Towards Hybrid Classical-Quantum Computation Structures in Wirelessly-Networked Systems

Minsung Kim
Princeton University
NASA Ames Research Center, QuAIL
USRA Research Institute for Advanced Computer Science

Davide Venturelli
NASA Ames Research Center, QuAIL
USRA Research Institute for Advanced Computer Science

Kyle Jamieson
Princeton University

ABSTRACT

With unprecedented increases in traffic load in today's wireless networks, design challenges shift from the wireless network itself to the computational support behind the wireless network. In this vein, there is new interest in quantum-computing approaches because of their potential to substantially speed up processing, and so improve network throughput. However, quantum hardware that actually exists today is much more susceptible to computational errors than silicon-based hardware, due to the physical phenomena of decoherence and noise. This paper explores the boundary between the two types of computation—classical-quantum hybrid processing for optimization problems in wireless systems—envisioning how wireless can simultaneously leverage the benefit of both approaches. We explore the feasibility of a hybrid system with a real hardware prototype using one of the most advanced experimentally available techniques today, *reverse quantum annealing*. Preliminary results on a low-latency, large MIMO system envisioned in the 5G New Radio roadmap are encouraging, showing approximately 2–10× better performance in terms of processing time than prior published results.

CCS CONCEPTS

• **Networks → Wireless access points, base stations and infrastructure**; • **Computer systems organization → Quantum computing**.

KEYWORDS

Wireless Networks, Quantum Computing, Hybrid Design, Quantum Annealing, MIMO Detection

ACM Reference Format:
Minsung Kim, Davide Venturelli, and Kyle Jamieson. 2020. Towards Hybrid Classical-Quantum Computation Structures in Wirelessly-Networked Systems. In *Proceedings of the 19th ACM Workshop on Hot Topics in Networks (HotNets '20), November 4–6, 2020, Virtual Event, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3422604.3425924

1 INTRODUCTION

The demand for faster and more capable video, audio, and teleconferencing applications over the past decade has resulted in sharp increases of wireless traffic loads at base stations. To increase the

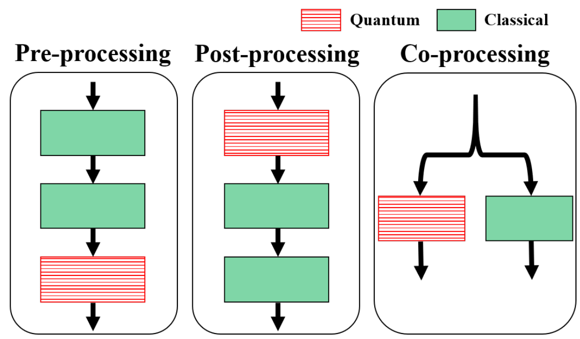

Figure 1: Fundamental computation structures for the coordination of classical-quantum processing units.

spectral efficiency of the entire network, there are many techniques available, most notably the 5G suite of technologies, including Massive MIMO, dense small cells, and millimeter wave communication. These techniques can indeed increase spectral efficiency to a point, but to scale up in terms of users and traffic loads, they quickly demand exponentially more *computational* throughput, at low *computational* latencies [2]. The latter challenge arises from the limited available processing time before a radio must turn-around response to incoming data, a part of the link layer's automatic repeat request functionality in almost all wireless networked systems, such as wireless local area and 4G LTE cellular networks [11, 21, 22, 55]. Forthcoming ultra-low latency designs, envisioned in 5G *New Radio* and beyond [45], tighten these requirements even further.

There has been much recent effort dedicated to faster and more parallel integration of silicon processors into individual chip packages. Indeed, parallel computing is mainstream for high-performance computation at wireless base stations [3, 37, 56]. However, rental of physical space and physical cooling dominate the cost of operating a base station and so the consensus architecture in cellular wireless networks is based on a principle of aggregating many base stations' processing together in a centralized data center via low-latency optic fiber, an architecture called a *centralized radio access network* (C-RAN) [9, 35, 47]. However, even in a C-RAN architecture, practical constraints (in algorithms of wireless networks and/or computer architectures) continue to limit speedup gains on classical computing platforms [19, 20, 30].

This paper takes a long view on these issues: we hypothesize that the aforementioned growth trends in wireless network demand will continue, and so in one to two decades, silicon-based computational resources will likely not suffice, since their clock speed has already

HotNets '20, November 4–6, 2020, Virtual Event, USA

ACM ISBN 978-1-4503-8145-1/20/11...$15.00
https://doi.org/10.1145/3422604.3425924

reached a plateau [10], and routing, power, and heat will constrain parallelism in the future.

To overcome these fundamentals, researchers are beginning to explore quantum computing (or quantum-inspired algorithms [15, 39, 48]) as an alternative to classical computers to solve intractable optimization problems in wireless systems [8, 14, 29, 53]. While error-corrected quantum computation may eventually enable asymptotic speedups over the best possible classical algorithms [42], at its current level of maturity, universal large-scale quantum computers are still many years away and so quantum computers are termed *Noisy Intermediate-Scale Quantum* (NISQ) devices. The power of near-term quantum algorithms [13, 18, 27, 38], which are heuristic methods that tolerate relatively high levels of noise, is currently unknown and the wide belief is that it should be investigated empirically, due to the impossibility of simulating the hardware accurately. Despite the lack of theoretical guidance, the payoff of understanding these methods' performance in these early days could be transformative, similar to what has recently happened in the field of deep learning, thanks to the appearance of computational devices that allowed a sufficiently large models to be trained and tested in practice [32].

The overarching goal of this paper is to give a preliminary assessment of the viability of hybrid classical-quantum computational structures (as shown in Figure 1) on NISQ computers for wireless networks, and to stimulate novel designs of quantum-enabled base stations (in cellular networks) or access points (in Wi-Fi local-area networks) in light of expected improvements in engineering and systems integration of this nascent technology in the next few years. Our two contributions are as follows:

1) We outline a vision of classical-quantum hybrid processing in wireless networked systems that aims to take advantage of both classical and quantum processing. For this paradigm, we share lessons that we learn from possible designs. For example, we explore the feasibility of a classical-quantum hybrid algorithm enabled by a new quantum technique on an analog quantum annealing device.

2) On the quantum annealing device we compare the performance of a *Reverse Annealing* (RA) classical-quantum hybrid protocol against two fully quantum solvers: *Forward Annealing* (FA) and a newly developed *Forward-Reverse* (FR) annealing. To our knowledge this is the first time reverse annealing techniques have been attempted on wireless applications.

While many computationally heavy problems exist that cause bottlenecks in 5G wireless networked systems (such as large-scale channel coding, resource allocation, scheduling, and pre-coding), we pick *Large MIMO detection*, an essential technique to increase wireless throughput that enables parallel data streams for many users (*spatial multiplexing*) in the same wireless spectrum. However, to make full use of spatial multiplexing, much more sophisticated receiver designs with (near) optimal detectors are required [20, 40, 46, 55]. Preliminary work has shown the possibility of reducing these problems to a form amenable to solution on a quantum computer, and provides performance baselines [7, 8, 14, 29].

Our experimental results show that RA starting from a candidate solution obtained by a fast greedy search outperforms all other tested methods for all channel modulations in terms of required compute time to finding the optimal decoding. For reference, for an eight-user, 16-QAM detection/decoding problem, our version of RA achieves approximately up to 10× higher success probability than the previously published results for FA. The application-specific classical solvers and the compilation parameters are standard and have not been tailored, so this result can be improved further without any hardware advances. In this regard, we discuss possible design thoughts towards our eventual design.

2 BACKGROUND AND RELATED WORK

Quantum computers for optimization: Quantum computers feature hardware that uses unique information processing capabilities based on quantum mechanics to perform calculations. While there remain much to investigate on the ultimate power of quantum computing, optimization is one of the key applications that the quantum computing community has identified as interesting in the short-term due to two main approaches that have shown promise in NISQ devices: *Quantum Annealing* (QA) [27, 38] and *Quantum Approximate Optimization Algorithms* (QAOA) [13, 18]. While QA and QAOA are different hardware (the former is analog, the latter digital) they have in common that both methods work on classical combinatorial problems which are often expressed as an *Ising Model*. For the purpose of this work, this model is trivially equivalent to a *Quadratic Unconstrained Binary Optimization* (QUBO) form [6, 54], which can be expressed as finding the bitstring $\hat{q}_1, \ldots, \hat{q}_N$ that minimizes the cost function:

$$E(\{q_1, \ldots, q_N\}) = \sum_{i \le j}^{N_v} Q_{ij} q_i q_j, \tag{1}$$

where N_v is variable count and $\mathbf{Q} \in \mathbb{R}^{N \times N}$ is upper triangular matrix, each of elements representing QUBO coefficients and each of binary variables q either 0 or 1. Quantum-inspired computing devices are often called *Ising Machines* (see e.g. [23]), implementing hardware that physically encodes the objective function 1 as a measurable observable of a physical system. The workflow of both quantum and quantum-inspired methods can be usually described as follows: After reducing the problem of interest into the Ising/QUBO form and programming the form on the quantum hardware, *Quantum Processing Unit* (QPU) on the hardware solves it in a way typical of heuristics blackboxes.[1] In general, multiple calls N_s are made to the device and the best sample (e.g. the one with the lowest QUBO cost function) is selected as the final solution.

Classical-quantum hybrid approaches: Since quantum computers were realized in experimental hardware, there have already been some studies on classical-quantum hybrid approach despite its relative novelty. For instance, the sequential use of a classical method as input for a quantum one (or viceversa) is a straightforward hybrid approach. In [50], quantum annealing was used to generate solution candidates checked by classical computing that more easily deal with hard constraints. Classical computing can also ease the problem by prefixing some variables as part of iterative loops [28] or as a pre-processing [16, 33, 34]. Other examples of hybridization include locality reduction simplifications [24, 25], problem decompositions [44, 58]. A solver block design consisting of multiple quantum annealing processors hybridized with Tabu search is also

[1] For details on how it works in some concrete use cases, see for instance Refs [5, 12, 13, 18, 27, 38].

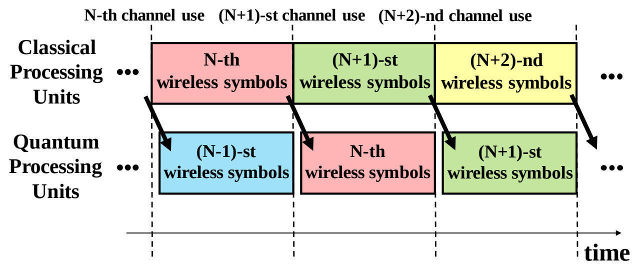

Figure 2: Example pipeline design of hybrid computational structure for successive wireless channel uses.

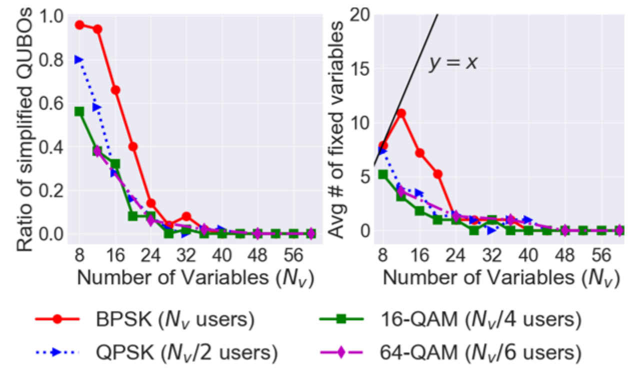

Figure 3: Empirical results of the simplifying QUBO scheme for 50 instances of MIMO detection across different problem sizes and modulations: *(Left)* ratio of simplified QUBOs and *(Right)* average number of fixed variables in the simplified cases. The scheme achieves nearly no-effect for problems over 32-40 variables.

in the commercial offerings of D-Wave Systems [1]. A qualitatively novel hybrid scheme introduced recently is known as reverse annealing [49, 52], which we further explore with our initial prototype in Section 4 for MIMO detection problems.

3 DESIGN CHALLENGES

This section introduces three design challenges we faced in our exploration, as well as early design attempts that failed to deliver significant advantage by means of hybridization, for illustrative and educational purposes.

Challenge 1: Hybridization design. The most crucial challenge consists in choosing the partitioning scheme of the computation between classical and quantum processing units. Pre-processing, post-processing and co-processing orchestration between quantum and classical modules need to be well specified. To our best knowledge, none of classical-quantum hybrid techniques have been empirically tested on problems in wirelessly networked systems, on a real hardware prototype.

Challenge 2: Optimal parameters. The designed hybrid system will include specifications on parameters crucial to the performance of the quantum and classical optimizers. The best (interdependent) parameter values for each processing unit, *i.e.*, those that maximize end-to-end performance of the system, need to be identified.

Challenge 3: Pipelining classical and quantum computation. The eventual goal of a hybrid classical-quantum architecture is to enable *ultra-high throughput* networks. To this end, how to best divide sequential computational tasks into classical *or* quantum units and then assign those units to staged processing units is crucial. Fortunately, the nature of the sequential arrival of traffic over a wireless link lends itself to such pipelining, as illustrated in Figure 2 where data bits from successive "channel uses" are processed in stages of the computational pipeline. However, typically more complex considerations are required for a pipelined system such as balancing, buffering, and costs and these become even more challenging in a classical-quantum hybrid system.

3.1 Initial Attempts

In this section, we introduce our two initial attempts of hybrid algorithms for wireless applications that achieve no or very limited gain with respect to the non-hybrid use of the quantum machine. However, note that while these schemes were not beneficial in our prototype, they may work with more sophisticated algorithms and/or with next-generation quantum devices.

Simplifying the QUBO form: A QUBO may be able to be simplified through the execution of a simple classical algorithm that would determine the value of some binary variables before quantum processing, depending on the value of its coefficients [34]. For $Q_{ii} > 0$ $(\forall i)$, if the absolute sum of negative coefficients of $(\forall k)$ Q_{ik} and Q_{ki} $(< 0$ and $k \neq i)$ is less than Q_{ii}, then the i-th variable can be fixed to 1. In a similar way, certain QUBO variables can be prefixed into 0. Prefixing one logical variable reduces the search space by a factor of two. We test this scheme varying the problem size and modulations in Figure 3. We observe that no simplifications are detected when the MIMO problems have over 32–40 variables, regardless of the chosen modulations. Given that computationally-heavy optimization problems in envisioned 5G wireless networks contain more than 50 or even hundreds of variables, this pre-processing scheme does not seem particularly useful.

Soft information to narrow the search space: Pre-knowledge of variables (wireless symbols) that are very likely to be (un)assigned a certain value in the unknown global optimum solution might ease the problem. This pre-knowledge is equivalent to soft-information in wireless networks which are already applied in many applications and can be obtained in various ways [20, 28, 31, 51, 57]. We test a new conservative algorithm of utilizing the pre-knowledge, by adding constraints as a classical pre-processing to force following quantum processing to search only spaces of more promising QUBO variables or symbols. For instance in Figure 4, assuming we learn highly possible spaces (green color-coded), then we can add constraint terms to the original QUBO to avoid unlikely spaces (red color-coded) without harming the global optimum (ideally). This approach seemingly looks useful, but it is difficult to find proper constraint factors on noisy, analog quantum machines especially for problems of more variables and multiple constraint factors and our empirical investigations have shown that it is not currently practical.

4 AN EXPLORATORY STUDY

In this section, we present an early-stage framework of classical-quantum processing for a specific problem (MIMO detection) and

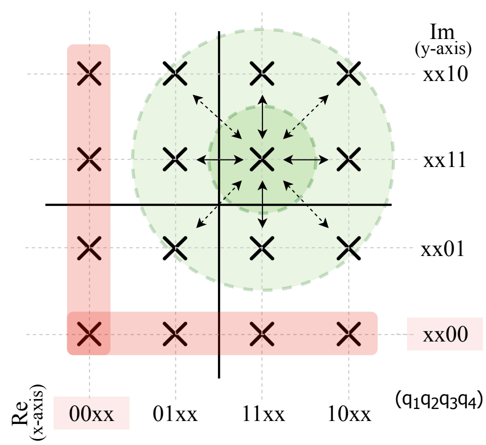

Figure 4: An example of adding constraints based on a pre-knowledge of symbols on a gray-coded constellation of 16-QAM modulation. Given that $q_1q_2q_3q_4$ is very likely to be close to 1111, adding constraint terms $C_1 \cdot (q_1 - 1) \cdot (q_2 - 1)$ and $C_2 \cdot (q_3 - 1) \cdot (q_4 - 1)$ to the QUBO forces the searching process to avoid unlikely symbols (red color-coded).

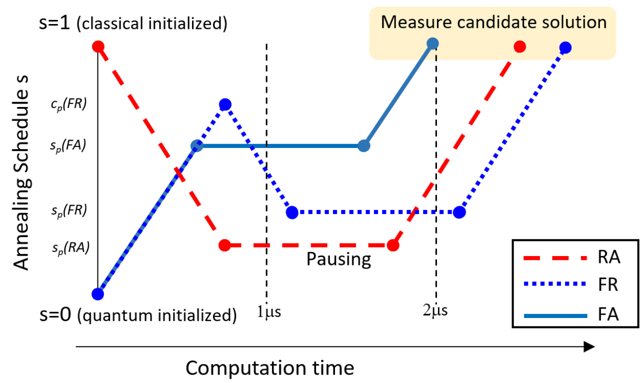

Figure 5: The three flavors of quantum annealing (FA, RA and FR) investigated in this study. The annealing schedule s determines the inverse strength of the signal that uses quantum effects to compute. If $s = 0$, the quantum annealer realizes a fully quantum state that would return a random bitstring if measured. For intermediate $0 \le s \le 1$, the annealer is optimizing exploiting quantum fluctuations to explore the feasible solutions. At $s = 1$ quantum fluctuations are suppressed and the system can be treated as a classical memory register with a stored result.

method. We introduce our initial prototype design in Sec 4.1 and its implementation on the analog-based quantum device in Sec 4.2. Results are in Sec 4.3.

4.1 Prototype Design

Our hybrid prototype design consists of two sequential modules; a Greedy Search (GS) classical algorithm and then a Reverse Annealing (RA) algorithm implemented on a *D-Wave 2000Q* quantum annealer. GS is a very simple deterministic QUBO solver featuring linear complexity. RA is a variation of quantum annealing (see Figure 5) where computation starts from a programmable classical initial state and thus enables sequential classical-quantum processing. Specifically:

(1) Classical Module: initial greedy search. Initially, GS solves the QUBO with a candidate solution determined by greedy descent [52].The bits are sorted in ascending order by the magnitude of $|\frac{1}{2}Q_{ii} + \frac{1}{4}\sum_{k=1}^{i-1} Q_{ki} + \frac{1}{4}\sum_{k=i+1}^{N} Q_{ik}|$.[2] The first bit is assigned $q_i = 0$ if the corresponding magnitude is positive and 1 otherwise. Then the procedure is iterated recursively on the remaining variables by assigning the value that minimize the energy of the QUBO form considering only the variables that are set, until all variables are determined. As expected, the greedy solution is often not the global optimum, but it is a good initial guess that requires nearly negligible computation time and resources. By choosing the simplest classical module, we focus on viability of hybrid design, excluding the discussion on design challenges introduced in Section 3.

(2) Quantum Module: reverse annealing. RA is a refined local quantum annealing starting from a known initial state, which has shown improved performance over FA for many different applications [41, 49, 52]. With reference to Figure 5, in our prototype, RA is initialized with the solution classical state of GS ($s = 1$) and starts annealing backward (R) up to s_p ($0.0 \le s_p \le 1.0$), pausing (P) for t_p microseconds[3], before switching to forward (F). The comparison between the RA and the FA programming steps can be stated in terms of the schedule steps [time (us), s $(0.0 - 1.0)$] follows:

- Forward Annealing (FA):
  $[0.0, 0.0] \xrightarrow{F} [s_p, s_p] \xrightarrow{P} [s_p + t_p, s_p] \xrightarrow{F} [t_a + t_p, 1.0]$
- Reverse Annealing (RA):
  $[0.0, 1.0] \xrightarrow{R} [1.0 - s_p, s_p] \xrightarrow{P} [1.0 - s_p + t_p, s_p] \xrightarrow{F} [2(1.0 - s_p) + t_p, 1.0]$,

where t_a is *anneal time*, t_p is anneal *pausing time*, and the mid-point phase s_p denotes *pause location* for FA and *switch+pause location* for RA. These are all parameters that need to be optimized. While total duration of FA depends on anneal time t_a, RA total duration depends on switch and pause location s_p.

Forward reverse annealing. In addition to FA, we test a newly-developed fully-quantum forward reverse annealing (FR) as another comparison scheme. The combination of FA+RA (*i.e.,* FA→ solution state →RA) has been tested before to solve Low-Rank Matrix Factorization [41], in two steps, but in our implementation we program the annealer to execute a forward reverse annealing in a single step, where RA initial state is determined by the FA procedure without doing a measurement, initializing the reverse annealing with the state at s=c_p. FR's anneal schedule is the following:

- Forward Reverse Annealing (FR):
  $[0.0, 0.0] \xrightarrow{F} [c_p, c_p] \xrightarrow{R} [2c_p - s_p, s_p] \xrightarrow{P} [2c_p - s_p + t_p, s_p] \xrightarrow{F} [2c_p - 2s_p + t_p + t_a, 1.0]$.

[2]In other words, the bits are sorted by the absolute magnitude of matrix's diagonal elements in the Ising model.

[3]It has been shown that the annealing pause brings out improvements for FA [26, 29, 36] and for RA [52].

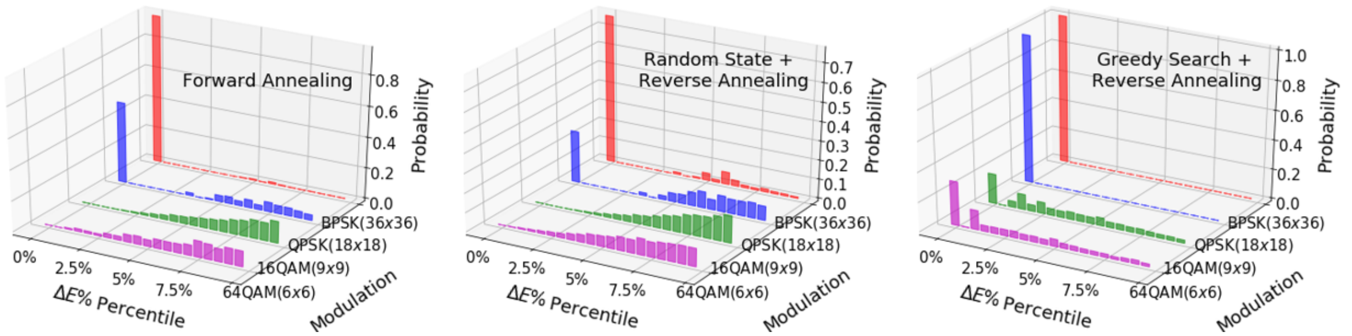

Figure 6: Average distribution of cost function value percentile out of 200,000-600,000 anneal samples of 20 instances of 36-variable decoding problems for different modulations and algorithms: *(Left)* forward annealing or QuAMax [29], *(Center)*: reverse annealing starting at a randomly picked initial state, *(Right)* reverse annealing starting at the result state of greedy search (hybrid processing with the simplest classical solver).

4.2 Implementation

We implement our prototype on the D-Wave 2000Q, a state-of-the-art analog-based quantum device to explore the possibility of improvement over pure quantum processing. We set the anneal time to be $t_a = 1\mu s$ for FA (the minimum compute time allowed by the hardware) and the pausing time $t_p = 1\mu s$ for FA, FR, and RA, consistently to the guidance in the literature for best performance. For s_p and c_p, we set its range from 0.25–0.99 (in steps of 0.04). We collect statistically enough anneal samples per setting (N_s at least 10,000) to analyze the performance. We synthesize 10-20 (QUBO) instances of random MIMO detection for various user numbers and modulations (BPSK, QPSK, 16-QAM, and 64-QAM) with unit gain signal and unit gain wireless channel with random phase. The method of reducing MIMO detection into QUBO form based on mapping rule between QUBO variables and wireless symbols was introduced in Ref. [29] and we apply the same mapping. In the experiments, we exclude the wireless noise (AWGN). For FA, we use data from the paper [29], and different algorithms are evaluated for the same QUBO instances as FA.

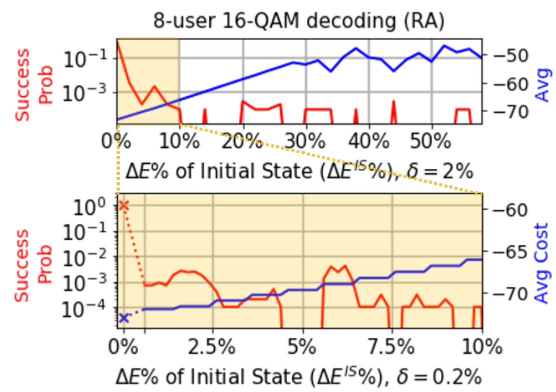

Figure 7: Expectation value of the cost function and success probability out of RA samples for a 8-user 16-QAM decoding instance across different $\Delta E^{IS}\%$. Shaded parts are a close up of the results for $\Delta E^{IS}\%$ below 10% (No initial candidate achieved less than 0.4%).

4.3 Experimental Results

A fundamental metric to evaluate heuristics-based solvers is the distribution of QUBO value of the samples. We define the quality of solution as the percentile of solution cost E_s compared to the best possible solution whose cost is E_g:

$$\Delta E\% = 100 \cdot \left[(|E_g| - |E_s|)/|E_g| \right] .$$

$\Delta E\% = 0\%$ indicates that the global optimum has been found. Figure 6 shows the empirical distribution $\Delta E\%$ out of all anneal samples of a 36-variable MIMO detection solved by FA and RA with median best parameter setting. The lower $\Delta E\%$ means the closer gap between E_s and E_g, which implies closer Euclidean distance to the optimal detection symbol. If the state that initializes RA is randomly selected (Figure 6, *center*) then the method works worse than FA (Figure 6, *left*), skewing the distribution towards low quality solutions. The distribution obtained by running RA after the GS initialization (Figure 6, *right*) shows the most promise and hence it is used for our classical-quantum hybrid prototype design.

Impact of the quality of the initial state. We observe that most of the solutions returned by GS are already scoring approximately $\Delta E\% \leq 10\%$. In order to design our hybrid architecture, we study the correlation between the quality of RA's initial state ($\Delta E^{IS}\%$) and the overall solver performance after the RA. We obtain sample states of various $\Delta E^{IS}\%$ using over 750,000 samples. Figure 7 shows the success probability of finding the optimal solution and average cost computed out of RA samples in one typical 8-user 16-QAM detection instance as a function of $\Delta E^{IS}\%$ (binned in steps of $\delta = 2\%$). We observe that, unsurprisingly, the probability of success and the expectation value for the cost function is generally better if the $\Delta E^{IS}\%$ is low.

Impact of switch and pause location (s_p). The performance is dependent on the parameter "switch and pause location" s_p ($0 \leq s_p \leq 1$). s_p should not be too close to 1, since quantum fluctuations require to be strong enough to perturb the initialized state. At the same time, s_p cannot be too close to 0, since the information related to the initial state would be wiped out by too strong fluctuations, countering all possible advantages of RA as a "refined local search". Recall that the choice of s_p affects the total duration of the computation. For our benchmark, we use a commonly-accepted metric for performance of quantum heuristics, *time-to-solution* (TTS) [43], which indicates the median required time (us) to find the global optimum with *target*

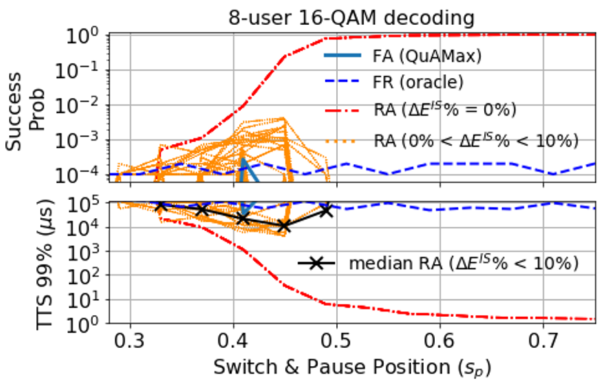

Figure 8: Success probability and TTS of RA compared against FA and FR for a 8-user 16-QAM decoding instance, initialized with different methods and candidate solutions of various quality ($\Delta E^{IS}\%$). The performance is reported as a function of the parameter (s_p).

confidence $C_t\%$:

$$TTS(C_t\%) = duration \cdot \frac{\log(1 - C_t/100)}{\log(1 - p^\star)}, \quad (2)$$

where $p^\star$ is the probability to find the ground-state during a single execution of the solver (FA, FR, and RA). We plot their $p^\star$ and TTS at $C_t = 99\%$ across different s_p in Figure 8. The red dashed line denotes the RA run with the ground state as a RA initial state ($\Delta E^{IS}\% = 0\%$) and each dotted yellow line denotes RA runs initialized with various initial states of $0 < \Delta E^{IS}\% < 10\%$ ($\delta = 0.2\%$). The initial states featuring $\Delta E^{IS}\% \geq 10\%$ are out of our consideration since they are very rarely generated. While FA cannot find the global optimum for the pause locations tested other than $s_p = 0.41$, RA is successful for s_p chosen in the interval 0.33 to 0.49. In the case of FR, $p^\star$ (and thus TTS) result is reported for the best found c_p out of an exhaustive search (the "oracle" scheme). The performance of the method for this instance happens to be worse than FA and RA despite the c_p optimization.

5 CONCLUSION

In this paper, we outlined the landscape of hybrid classical-quantum processing in wireless networks. We share lessons that we learn from possible designs and explore the feasibility of a hybrid design enabled by reverse annealing. Our experimental results on a real hardware prototype show RA achieves improved performance comparing against two fully quantum solvers, even with the simplest classical solver which frequently fails to find the global optimum. Of course, the current state of maturity of the quantum annealing technology, as well as of other quantum-hardware, still forbids real-world deployment at scale due to high costs and system integration overheads that make the use impractical [29]. Further, the results reported are mostly illustrative since they relate to performance on a single typical problem instance. Nevertheless, the speedup observed is a promising indication that a full-scale bench-marking of hybrid reverse annealing solvers is worthwhile to advance the state of art.

One important next step consists in exploring RA-based hybrid designs with application-specific solvers. The combination of application-specific classical solvers and RA is very likely to improve over the GS initialization. Classical approximate solvers for possible combinations with RA include (but are not limited to) linear solvers and tree search-based solvers. Linear solvers (*e.g.,* zero-forcing) can likely achieve better initialization quality $\Delta E^{IS}\%$ than GS, requiring matrix inversion to nullify the wireless channel effect and thus slightly longer compute time, but their process cannot be parallelized. Tree-based solvers (*e.g.,* FCSD [4] and K-best SD [17]) have tunable complexity, enabling parallelism, which could provide some control over $\Delta E^{IS}\%$. However, this flexibility comes at the expense of more complex hybrid designs.

ACKNOWLEDGEMENTS

We thank the anonymous reviewers of this paper for their extensive technical feedback, which has enabled us to significantly improve the work. We also thank the NASA Quantum AI Laboratory (QuAIL), and the Princeton Advanced Wireless Systems (PAWS) Group for useful discussions. This research is supported by National Science Foundation (NSF) Award CNS-1824357 and CNS-1824470, and an award from the Princeton University School of Engineering and Applied Science Innovation Fund. Minsung Kim was also supported by the USRA Feynman Quantum Academy funded by the NAMS R&D Student Program at NASA Ames Research Center (FA8750-19-3-6101).

REFERENCES

[1] 2020. D-Wave Hybrid Solver Service: An Overview. *D-Wave Technical Report 14-1039A-A* (2020).

[2] Jeffrey G Andrews, Stefano Buzzi, Wan Choi, Stephen V Hanly, Angel Lozano, Anthony CK Soong, and Jianzhong Charlie Zhang. 2014. What will 5G be? *IEEE Journal on selected areas in communications* 32, 6 (2014), 1065–1082.

[3] Mark Baker. 2000. Cluster computing white paper. *arXiv preprint cs/0004014* (2000).

[4] Luis Barbero and John Thompson. 2008. Fixing the complexity of the sphere decoder for MIMO detection. *IEEE Transactions on Wireless Communications* 7, 6 (June 2008), 2131–2142.

[5] Sergio Boixo, Troels F Rønnow, Sergei V Isakov, Zhihui Wang, David Wecker, Daniel A Lidar, John M Martinis, and Matthias Troyer. 2014. Evidence for quantum annealing with more than one hundred qubits. *Nature Physics* 10, 3 (2014), 218.

[6] Endre Boros, Peter L Hammer, and Gabriel Tavares. 2007. Local search heuristics for quadratic unconstrained binary optimization (QUBO). *Journal of Heuristics* 13, 2 (2007), 99–132.

[7] Panagiotis Botsinis, Soon Xin Ng, and Lajos Hanzo. 2013. Quantum search algorithms, quantum wireless, and a low-complexity maximum likelihood iterative quantum multi-user detector design. *IEEE access* 1 (2013), 94–122.

[8] Panagiotis Botsinis, Soon Xin Ng, and Lajos Hanzo. 2014. Fixed-complexity quantum-assisted multi-user detection for CDMA and SDMA. *IEEE Transactions on Communications* 62, 3 (2014), 990–1000.

[9] Aleksandra Checko, Henrik L Christiansen, Ying Yan, Lara Scolari, Georgios Kardaras, Michael S Berger, and Lars Dittmann. 2014. Cloud RAN for mobile networks—A technology overview. *IEEE Communications surveys & tutorials* 17, 1 (2014), 405–426.

[10] R. Courtland. 2016. Transistors could stop shrinking in 2021. *IEEE Spectrum* 53, 9 (2016), 9–11.

[11] Erik Dahlman, Stefan Parkvall, and Johan Skold. 2013. *4G: LTE/LTE-advanced for mobile broadband*. Academic press.

[12] Vasil Denchev, Sergio Boixo, Sergei Isakov, Nan Ding, Ryan Babbush, Vadim Smelyanskiy, John Martinis, and Hartmut Neven. 2016. What is the Computational Value of Finite Range Tunneling? *Physical Review X* 6 (2016), 031015. https://journals.aps.org/prx/abstract/10.1103/PhysRevX.6.031015

[13] Edward Farhi, Jeffrey Goldstone, and Sam Gutmann. 2014. A quantum approximate optimization algorithm. *arXiv preprint 1411.4028* (2014).

[14] Hongyuan Gao and Ming Diao. 2009. Quantum particle swarm optimization for MC-CDMA multiuser detection. In *2009 International Conference on Artificial Intelligence and Computational Intelligence*, Vol. 2. IEEE, 132–136.

[15] Hans-Otto Georgii. 2011. *Gibbs measures and phase transitions*. Vol. 9. Walter de Gruyter.

[16] Fred Glover, Mark Lewis, and Gary Kochenberger. 2018. Logical and inequality implications for reducing the size and difficulty of quadratic unconstrained binary optimization problems. *European Journal of Operational Research* 265, 3 (2018), 829–842.

[17] Zhan Guo and P. Nilsson. 2006. Algorithm and implementation of the K-best sphere decoding for MIMO detection. *IEEE Journal on Selected Areas in Communications* 24, 3 (March 2006), 491–503.

[18] Stuart Hadfield, Zhihui Wang, Eleanor G. Rieffel, Bryan O'Gorman, Davide Venturelli, and Rupak Biswas. 2017. Quantum Approximate Optimization with Hard and Soft Constraints. In *Workshop on Post Moores Era Supercomputing (PMES)*. https://doi.org/10.1145/3149526.3149530

[19] Mark D Hill and Michael R Marty. 2008. Amdahl's law in the multicore era. *Computer* 41, 7 (2008), 33–38.

[20] C Husmann, G Georgis, K Nikitopoulos, and K Jamieson. 2017. FlexCore: Massively Parallel and Flexible Processing for Large MIMO Access Points. In *Proc. of the USENIX NSDI Symp.*

[21] IEEE [n.d.]. IEEE Standard 802.11 Amendment 4: Enhancements for very high throughput for operation in bands below 6 GHz (802.11ac-2013).

[22] IEEE 2012. IEEE Standard 802.11: Wireless LAN medium access and physical layer specifications.

[23] Takahiro Inagaki, Yoshitaka Haribara, Koji Igarashi, Tomohiro Sonobe, Shuhei Tamate, Toshimori Honjo, Alireza Marandi, Peter L McMahon, Takeshi Umeki, Koji Enbutsu, et al. 2016. A coherent Ising machine for 2000-node optimization problems. *Science* 354, 6312 (2016), 603–606.

[24] Hiroshi Ishikawa. 2009. Higher-order clique reduction in binary graph cut. *IEEE Conference on Computer Vision and Pattern Recognition* (2009).

[25] Hiroshi Ishikawa. 2014. Higher-order clique reduction without auxiliary variables. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1362–1369.

[26] Zoe Gonzalez Izquierdo, Shon Grabbe, Stuart Hadfield, Jeffrey Marshall, Zhihui Wang, and Eleanor Rieffel. 2020. Ferromagnetically shifting the power of pausing. *arXiv preprint arXiv:2006.08526* (2020).

[27] Tadashi Kadowaki and Hidetoshi Nishimori. 1998. Quantum annealing in the transverse Ising model. *Phys. Rev.* E, 58 (1998), 5355–5363.

[28] Hamed Karimi and Gili Rosenberg. 2017. Boosting quantum annealer performance via sample persistence. *Quantum Information Processing* 16, 7 (2017), 166.

[29] Minsung Kim, Davide Venturelli, and Kyle Jamieson. 2019. Leveraging quantum annealing for large MIMO processing in centralized radio access networks. In *Proceedings of the ACM Special Interest Group on Data Communication (SIGCOMM'19)*. ACM, 241–255.

[30] James King, Sheir Yarkoni, Jack Raymond, Isil Ozfidan, Andrew D King, Mayssam Mohammadi Nevisi, Jeremy P Hilton, and Catherine C McGeoch. 2019. Quantum annealing amid local ruggedness and global frustration. *Journal of the Physical Society of Japan* 88, 6 (2019), 061007.

[31] Erik G Larsson and Joakim Jalden. 2008. Fixed-complexity soft MIMO detection via partial marginalization. *IEEE transactions on Signal Processing* 56, 8 (2008), 3397–3407.

[32] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *nature* 521, 7553 (2015), 436–444.

[33] M Lewis and F Glover. 2017. Quadratic Unconstrained Binary Optimization Problem Preprocessing: Theory and Empirical Analysis. *arXiv Preprint arXiv:1705.09844* (May 2017).

[34] Mark Lewis and Fred Glover. 2017. Quadratic unconstrained binary optimization problem preprocessing: Theory and empirical analysis. *Networks* 70, 2 (2017), 79–97.

[35] Yonghua Lin, Ling Shao, Zhenbo Zhu, Qing Wang, and Ravie K Sabhikhi. 2010. Wireless network cloud: Architecture and system requirements. *IBM Journal of Research and Development* 54, 1 (2010), 4–1.

[36] Jeffrey Marshall, Davide Venturelli, Itay Hen, and Eleanor Rieffel. 2018. The power of pausing: advancing understanding of thermalization in experimental quantum annealers. *arXiv:1810.05881* (2018).

[37] Rudolf Mathar and Michael Schmeink. 2001. Optimal base station positioning and channel assignment for 3G mobile networks by integer programming. *Annals of Operations Research* 107, 1-4 (2001), 225–236.

[38] C C McGeoch. 2014. Adiabatic quantum computation and quantum annealing: Theory and practice. *Synthesis Lectures on Quantum Computing* 5, 2 (2014), 1–93.

[39] Nicholas Metropolis and Stanislaw Ulam. 1949. The monte carlo method. *Journal of the American statistical association* 44, 247 (1949), 335–341.

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

[41] Daniele Ottaviani and Alfonso Amendola. 2018. Low rank non-negative matrix factorization with d-wave 2000q. *arXiv preprint arXiv:1808.08721* (2018).

[42] John Preskill. 2018. Quantum Computing in the NISQ era and beyond. *arXiv preprint arXiv:1801.00862* (2018).

[43] Troels F Rønnow, Zhihui Wang, Joshua Job, Sergio Boixo, Sergei V Isakov, David Wecker, John M Martinis, Daniel A Lidar, and Matthias Troyer. 2014. Defining and detecting quantum speedup. *Science* 345, 6195 (2014), 420–424.

[44] Gili Rosenberg, Mohammad Vazifeh, Brad Woods, and Eldad Haber. 2016. Building an iterative heuristic solver for a quantum annealer. *Computational Optimization and Applications* 65, 3 (2016), 845–869.

[45] M. Shafi, A. F. Molisch, P. J. Smith, T. Haustein, P. Zhu, P. De Silva, F. Tufvesson, A. Benjebbour, and G. Wunder. 2017. 5G: A Tutorial Overview of Standards, Trials, Challenges, Deployment, and Practice. *IEEE Journal on Selected Areas in Communications* 35, 6 (June 2017), 1201–1221. https://doi.org/10.1109/JSAC.2017.2692307

[46] C Shepard, H Yu, N Anand, L Li, T Marzetta, R Yang, and L Zhong. 2012. Argos: Practical Many-Antenna Base Stations. In *Proc. of the ACM MobiCom Conf.*

[47] Karthikeyan Sundaresan. 2013. Cloud-driven architectures for next generation small cell networks. In *Proceedings of the eighth ACM international workshop on Mobility in the evolving internet architecture*. ACM, 3–4.

[48] Robert H Swendsen and Jian-Sheng Wang. 1986. Replica Monte Carlo simulation of spin-glasses. *Physical review letters* 57, 21 (1986), 2607.

[49] Jason P Terry, Prosper D Akrobotu, Christian FA Negre, and Susan M Mniszewski. 2020. Quantum isomer search. *Plos one* 15, 1 (2020), e0226787.

[50] Tony T Tran, Minh Do, Eleanor G Rieffel, Jeremy Frank, Zhihui Wang, Bryan O'Gorman, Davide Venturelli, and J Christopher Beck. 2016. A hybrid quantum-classical approach to solving scheduling problems. In *Ninth annual symposium on combinatorial search*.

[51] D Tse and P Viswanath. 2005. *Fundamentals of Wireless Communication*. Cambridge Univ. Press.

[52] Davide Venturelli and Alexei Kondratyev. 2018. Reverse Quantum Annealing Approach to Portfolio Optimization Problems. *arXiv:1810.08584* (2018).

[53] Chi Wang, Huo Chen, and Edmond Jonckheere. 2016. Quantum versus simulated annealing in wireless interference network optimization. *Nature Scientific reports* 6 (2016), 25797.

[54] D Wang and R Kleinberg. 2009. Analyzing quadratic unconstrained binary optimization problems via multicommodity flows. *Discrete Applied Mathematics* 157, 18 (2009), 3746–3753.

[55] Qing Yang, Xiaoxiao Li, Hongyi Yao, Ji Fang, Kun Tan, Wenjun Hu, Jiansong Zhang, and Yongguang Zhang. 2013. BigStation: Enabling scalable real-time signal processing in large MU-MIMO systems. *ACM SIGCOMM Computer Communication Review* 43, 4 (2013), 399–410.

[56] Xiang Zeng, Rajive Bagrodia, and Mario Gerla. 1998. GloMoSim: a library for parallel simulation of large-scale wireless networks. In *Proceedings. Twelfth Workshop on Parallel and Distributed Simulation PADS'98 (Cat. No. 98TB100233)*. IEEE, 154–161.

[57] Jun Zheng and Bhaskar D Rao. 2006. LDPC-coded MIMO systems with unknown block fading channels: soft MIMO detector design, channel estimation, and code optimization. *IEEE Transactions on signal processing* 54, 4 (2006), 1504–1518.

[58] Ilia Zintchenko, Matthew B Hastings, and Matthias Troyer. 2015. From local to global ground states in Ising spin glasses. *Physical Review B* 91, 2 (2015), 024201.