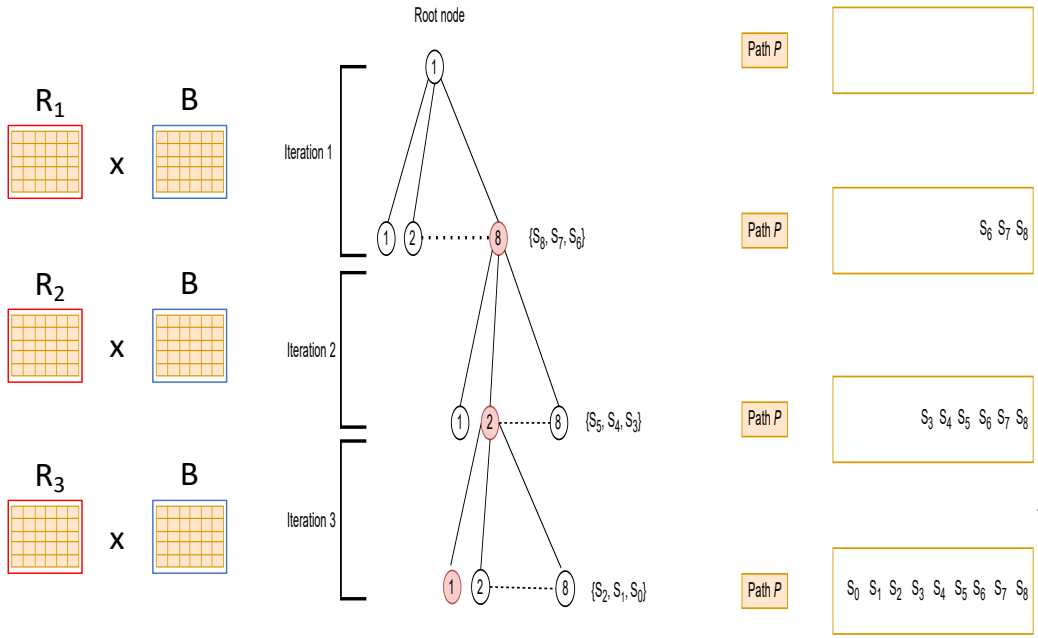
$$=$$

$$\overbrace{\underbrace{\sum_{k=1}^{L} g_k(s_{M-1}, \ldots, s_{M-k})}_{\text{E(P)}} + \underbrace{\sum_{k=L+1}^{L_i} g_k(s_{M-1}, \ldots, s_{M-k})}_{\text{non-computed part}}.}^{} \tag{5}$$

Multi-level technique

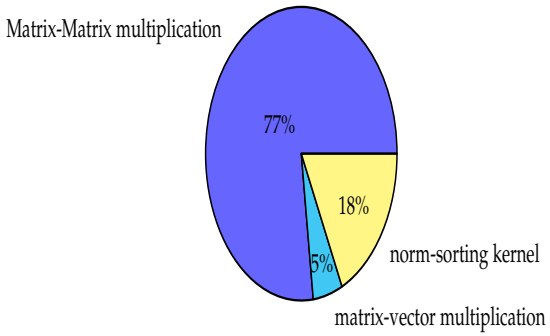
- Two main steps at each iteration
  - Matrix-matrix multiplication
  - Sorting phase using a reduction process

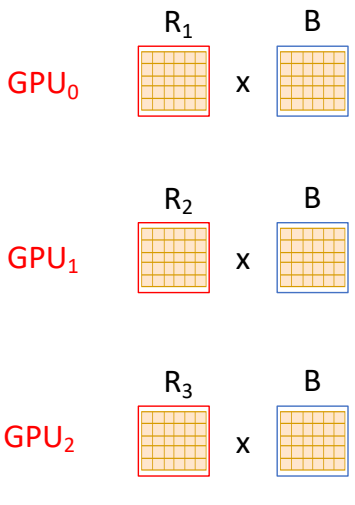


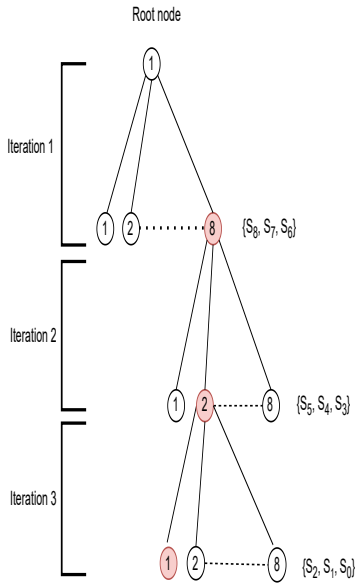


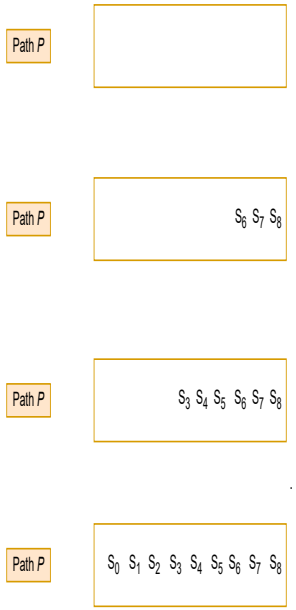
Parallel Multi-level technique

- Multi-GPU version
- Batched version









Performance and Complexity Results

- Achieving near optimal sphere decoder results with low fixed complexity.

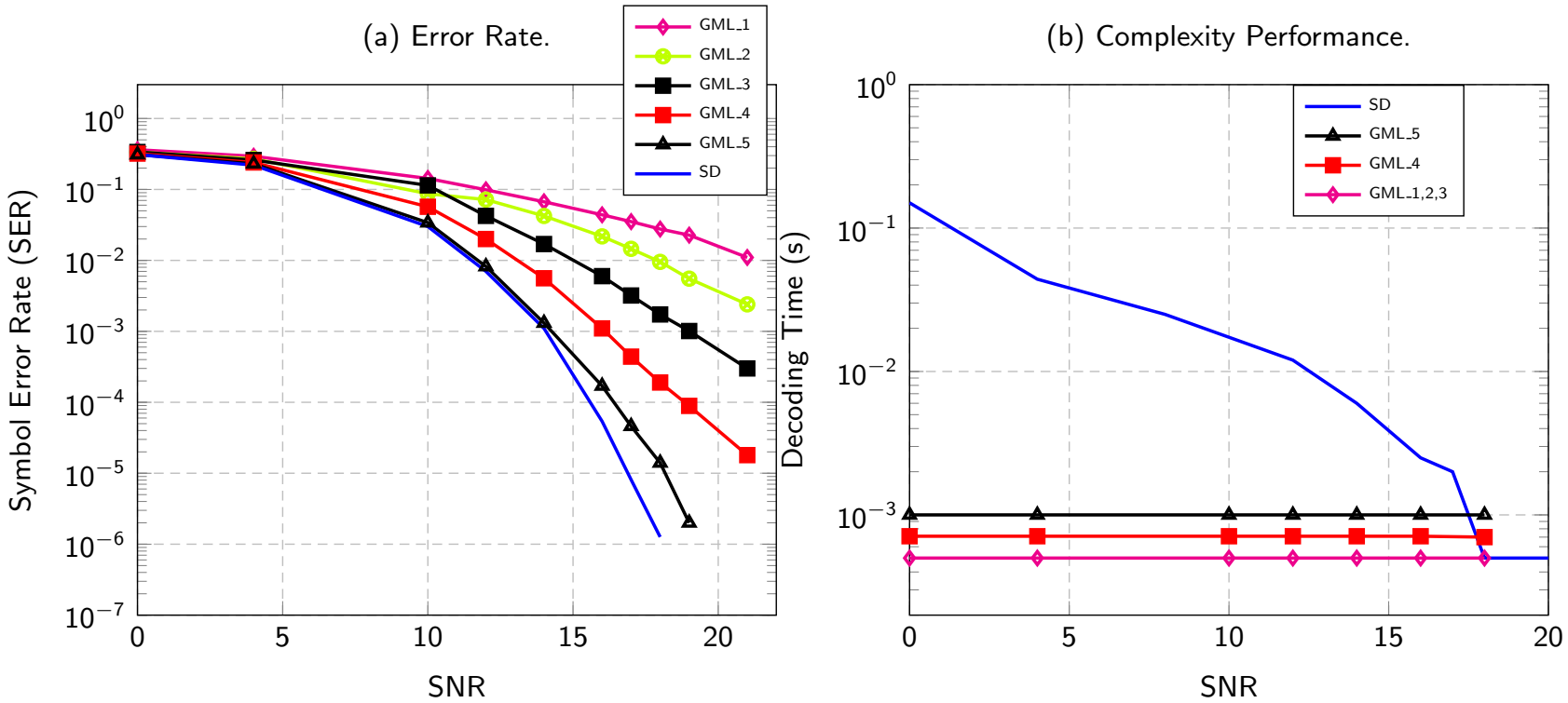


Figure 6: Comparing SD results with our multi-level approach (GML) for a 11×11 MIMO system with 16-QAM modulation.

Performance and Complexity Results

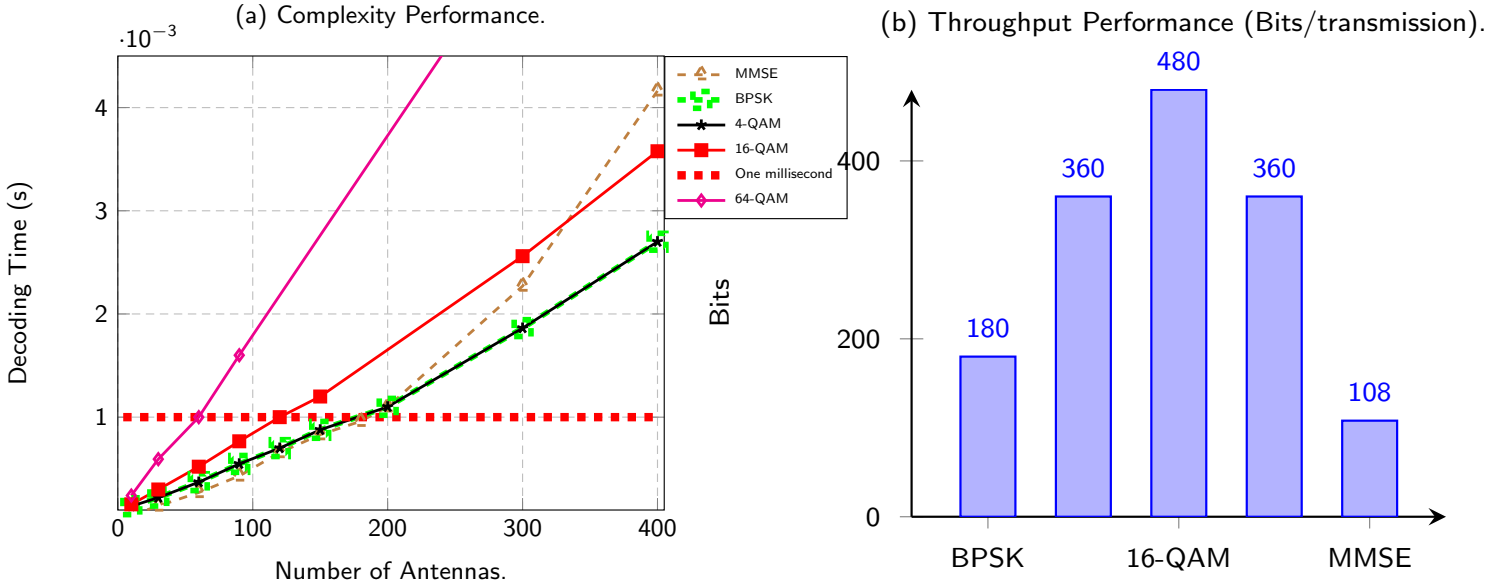


Figure 7: Complexity, modulation, and throughput versus the number of antennas for our GML approach.

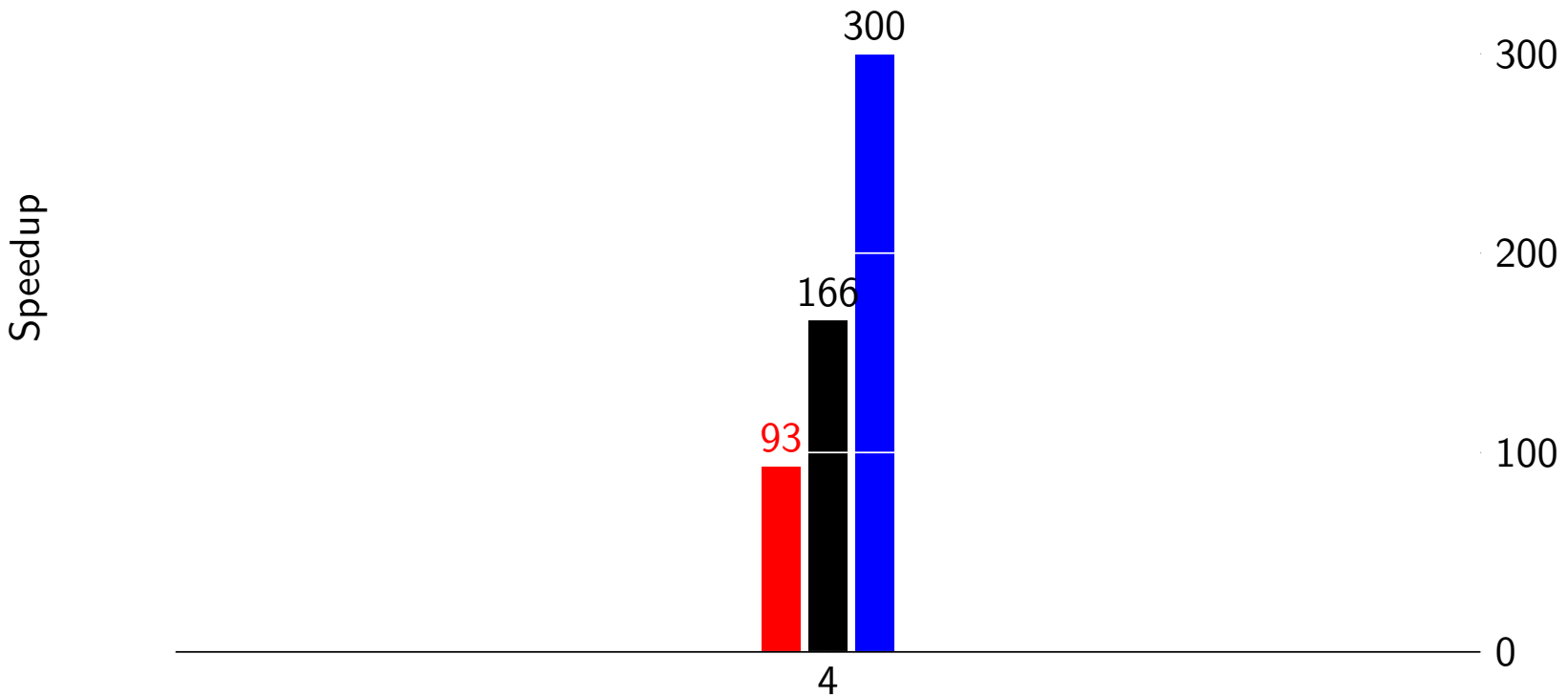
With Ultra-low latency of 5G (1ms)

- Our approach supports Up to 60 antennas using 64-QAM and 120 antennas with 16-QAM.
- 4.5× throughput increase compared to linear MMSE.

GML_3

- Up to 93 times faster than a similar reference CPU implementation on Intel IceLake.

(b) Speedup of our GPU multi-level approaches.



Scalability Results (2.3 ms latency)

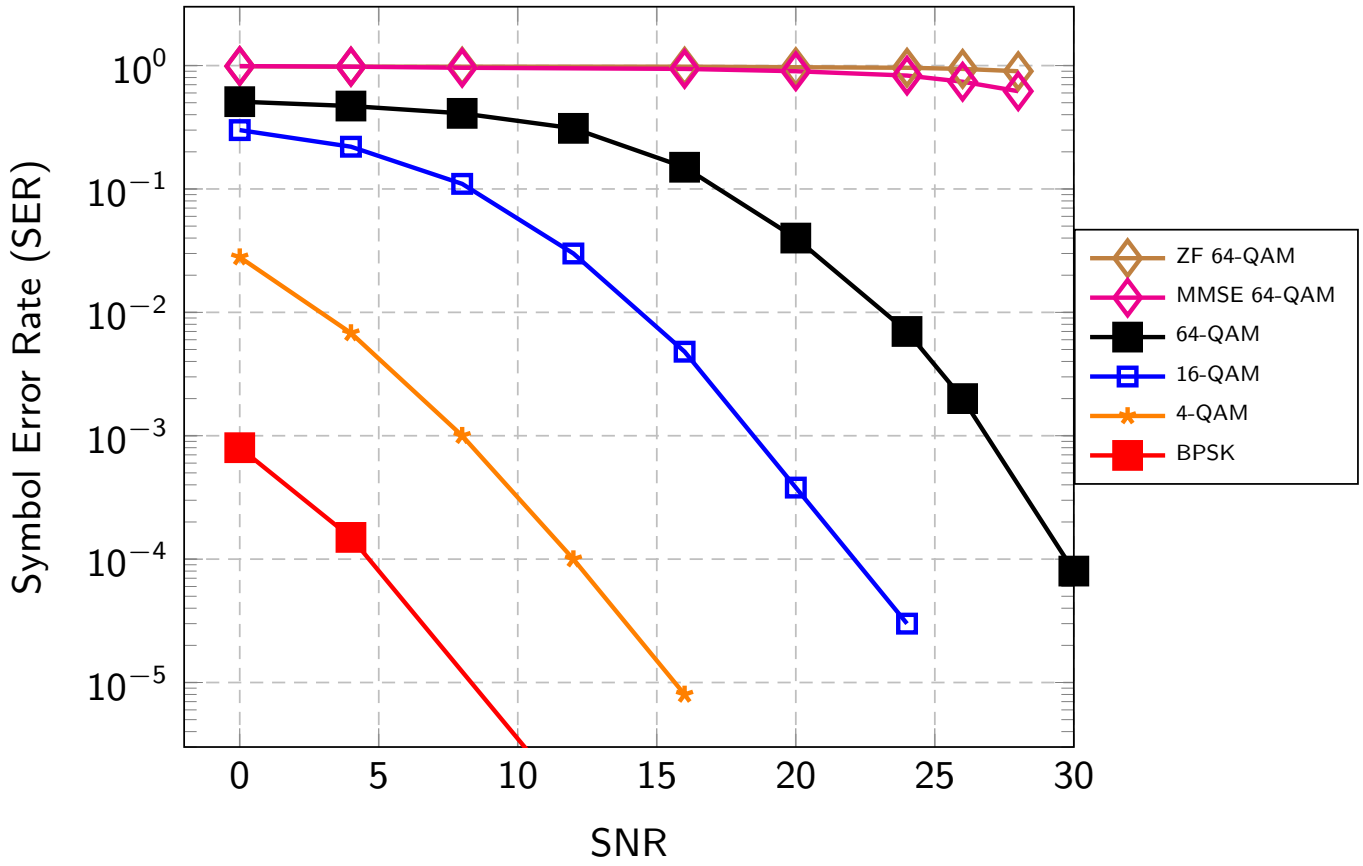


Figure 10: Bits per transmission Vs. modulation for a 128×128 MIMO system with three levels.

Throughput performance

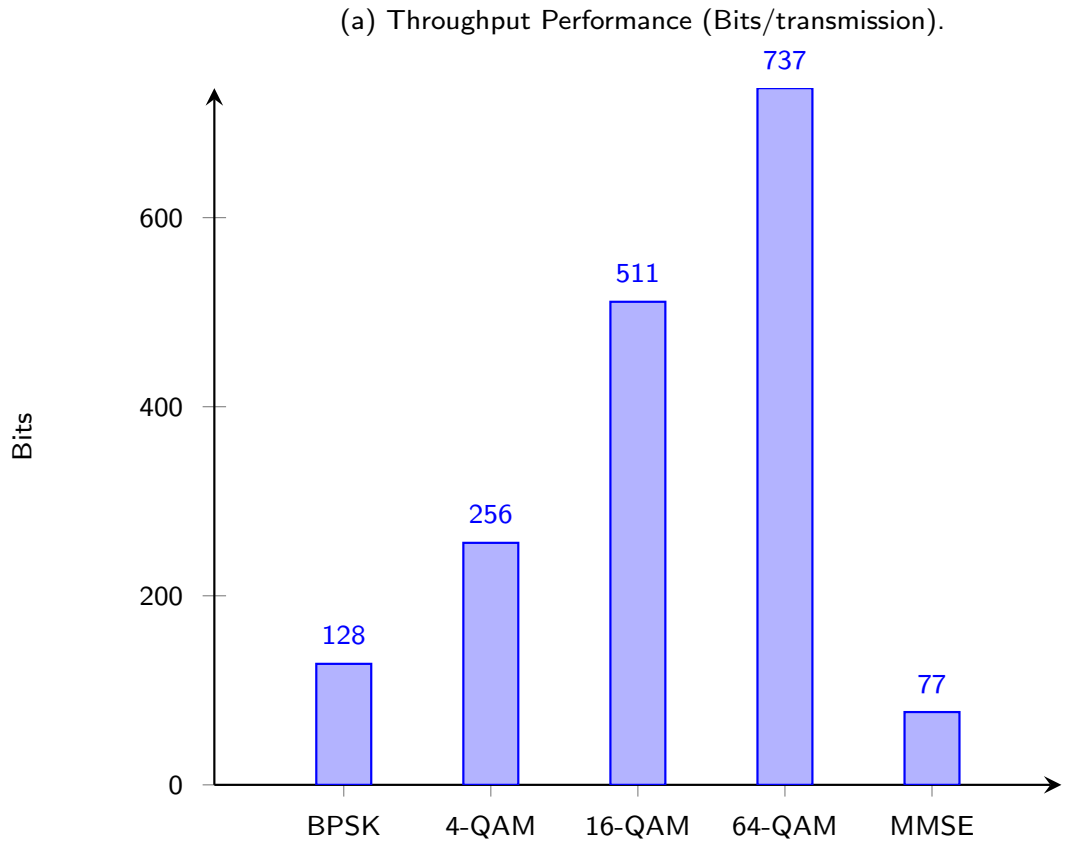


Figure 11: Throughput Vs. modulation for a 128×128 MIMO system (SNR=22 dB).

Conclusion

- Up to 8× throughout improvement compared to Linear MMSE algorithm at a practical SNR.
- The importance of designing new algorithms on new HPC hardware is critical to meet the requirements for next-generation wireless communication networks.

Approach	Latency	Nb antennas	Low Error rate	SNR
Multi-sphere [2]	>10 ms	16	++	25 dB
Flexcore [1]	>10 ms	12	++	22 dB
MMSE	<10 ms	±600	-	35 dB
Our approach	**< 10 ms**	**400**	**+++**	**21 dB**

Table 1: Our approach vs. existing works for uncoded MIMO system with 64-QAM modulation.

Thank you!