FINER-GRAINED DECOMPOSITION FOR PARALLEL QUANTUM MIMO PROCESSING

Minsung Kim and Kyle Jamieson
Princeton University

ABSTRACT

Exploiting (near-)optimal MIMO signal processing algorithms in the next generation (NextG) cellular systems holds great promise in achieving significant wireless performance gains in spectral efficiency and device connectivity, to name a few. However, it is extremely difficult to enable optimal processing methods in the systems, since the required computational amount increases exponentially with more users and higher data rates, while available processing time is strictly limited. In this regard, quantum signal processing has been recently identified as a promising potential enabler of the (near-)optimal algorithms in the systems, since quantum computing could dramatically speed up the computation via non-conventional effects based on quantum mechanics. Given existing quantum decoherence and noise on quantum hardware, parallel quantum optimization could accelerate the process even further at the expense of more qubit usage. In this paper, we discuss the parallelization of quantum MIMO processing and investigate a spin-level preprocessing method for relatively finer-grained decomposition that can support more flexible parallel quantum signal processing, compared to the recently reported symbol-level decomposition method. We evaluate the method on the state-of-the-art analog D-Wave Advantage quantum processor.

Index Terms— Quantum MIMO Processing, Parallel Quantum Annealing, Problem Decomposition

1. INTRODUCTION

One of the representative challenging processing jobs in the physical layer (PHY) of wireless systems is *multi-user (MU) precoding* and *decoding* at *Multiple-Input Multiple-Output* (MIMO) base station systems. The precoding and decoding are dual techniques to enable parallel streams to service multiple users concurrently on a single time-frequency resource, in downlink and uplink, respectively. As the MIMO dimension increases (*i.e.,* larger parallel streams), significant gains in spectral efficiency or connectivity become achievable efficiently, and therefore MU-MIMO becomes an essential block in NextG wireless systems. However, these techniques start to suffer from the trade-off between linear sub-optimal processing versus non-linear (near-)optimal processing when the wireless systems aim larger MIMO dimensions for the NextG performance, for the following reasons.

Linear processing is a MIMO signal processing method that is deployed in the current systems due to its straightforward implementation and low computational complexity [1, 2, 3]. While its precoding and decoding performance is sub-optimal, even linear methods like *Minimum Mean Square Error* (MMSE) can achieve the near-optimal performance, when a base station (BS) equipped with many antennas (also known as *massive* MIMO base station systems) serves relatively small number of users at a time (*e.g.,* 8×64 MIMO; 8 users and 64 base station antennas). However, this required high ratio between BS antenna counts and user counts is becoming a limiting factor to large MIMO dimensions that are desirable for the NextG performances, since the number of parallel streams is decided by the number of concurrently served users. With given BS antenna counts, merely increasing MIMO dimensions to serve more users will cause severely high error rates with sub-optimal linear processing algorithms [4, 5].

Maximum Likelihood (ML) processing can improve the precoding and decoding performance significantly for large MIMO dimensions even when user counts approach to BS antenna counts (*i.e.,* towards maximum MIMO dimensions). ML processing is the best possible method which guarantees theoretically optimal performance in terms of the minimum error rates. However, its required computational amount increases at an exponential rate both with the number of the concurrently serviced users and with data rate of each user. Thus, it is challenging to enable the ML processing for large MIMO dimensions in current wireless systems, where both computing resources and allowed processing time are limited.

Quantum-Accelerated MIMO ML Processing. Recently, quantum computing has been identified as a promising potential enabler of *fast* near-ML processing in MIMO wireless systems with its great acceleration potential via non-conventional computation based on quantum mechanics [6, 7, 8, 9, 10, 11]. On current and near-future quantum hardware, users can utilize only limited quantum fluctuations due to existing quantum decoherence and noise. Thus, how to make use of given limited quantum fluctuations to solve hard optimization problems is an important challenge. In this regard, parallel quantum optimization could be a promising strategy, accelerating the process even further at the expense of more qubit usage. While recent work reports parallel quantum ML processing based on the user symbol decomposition [12], it supports rather inflexible parallelism, since it uses a coarse-grained decomposition method based on wireless symbols that depend on the modulation size. In this work, we investigate a binary spin-level decomposition method, exploiting quantum input Ising models, for relatively finer-grained decomposition in order to support more flexibly parallel quantum ML processing whose available parallelism is not affected by the modulation.

2. BACKGROUND: QUANTUM ANNEALING

Quantum computing is a new type of computing method based on quantum mechanics. While there are several models of quantum computers that are available, this paper focuses on quantum annealing machines [13, 14, 15] due to the sufficient available qubit counts for real-world application experiments. Quantum annealers are analog heuristic optimizer machines that leverage *quantum annealing* (QA) algorithms based on quantum effects like quantum tunneling [16] in order to traverse all possible states among search space of input optimization problems (from an initial quantum *superposition* state) and thus to find their ground states (*i.e.* the state corresponding to global optimum) at the end of computation in ideal cases.

Currently, QA solves only a certain type of combinatorial optimization problems called *Ising Models* whose variables are *spins* with each of spin s_i either -1 or +1 (*i.e.,* binary). The objective function consists of only linear and quadratic terms, and the goal of the model is to find the spin vector (or configuration) consisting of N_V spins, or $\mathbf{s} = \{s_1, \ldots, s_{N_V}\}$ that minimizes the cost function called *Ising energy* E:

$$E(\{s_1, \ldots, s_{N_V}\}) = \sum_{i \leq j}^{N_V} \mathbf{M}_{(i,j)} s_i s_j. \quad (1)$$

$\mathbf{M} \in \mathbb{R}^{N_V \times N_V}$ is an upper triangular matrix whose elements are Ising coefficients that represent input optimization problems. On QA devices, non-diagonal coefficients g_{ij} are (anti-)ferromagnetic couplings that indicate a preference of correlation between i^{th} and j^{th} spins (*i.e.,* on $s_i s_j = \pm 1$), while diagonal coefficients f_i are local magnetic fields that indicate each spin's preference on $s_i = \pm 1$.

3. QAML: QA FOR MIMO ML PROCESSING

When a base station with N_r antennas serves N_t users simultaneously, wireless signal that the base station receives from the served users can be expressed as $\mathbf{y} = \mathbf{H}\bar{\mathbf{v}} + \mathbf{n} \in \mathbb{C}^{N_r}$, where $\mathbf{H} \in \mathbb{C}^{N_r \times N_t}$ is a channel matrix, $\bar{\mathbf{v}} \in O^{N_t}$ is a transmitted signal from users with O being all possible modulation symbols per user, and $\mathbf{n} \in \mathbb{C}^{N_r}$ is an AWGN vector. Since the base station does not have knowledge on each component in $\mathbf{y}$, it needs to estimate the signal $\hat{\mathbf{v}}$ (ideally $\bar{\mathbf{v}} = \hat{\mathbf{v}}$) based on the perturbed received signal $\mathbf{y}$ and estimated $\hat{\mathbf{H}}$ through pilot symbols. This signal processing is uplink MU-MIMO *decoding* or *detection*, while downlink MU-MIMO precoding is a dual technique of this [17]. Assuming $\hat{\mathbf{H}} = \mathbf{H}$, *Maximum Likelihood* (ML) formulation to acquire the best estimated $\hat{\mathbf{v}}$ is

$$\hat{\mathbf{v}} = \arg \min_{\mathbf{v} \in O^{N_t}} \|\mathbf{y} - \mathbf{H}\mathbf{v}\|^2. \quad (2)$$

The ML processing obtains theoretically optimal solutions, but becoming prohibitive for larger N_t and/or $|O|$, due to the exponentially increasing amount of computation, with $|O|$ being the modulation size. QA could potentially accelerate the computation required for ML processing [11]. In order to solve the ML equation using QA, the ML form (Eq. 2) needs to be translated into the equivalent Ising model (Eq. 1) that is a programmable form on the hardware. After replacing each symbol $v \in O$ in the objective function of Eq. 2 (each element of $\mathbf{v}$) with a symbol-spin(s) linear mapping expression (*e.g.,* ref [7]), the norm expansion will result in an equivalent Ising model whose ground state corresponds to the ML solution. The required N_V is $N_t log_2(|O|)$.

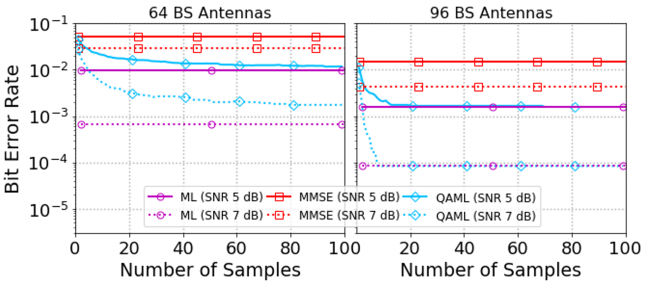


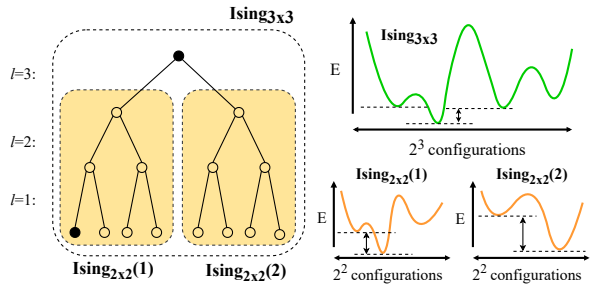
Fig. 1: QAML's bit error rate performance of 64-user BPSK MIMO detection, comparing with MMSE. In the tested QAML, each anneal (sampling) takes 2 μs.

The next step is to program the model coefficients f_i and g_{ij} on the quantum annealer and then the system can run a QA algorithm to solve the problem in a typical way of probabilistic heuristic optimization blackboxes, where each anneal results in a solution candidate or a *sample*. Typically, multiple anneals are conducted per system run to collect multiple samples, and the best sample with the lowest cost $E(\mathbf{s})$ is selected as the final solution. In this paper, we call this series of QA-related processing for ML optimization, *'QAML'*. Figure 1 compares Bit Error Rate (BER) performance of MIMO QAML detection with MMSE for different N_r and SNRs. The figure shows that while the linear MMSE results in poor BER, QAML's BER converges to the optimal ML performance as more samples are collected, showing some promise of the QAML approach. However, note that there exist many unresolved issues on practical (parallel) QAML which will not be discussed in this paper (but available in the references). Similar physics-inspired computing approaches are available in [18, 4, 19, 20].

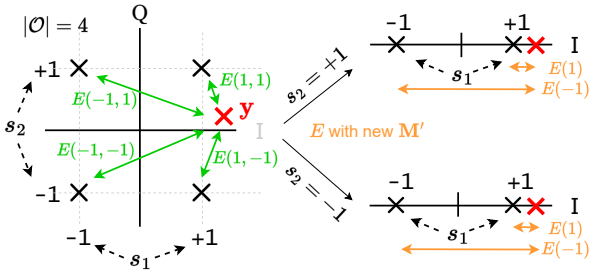
4. QAML PARALLELIZATION

With the trend of exponentially increasing qubit counts on QA devices, parallelizing QA optimization will become more and more important to make better use of increasing qubits and thus accelerate overall optimization process. However, how to make use of extra qubits for task parallelism in QA optimization (cf. data parallelism) is an open question which has not been often discussed and studied so far.

Same Instance Multi-Programming Approach. One of the most straightforward methods of QA parallelization is to program the same (QAML) Ising problem onto hardware multiple times to collect more samples per anneal call [21]. Considering that QA processing is a probabilistic technique, more collected samples are favorable to ensure that the global optimum ML sample is collected at least once, since only the best anneal result among them will be filtered at the end.



(a) Pictorial Ising decomposition.



(b) Example impact on 1×1 QPSK MIMO detection.

Fig. 2: Ising search decomposition. By fully expanding hard variables, sub-problems are formulated. Only one of them retains the ground state of the original problem, but each of them is generally easier to solve than the original one.

Search/Problem Decomposition Approach. Decomposition methods split a ML Ising problem into multiple different sub-problems that can be processed in parallel. Recent work [12] reports parallel QAML based on a decomposition that is inspired from the conventional parallel tree search algorithm [22]. However, since its decomposition is based on a symbol $v \in \mathcal{O}$, the algorithm can support only rather coarse-grained decomposition that depends on the modulation $\mathcal{O}$, which limits parallelism flexibility of QAML. The symbol-decomposed method for QAML requires $\lceil (N_V - log_2(|\mathcal{O}|)N_{fs}) \cdot |\mathcal{O}|^{N_{fs}} \rceil$ spin variables (or *logical* qubits) for fully-parallel processing, where N_{fs} is the controllable number of fully-expanded user symbols.

5. SPIN-BASED ISING SEARCH DECOMPOSITION

We investigate a spin-based *Ising search decomposition*, which is a classical preprocessing module for the following QAML processing. By leveraging binary spin variables in generated Ising forms (instead of wireless symbols) for decomposition, the method can support relatively finer-grained decomposition than the symbol-based decomposition method for parallel QAML. For example, with 16-QAM, the symbol-based decomposition supports $16, 16^2, 16^3, \cdots$ (power of modulation size) parallelism for fully parallel processing, while Ising search decomposition supports $2, 2^2, 2^3, \cdots$ (power of two) parallelism, regardless of the modulation, which implies more flexible parallelization. Table 1 shows the comparisons for the required physical qubit counts for fully parallel QPSK MIMO QAML on the Pegasus-topology QA hardware [15].

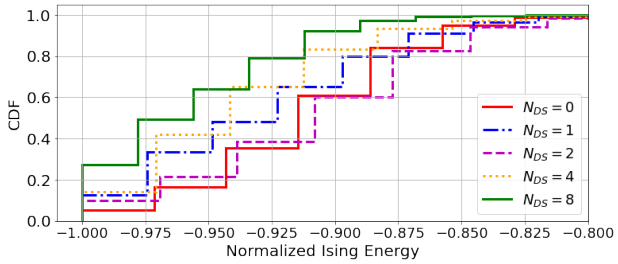
Table 1: The required numbers of *physical* qubits for fully parallel QAML processing for 16×16 QPSK MIMO are shown, considering the limited connectivity on the annealer hardware to program fully-connected ($\forall g_{ij} \neq 0$) ML Ising forms.

Decom. Method & Level	Required Qubits
(Spin-based) N_{DS}: 0 / 1 / 2 / 3 / 4	154 / 300 / 456 / 880 / 1,696
(Symbol-based) N_{fs}: 0 / 1 / 2 / 3 / 4	154 / 456 / 1,696 / 6,272 / 23,040

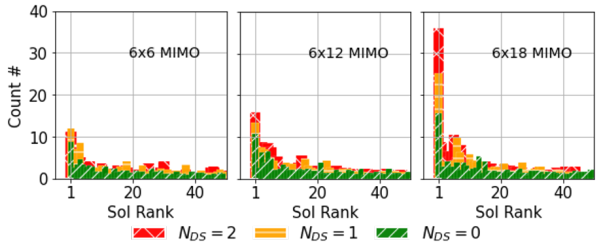
For parallel QAML processing, a system decomposes a ML Ising model into multiple sub-Ising forms by prefixing the hard spin variables. The resulted sub-Ising forms with updated $\mathbf{M}'$ (with reduced variable counts) can be explored simultaneously using parallel QAML. In order to retain the unknown ground state (ML Sol.), each prefixed spin s_i has to be divided into both $s_i = +1$ and $s_i = -1$ (full expansions) as shown in Figure 2(a). Each of the generated sub-problems is generally a easier problem compared to the original one, while only one of them holds the original ground state.

The underlying principle of potential optimization improvements in the decomposition-based parallel QAML (other than reduced variable counts) is that each spin variable has different impacts on E. With a toy example of 1×1 QPSK in Figure 2(b), where spin s_1 corresponds to symbol's I-plane (x-axis), while spin s_2 to symbol's Q-plane (y-axis), let us assume the received signal $\mathbf{y}$ is near x-axis and far from y-axis. In this case, while the different s_1 values ($s_1 = +1$ vs. $s_1 = -1$) lead to large E gaps (*i.e.,* $E(-1, s_2)$ vs. $E(1, s_2)$), the different s_2 values lead to small E gaps (*i.e.,* $E(s_1, -1)$ vs. $E(s_1, 1)$). In other words, spin s_2 is an intuitively harder detection variable for QAML. If the system can choose s_2 as a difficult variable based on the input $\mathbf{M}$, it can prefix its value into $+1$ and -1 for the decomposition, resulting in two Ising forms (only with the s_1 spin) that now have large gaps between $E(1)$ and $E(-1)$ (like two problems with BPSK modulation). For both Ising forms, the solver will likely choose $s_1 = 1$ as the solution relatively easily, which implies that the samples with prefixed s_2 considered will be $\mathbf{s} = \{1, 1\}$ from one sub-problem and $\{1, -1\}$ from the other. Based on the original $\mathbf{M}$, the former that has lower E will be selected as the final solution.

Now we describe the proposed algorithm of Ising search decomposition. Since it is generally observed that the values of diagonal elements $|f_i|$ are larger than non-diagonal elements $|g_{ij}|$ in Eq. 1 of ML [7], we assume the diagonal elements are more important factors that decide the impact on E. First, with $\mathbf{M}_{N_V \times N_V}$ formed, the system chooses the most difficult spin variable to detect (s_i with $1 \leq i \leq N_V$) with $min|f_i|$. Second, the system prefixes the selected s_i into one with $s_i = +1$ and the other with $s_i = -1$, as a full expansion. Then, two different updated $\mathbf{M}'_{(N_V-1)\times(N_V-1)}$ are formed. For the further decomposition, s_i ($1 \leq i \leq N_V - 1$) is chosen as the most difficult spin variable between two decomposed Ising forms using the ratio between $min|f_i|$ and $max|f_i|$. This process iterates for N_{DS} times, with N_{DS} the controllable number of fully-expanded variables for the decomposition, resulting in



(a) Normalzied E distribution (32×32 QPSK MIMO).



(b) Across BS antenna counts (6-user 16-QAM MIMO).

Fig. 3: Impact of Ising search decomposition on parallel QAML with various decomposition levels. In general, further decomposition and parallelism (higher N_{DS}) leads to better QAML result distribution, while more qubits are required.

$2^{N_{DS}}$ sub-Ising forms. After the decomposition, the generated sub-Ising forms can be solved by QAML in parallel (*i.e.,* $2^{N_{DS}}$ parallelism), requiring $(N_V - N_{DS}) \cdot 2^{N_{DS}}$ spin variables for fully-parallel processing.

6. PRELIMINARY EXPERIMENTS

We evaluate the impact of spin-based Ising search decomposition on parallel QAML on the state-of-the-art quantum annealer (D-Wave Advantage [15]) with standard forward annealing schedule. We collect 5,000 anneal samples per setting with relatively-small a few tens of instances. For the purpose of heuristics analysis, we focus on a sub-Ising form that contains the original global optimum among $2^{N_{DS}}$ sub-problems. To test largely parallel problems (*e.g.,* ones with high N_{DS}), separate anneal cycles are used to mimic parallel processing and its performance, since the current QA hardware has limited qubit counts for high-scale parallelism.

We first test parallel QAML processing for 32×32 MIMO detection (SNR 20 dB) with QPSK, varying Ising search decomposition levels (N_{DS}). Figure 3(a) plots CDF of resulted anneal samples across their corresponding E normalized by the absolute global optimum cost, where -1.0 denotes the ML solution. It is observed that as more decomposition levels are applied, the following QAML generally results in samples that are closer to the global optimum (*i.e.,* better anneal quality).

Next, we microbenchmark parallel QAML for 16×16 MIMO detection (SNR 20 dB) with QPSK, comparing the spin-based decomposition against the symbol-based one using the probability of achieving ML solution per anneal or P_{ML}. This metric is directly related to the required time-to-solution or *TTS* (*i.e.,* higher P_{ML} requires less compute time for near-optimal QAML performance), while we leave comprehensive TTS evaluations for our future extended work, since different preprocessing times also need to be considered for overall compute times. Figure 4 shows P_{ML} across applied parallelism. We observe that the spin-based method can support relatively finer-grained decomposition making some unavailable parallelism numbers in the symbol-based method available. This can be further highlighted when with higher-order modulations like 16- or 64-QAM (*e.g.,* 1, 2, 4, 8, $\cdots$ vs. 1, 64, 4096, $\cdots$). Considering the limited qubit count on the hardware (while it keeps increasing), supporting more flexibly parallel QAML will be a good benefit, to further accelerate QAML processing efficiently, although P_{ML} improvements across parallelism are not clearly observed with these tested instances in Figure 4, probably due to the heuristic nature of QA and rather small tested instances and parallelism.

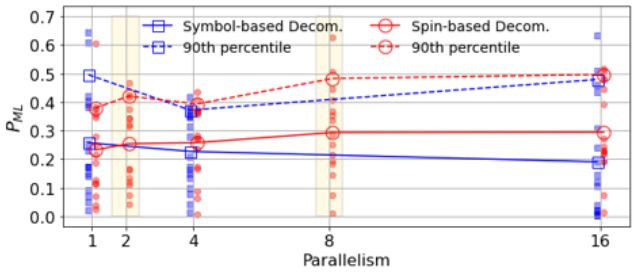


Fig. 4: Microbenchmark of parallel QAML with P_{ML} across parallelism with different decomposition methods. While plain lines report mean, symbols do instances. Shadings highlight additionally available parallelism in the spin decomposition.

Lastly, we also test 16-QAM MIMO parallel QAML with a hybrid QA algorithm [8] with 20 instances (SNR 16 dB). Figure 3(b) reports average occurrences for solution ranks (ordered by E) for different BS antenna counts and various N_{DS}, where 300 anneals are conducted per instance. Interestingly, we observe that parallel QAML obtains more clear (P_{ML}) gains across N_{DS} when with more BS antenna counts.

7. CONCLUSION

This short paper introduces a decomposition technique for parallel QAML. Using translated Ising forms, the method can provide relatively finer-grained decomposition, allowing more flexible parallel QAML, compared to the recently studied wireless symbol-based decomposition approach. Such preprocessing can be applied to parallel optimization on any Ising machines such as optical coherent Ising machines and digital annealers whose inputs are Ising models. Furthermore, the method can be used for any applications to make use of parallel QA, since QA-translated input variables are always spins which are not application-specific ones (cf. wireless symbol).

Acknowledgement

This work is supported by the NSF Grant Nos. CNS-1824357 and a Princeton School of Engineering Innovation Fund award. QA experiments are conducted using D-Wave Systems' QPU access time donation to the PAWS laboratory.

8. REFERENCES

[1] Jian Ding, Rahman Doost-Mohammady, Anuj Kalia, and Lin Zhong, "Agora: Real-time massive mimo baseband processing in software," in *Proceedings of the 16th CoNEXT*, 2020, pp. 232–244.

[2] Clayton Shepard, Hang Yu, Narendra Anand, Erran Li, Thomas Marzetta, Richard Yang, and Lin Zhong, "Argos: Practical many-antenna base stations," in *Proceedings of ACM MobiCom*, 2012, pp. 53–64.

[3] Qing Yang, Xiaoxiao Li, Hongyi Yao, Ji Fang, Kun Tan, Wenjun Hu, Jiansong Zhang, and Yongguang Zhang, "BigStation: Enabling scalable real-time signal processing in large MU-MIMO systems," *ACM SIGCOMM Computer Communication Review*, vol. 43, no. 4, pp. 399–410, 2013.

[4] Minsung Kim, Salvatore Mandrà, Davide Venturelli, and Kyle Jamieson, "Physics-inspired heuristics for soft mimo detection in 5g new radio and beyond," in *ACM MobiCom*, 2021, pp. 42–55.

[5] Konstantinos Nikitopoulos, Juan Zhou, Ben Congdon, and Kyle Jamieson, "Geosphere: Consistently turning MIMO capacity into throughput," in *Proc. of the ACM SIGCOMM Conf.*, 2014, pp. 631–642.

[6] JC De Luna Ducoing and Konstantinos Nikitopoulos, "Quantum annealing for next-generation mu-mimo detection: Evaluation and challenges," in *ICC 2022-IEEE International Conference on Communications*. IEEE, 2022, pp. 637–642.

[7] Minsung Kim, Davide Venturelli, and Kyle Jamieson, "Leveraging quantum annealing for large mimo processing in centralized radio access networks," in *ACM SIGCOMM*, 2019, pp. 241–255.

[8] Minsung Kim, Davide Venturelli, and Kyle Jamieson, "Towards hybrid classical-quantum computation structures in wirelessly-networked systems," in *Proceedings of the 19th ACM Workshop on Hot Topics in Networks*, 2020, pp. 110–116.

[9] Zsolt I Tabi, Ádám Marosits, Zsófia Kallus, Péter Vaderna, István Gódor, and Zoltán Zimborás, "Evaluation of quantum annealer performance via the massive mimo problem," *IEEE Access*, vol. 9, pp. 131658–131671, 2021.

[10] Jingjing Cui, Yifeng Xiong, Soon Xin Ng, and Lajos Hanzo, "Quantum approximate optimization algorithm based maximum likelihood detection," *IEEE Transactions on Communications*, vol. 70, no. 8, pp. 5386–5400, 2022.

[11] Minsung Kim, Srikar Kasi, P Aaron Lott, Davide Venturelli, John Kaewell, and Kyle Jamieson, "Heuristic quantum optimization for 6g wireless communications," *IEEE Network*, vol. 35, no. 4, pp. 8–15, 2021.

[12] Minsung Kim, Davide Venturelli, John Kaewell, and Kyle Jamieson, "Warm-started quantum sphere decoding via reverse annealing for massive iot connectivity," in *Proceedings of the 28th Annual International Conference on Mobile Computing And Networking*, 2022, pp. 1–14.

[13] Vasil Denchev, Sergio Boixo, Sergei Isakov, Nan Ding, Ryan Babbush, Vadim Smelyanskiy, John Martinis, and Hartmut Neven, "What is the computational value of finite range tunneling?," *Physical Review X*, vol. 6, pp. 031015, 2016.

[14] Catherine C McGeoch, "Adiabatic quantum computation and quantum annealing: Theory and practice," *Synthesis Lectures on Quantum Computing*, vol. 5, no. 2, pp. 1–93, 2014.

[15] Catherine McGeoch and Pau Farré, "Advantage processor overview," *D-Wave Technical Review*.

[16] Tadashi Kadowaki and Hidetoshi Nishimori, "Quantum annealing in the transverse Ising model," *Phys. Rev.*, vol. E, no. 58, pp. 5355–5363, 1998.

[17] Srikar Kasi, Abhishek Kumar Singh, Davide Venturelli, and Kyle Jamieson, "Quantum annealing for large mimo downlink vector perturbation precoding," in *IEEE ICC*, 2021, pp. 1–6.

[18] Abhishek Kumar Singh, Kyle Jamieson, Peter L McMahon, and Davide Venturelli, "Ising machines' dynamics and regularization for near-optimal mimo detection," *IEEE Transactions on Wireless Communications*, vol. 21, no. 12, 2022.

[19] Jaijeet Roychowdhury, Joachim Wabnig, and K Pavan Srinath, "Performance of oscillator ising machines on realistic mu-mimo decoding problems," 2021.

[20] Abhishek Kumar Singh, Davide Venturelli, and Kyle Jamieson, "A finite-range search formulation of maximum likelihood mimo detection for coherent ising machines," *arXiv preprint arXiv:2205.05020*, 2022.

[21] Elijah Pelofske, Georg Hahn, and Hristo N Djidjev, "Parallel quantum annealing," *Scientific Reports*, vol. 12, no. 1, pp. 1–11, 2022.

[22] Luis Barbero and John Thompson, "Fixing the complexity of the sphere decoder for MIMO detection," *IEEE Transactions on Wireless Communications*, vol. 7, no. 6, pp. 2131–2142, June 2008.