

Copyright Notice

© 2023 IEEE. Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works must be obtained from the IEEE.

This material is presented to ensure timely dissemination of scholarly and technical work. Copyright and all rights therein are retained by authors or by other copyright holders. All persons copying this information are expected to adhere to the terms and constraints invoked by each author's copyright. In most cases, these works may not be reposted without the explicit permission of the copyright holder.

Institute of Communication Networks and Computer Engineering
University of Stuttgart
Pfaffenwaldring 47, D-70569 Stuttgart, Germany
Phone: ++49-711-685-68026, Fax: ++49-711-685-67983
Email: mail@ikr.uni-stuttgart.de, <http://www.ikr.uni-stuttgart.de>

Tactile Network Resource Allocation enabled by Quantum Annealing based on ILP Modeling

Arthur Witt^{*} , Christopher Körber[†] , Andreas Kirstädter[‡] and Thomas Luu[§] 

^{*},[†] *Institute of Communication Networks and Computer Engineering, University of Stuttgart, Stuttgart, Germany*

[†] *Institute of Theoretical Physics II, Ruhr-University Bochum, Bochum, Germany*

[‡] *Fraunhofer Research Institution for Energy Infrastructures and Geothermal Systems IEG, Bochum, Germany*

[§] *Institute for Advanced Simulation (IAS-4) & JARA HPC, Forschungszentrum Jülich, Jülich, Germany*

{arthur.witt, andreas.kirstaedter}@ikr.uni-stuttgart.de, christopher.koerber@ieg.fraunhofer.de, t.luu@fz-juelich.de

Abstract—Agile networks with fast adaptation and reconfiguration capabilities are required for on-demand provisioning of various network services.

We propose a new methodical framework for short-time network optimization based on quantum computing (QC) and integer linear program (ILP) models, which has the potential of realizing a real-time network automation. We define methods to map a nearly real-world ILP model for resource provisioning to a quadratic unconstrained binary optimization (QUBO) problem, which is solvable on quantum annealer (QA).

We concentrate on the three-node network to evaluate our approach and its obtainable quality of solution using the state-of-the-art quantum annealer D-Wave Advantage™ 5.2/5.3. By studying the annealing process, we find annealing configuration parameters that obtain feasible solutions close to the reference solution generated by the classical ILP-solver CPLEX.

Further, we studied the scaling of the network problem and provide estimations on quantum annealer’s hardware requirements to enable a proper QUBO problem embedding of larger networks. We achieved the QUBO embedding of networks with up to 6 nodes on the D-Wave Advantage™. According to our estimates a real-sized network with 12 to 16 nodes require a QA hardware with at least 50000 qubits or more.

Index Terms—integer linear program, network automation, optical networks, quantum annealing, quantum computing, resource allocation

I. INTRODUCTION

A. Motivation

Optical wide-area networks are the backbone for public communication systems like 5G/6G mobile communication and different variants of fixed-access networks. The transport of internet protocol (IP) traffic requires a conversion from electrical to optical signals and vice versa. It is realized by power-hungry transceivers within the optical networks. Because traffic volume changes over time, adapting the network configuration to new demands is beneficial.

The authors gratefully acknowledge the Jülich Supercomputing Centre for funding this project by providing computing time through the Jülich UNified Infrastructure for Quantum computing (JUNIQU) on the D-Wave Advantage™ quantum system.

This work has been performed in the framework of the CELTIC-NEXT EUREKA project AI-NET ANTILLAS (Project ID C2019/3-3), and it is partly funded by the German Federal Ministry of Education and Research (Project ID 16 KIS 1312).

The authors thank C. C. Chang and D. Willsch for fruitful discussions.

The authors alone are responsible for the content of the paper.

Traffic temporal changes occur regularly on a diurnal scale [1], [2] but also have fluctuations at smaller timescales of seconds and less [3], [4]. Therefore, an economically-friendly adaptive operation of networks requires fast control algorithms for dynamic resource allocation, traffic engineering and restoration.

The underlying problem of network resource adaptation is a constrained optimization problem of integer variables. This kind of problem can be formulated as an integer linear program (ILP). Solving ILPs for network resource allocation can be time consuming, especially if a detailed modeling of the network architecture and operation strategies is required, which is usually the case. Therefore, an ILP-based network adaptation in short time periods is difficult to achieve—in [5] solutions are obtained within several hours. In a real network, problems like these are often solved with heuristic or meta-heuristic approaches which allow for a fast solvability but at the expense of modeling accuracy. These methods also require substantial effort to tailor the design of the heuristic to the specific problem at hand. Those comparatively fast heuristics require still several minutes to solve such problems [6], [7].

Quantum hardware that leverages the superposition of quantum bits enables massively parallel optimization protocols and could possibly overcome the time constraints for short time optimizations without the need for reduced model accuracy. For example, refs. [8] and [9] showed that classical solution approaches can be outperformed by quantum computing. A significant optimizer speedup on the order of magnitudes is, in principle, achievable. Only through recent technological progress has it become possible to map non-trivial optimization problems on quantum hardware. In particular, our work proposes a new ILP-based solution approach for the aforementioned network problem that can be solved on a quantum annealer (QA), a form of quantum computing architecture. While solutions obtained on a quantum hardware are not generally guaranteed to be optimal, this problem is an ideal use case for such algorithms as it is only required to obtain better solutions—which can be verified within microseconds.

B. Objectives

We test the feasibility of solving network optimization problems on quantum hardware. Specifically, we test the feasibility

to extract solutions of a small but non-trivial discrete optimization problem within seconds. Because today's networks are controlled centrally via software-defined networking (SDN), it is technically possible to incorporate this quantum framework in realistic network configuration setups for traffic engineering and restoration.

We use the new D-Wave Advantage™ 5.2/5.3 quantum annealer in Jülich for the evaluation of our network optimizing algorithm. We illustrate the essential steps in the general application of the QA to the network problem, and discuss the findings, discovered challenges, and restrictions of our approach for a three-node network. Additionally we provide a network scaling study which allows us to make quantitative statements related to the feasibility of our approach and QA-related requirements for solving real-world network problems.

C. Organization of the Work

Our work covers the interdisciplinary aspects of quantum-aided optimization and network resource allocation as problem. The allocation problem is introduced as a nearly real-world application of quantum computing. In particular, we specify the considered network architecture in Sec. II and formulate the resource allocation problem of wide-area networks. In Sec. III, we introduce our proposed algorithmic methods to solve a typical network resource allocation problem with quantum annealing. Therefore, we model the resource allocation problem as an ILP and explicitly derive the formulation of this ILP in terms of a QUBO representation required for the optimization procedure on the QA. In Sec. IV we evaluate our solution method according to a defined procedure and interpret the results that we obtained for a three-node network. We also discuss the stochastic effects of the quantum annealing process. In Sec. V we introduce the embedding of QUBO matrices on the D-Wave QA hardware and study the scalability of our approach with respect to increased network sizes and hardware limitations. This allows us to estimate the QA hardware requirements for real-sized networks. Finally, we recapitulate in Sec. VI. Additionally, we provide supplemental material in the appendices on the idea of quantum annealing (Appx. A) and a summary on the general approach of ILP to QUBO mapping (Appx. B).

II. RESOURCE ALLOCATION IN WIDE AREA NETWORKS AS COMBINATORIAL OPTIMIZATION PROBLEM

A. Network Architecture

An optical wide-area network is structured like a graph with a meshed topology consisting of network nodes that are linked by optical fiber systems on the graph's edges. An optical fiber system itself, denoted in short as a fiber link, is a sequence of fiber spans (each with 80km length) with amplifiers in between. Edges within a wide-area network allow transmissions in both directions. They are typically realized by separated fibers to avoid signal distortions due to crosstalk and other effects [10].

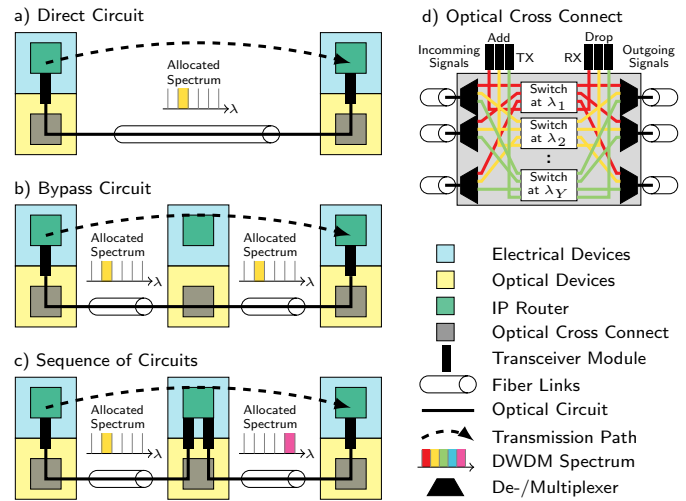


Fig. 1. a–c) Various ways of realizing transmission paths in wide-area networks with optical dense wavelength division multiplexing (DWDM) layer and d) architecture of a optical cross connects (OXC).

Individual signals, propagating in the same direction, can be transported simultaneously on a fiber using dense wavelength division multiplexing (DWDM). The fiber bandwidth is divided into a finite number of channels each with 50GHz bandwidth. An optical channel is specified by its central wavelength.

Network nodes V are composed of electrical IP routers in an upper layer and optical equipment, like optical cross connects (OXC), in a lower layer. Transition between these layers, i.e., the electrical-to-optical signal transformation and the other way round, are enabled by power-hungry transceivers. They allow transmission (TX) and reception (RX) via time-division multiplexing, but are used mostly unidirectionally. An optical signal generated by a transceiver utilizes a spectral bandwidth dictated by the granularity of that single optical channel. Within the OXC, optical signals between transceivers and connected fiber links are switched in a wavelength-selective way. In particular, redirected signals can be forwarded only at the same wavelength. Signals at the fiber link side are de-/multiplexed to apply DWDM signals on fiber links.

The transportation of data requires the realization of a transmission path. A transmission path is a list of traversed nodes which includes the source and target node. It is realized by a single circuit or a sequence of optical circuits. Optical circuits are terminated by transceivers and utilize the spectrum of a single optical channel. A *direct* optical circuit traverses a single fiber link whereas a *bypass* circuit traverses at least two fiber links (wavelengths are preserved). If transmission paths are realized by a sequence of optical circuits by a sequence of optical circuits, the allocated wavelength of subsequent circuits may differ. These variants of transmission path realizations and the functionality of OXCs are depicted in Fig. 1.

We will consider a simplified network architecture that is restricted as follows:

- **Single Rate System:** All optical circuits operate at a

single data rate of 100 Gbit/s with a maximal optical reach of 1000 km. These are typical values in wide-area networks and depend on the type of equipped transceivers.

- **Colorless Wavelength Assignment:** Fibers are typically over-provisioned during the construction phase of a network, such that spectral resources at fiber links are not limited. Therefore, we will ignore the wavelength assignment as it can be optimized in an additional step by a graph coloring algorithm.

Beside these considerations, multiple optical circuits can be realized along the same circuit path c (sequence of traversed nodes). As we ignore the wavelength assignment, we can summarize the specification of those circuits by its unique circuit path c and a circuit counter w_c , which will be used to determine the required capacity for a traffic flow or a migration of traffic flows along a circuit path.

B. Connectivity as a Network Service

Traffic flows from/towards connected data centers and other interconnecting networks ingress and egress the wide-area network at one of the IP routers. A traffic ingress/egress on the optical layer is only possible with additional considerations that are not covered here. Incoming traffic flows are combined by a migration process and considered as single flow if they have to traverse the same source node u and target node v for $u, v \in V$. So, we can define a unidirectional traffic demand d for each disjunct node pair (u, v) with $u \neq v$. The data volume for transport between those nodes—the demand value h_d —is expressed as a data rate. Demands are summarized in the set D and can be interpreted as a request of connectivity with a specific data rate.

Typically the demands will vary over time, such that a network is adapted by a frequent or on-demand traffic engineering to provide the required capacity and connectivity. This process is an essential part of the network operation and considers:

- **Traffic Routing:** A transmission path (i.e., sequence of traversed nodes) has to be selected for realizing a traffic flow between two nodes. Usually, the shortest path is selected as, in most cases, it provides the shortest latency. However, to potentially allocate network resources more efficiently, we consider alternative paths — which increases the amount of combinations for possible traffic flow migrations.
- **Resource Allocation:** Establishing a transmission path requires the activation of optical circuits, see Fig. 1. This activation includes the allocation of optical channels per traversed fiber link and transceivers. The (expensive) transceivers are node-wise limited and can be used for transmission or reception. The considerations for spectrum allocation are already described in the context of the colorless wavelength assignment.

C. Definition of Possible Transmission Path Realizations

Initially, we generate a set of circuit paths C that includes all possible direct circuit paths that traverse a single network edge and a finite number of possible alternatives in the form

of bypass circuit paths. We must therefore consider the optical reach of a circuit and the restrictions of the topology.

Based on a shortest path search, we can identify a set of one or more loop-free transmission paths L_d per demand d , that are listed by their distance in ascending order. It can be beneficial if some of the transmission paths are realized by a sequence of rather short circuit paths. If this is the case, it increases the flexibility to carry a combination of different traffic flows along a circuit path c by data migration. Therefore, we compute the set $R_{d,l}$ of possible circuit path patterns $r \in R_{d,l}$, that realize transmission path $l \in L_d$. The pattern r consists of a sequence of circuit paths or a single circuit path c , predefined in C . The first circuit path pattern in $R_{d,l}$ is composed of the shortest possible circuit paths, i.e. they traverse only a single edge of the network. The second pattern uses a minimal amount of circuit paths which incorporate multiple edges and therefore increased distances. Further patterns have mixed circuit path lengths, and so on. For simplicity we will define the union set T_d , where $t_d \in T_d$ are all circuit path patterns that can realize a transmission path for demand d , and the total union set T according to

$$T_d = \bigcup_{l \in L_d} R_{d,l} \quad \text{and} \quad T = \bigcup_{d \in D} T_d. \quad (1)$$

Circuit paths of C , that are not used in T are deleted in C .

The economic resource allocation within wavelength division multiplexing (WDM) networks, in this context, forms a combinatorial optimization problem with a non-polynomial (NP)-complete complexity.

III. ALGORITHMIC SOLUTION APPROACH

A. Framework

Optimizing the resource allocation problem of wide-area networks involves multiple interacting stages: a) the modeling of the network as an ILP problem, b) finding a representation of this problem in QA-compatible fashion, c) finding possible embeddings of the problem on the QA, d) evaluating the output of the QA, and e) performing a post-processing analysis to find feasible solutions that optimize the network. Some of these steps are time intensive but only need to be executed once, other steps need to be performed multiple times but can be executed within microseconds. Fig. 2 summarizes the concept of our approach for solving network resource allocation problems by quantum annealing.

Our solution approach relies on a method for solving integer linear programs (ILPs) on quantum annealers that was published firstly in [11]. The authors propose a strategy for mapping an integer linear program (ILP) to a quadratic unconstrained binary optimization (QUBO) problem and used it to solve a binary linear problem. In this work, we incorporate a mostly real-world network resource allocation problem. This allows us to study this mapping process and estimate the feasibility of incorporating this formalism for a real world use case. We also give an extended form of the generic matrix-wise definition of the QUBO representing non-binary ILP target variables.

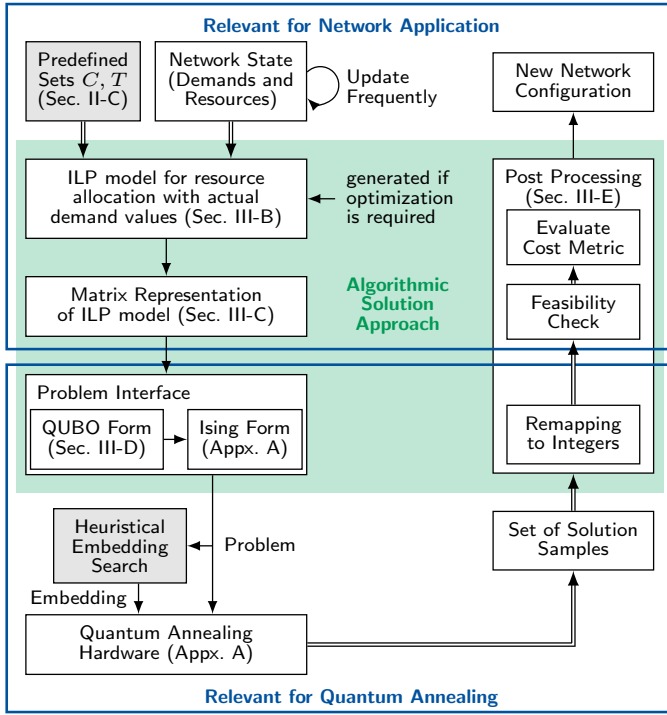


Fig. 2. Proposed framework for network resource allocation enabled by ILP modeling and quantum annealing at a glance. Gray boxes represent work steps which need to be executed once in a setup phase. White boxes form the control loop of regularly optimization, which is triggered frequently or on-demand.

B. Integer Linear Program for Network Optimization

Integer linear programs (ILPs) for resource allocation and service provisioning in wide area networks are often used as a reference method providing exact solutions for comparison with newly developed heuristic or meta-heuristic algorithms. Sometimes they are also used to study the possible benefit of new network operation strategies. Some examples are given in [12], [13] where the reduction of over-provisioned quality of services in networks was studied.

To study the applicability of solving a network resource allocation problem on quantum annealers we define the following terms based on [12]. Though the traffic volume h_d of a demand d varies over time, for the ILP it is held constant. The variations over time can be considered by frequently solving the ILP with updated values of h_d .

We now enumerate the parameters and constraints of our ILP.

Variables:

- $g_{t_d} \in \{0,1\}$: path selector, i.e., g_{t_d} equals 1 if a transmission path for demand d is realized by circuit configuration $t_d \in T_d$.
- $w_c \in \mathbb{N}$: the number of active, parallel circuits on circuit path c .

Constants:

- $\xi \in \mathbb{R}$: the data rate of a single optical circuit.
- $\eta_v \in \mathbb{N}$: the amount of transceivers installed at node v .

- $\rho_{c,t_d} \in \{0,1\}$: indicates whether circuit configuration t_d uses circuit path c .
- $\varphi_{v,c} \in \{0,1\}$: indicates whether node v is the source or target node of circuit path c .
- $h_d \in \mathbb{R}$: traffic volume of demand d .

Constraints:

$$\sum_{t_d \in T_d} g_{t_d} = 1 \quad \forall d \in D \quad (2)$$

$$-w_c + \sum_{d \in D} \sum_{t_d \in T_d} \rho_{c,t_d} \cdot \frac{h_d}{\xi} \cdot g_{t_d} \leq 0 \quad \forall c \in C \quad (3)$$

$$\sum_{c \in C} w_c \cdot \varphi_{v,c} \leq \eta_v \quad \forall v \in V \quad (4)$$

Objective:

$$\sum_{c \in C} w_c \rightarrow \min. \quad (5)$$

Equation (2) ensures that a demand is routed on exactly one path. The constraint (3) ensures that enough circuits are activated depending on the chosen paths. Constraint (4) activates a sufficient amount of transceivers to accommodate the active optical circuits. Finally, the objective (5) minimizes the number of active circuits and therefore also the number of active transceivers.

C. Network-related ILP Model in Matrix Form

An intermediate step of the ILP to QUBO mapping described in [11] is the reformulation of the ILP in matrix form. In general, an ILP can be defined by an objective function that incorporates a set of K constraints that are defined with M integer variables $x_m \geq 0$. We collectively describe these integer variables in vector form, $\mathbf{x} \in \mathbb{N}^M$. The objective function defines our optimization target, which in this case involves a minimization,

$$\mathbf{c}^\top \mathbf{x} \rightarrow \min, \quad (6)$$

where $\mathbf{c}^\top \in \mathbb{R}^M$ represents our cost weights. The constraints are given in the form of equalities or inequalities according to

$$\mathbf{A}\mathbf{x} + \mathbf{b} \leq 0 \quad (7)$$

with constant values $\mathbf{b} \in \mathbb{R}^K$ and the matrix \mathbf{A} with shape $K \times M$ that contains the linear weights $A_{k,m} \in \mathbb{R}$ for all constraints of the ILP. Typically, when solving ILPs, the inequalities in (7) are transformed into equalities by introducing integer slack variables $\mathbf{s} \in \mathbb{N}^K$ such that

$$\mathbf{A}\mathbf{x} + \mathbf{b} \leq 0 \Leftrightarrow \mathbf{A}\mathbf{x} + \mathbf{b} + \mathbf{s} = \mathbf{0}. \quad (8)$$

Thus, the network allocation ILP given in Sec. III-B is equivalent to

$$\mathbf{A} = \begin{bmatrix} \mathbf{G}^{|D| \times |T|} & \mathbf{0}^{|D| \times |C|} \\ \mathbf{H}^{|C| \times |T|} & -\mathbb{I}^{|C| \times |C|} \\ \mathbf{0}^{|V| \times |T|} & \boldsymbol{\varphi}^{|V| \times |C|} \end{bmatrix}, \mathbf{b} = - \begin{bmatrix} \mathbf{1}^{|D|} \\ \mathbf{0}^{|C|} \\ \boldsymbol{\eta}^{|V|} \end{bmatrix}, \mathbf{s} = \begin{bmatrix} \mathbf{0}^{|D|} \\ \mathbf{s}_c^{|C|} \\ \mathbf{s}_v^{|V|} \end{bmatrix},$$

$$\mathbf{x} = \begin{bmatrix} \mathbf{g}^{|T|} \\ \boldsymbol{\omega}^{|C|} \end{bmatrix}, \mathbf{c}^\top = [\mathbf{0}^{|T|} \quad \mathbf{1}^{|C|}]. \quad (9)$$

The term $|\cdot|$ in the superscript of the matrices refers to the size of the used sets C, D, T and V and thus defines the matrices' dimensions. The expression $\mathbf{0}$ refers to a vector or matrix that contains only zero values, while the entries of $\mathbf{1}$ are all one. \mathbb{I} is the identity matrix. Note that the column vector \mathbf{x} contains the ILP's variables $\mathbf{g} = [g_t]_{t \in T}$, $g_t \in \{0, 1\}$ and $\boldsymbol{\omega} = [\omega_c]_{c \in C}$, $\omega_c \in \{1, 2, \dots, \omega_{\max}\}$.

The rows of \mathbf{A} , \mathbf{b} and \mathbf{s} are ordered according to the constraints of the ILP. The first row with matrix

$$\mathbf{G} = \begin{bmatrix} \mathbf{1}^{1 \times |T_1|} & \mathbf{0}^{1 \times |T_2|} & \dots & \mathbf{0}^{1 \times |T_{|D|}|} \\ \mathbf{0}^{1 \times |T_1|} & \mathbf{1}^{1 \times |T_2|} & \dots & \mathbf{0}^{1 \times |T_{|D|}|} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}^{1 \times |T_1|} & \mathbf{0}^{1 \times |T_2|} & \dots & \mathbf{1}^{1 \times |T_{|D|}|} \end{bmatrix} \quad (10)$$

describes the possible path selection according to (2). The second row describes how the traffic volume h_d is distributed over the set of circuits C , cf. (3). As $h_d \in \mathbb{R}$ does not necessarily take on discrete values, we introduce a discretized version \bar{h}_d for our ILP by designating a desired number of representative digits $a \in \mathbb{N}$ in conjunction with the normalization to the single circuit capacity ξ ,

$$\bar{h}_d = \left\lceil \frac{h_d * 2^a}{\xi} \right\rceil \cdot \frac{1}{2^a}. \quad (11)$$

Here the expression $\lceil \cdot \rceil$ indicates a rounding to the next largest integer value. Together with the binary matrix $\boldsymbol{\rho}_d = [\rho_{c,t_d}]_{c \in C, t_d \in T_d}$ that describes the existence of a circuit c inside a circuit configuration t_d for demand d , the matrix \mathbf{H} is given as

$$\mathbf{H} = [\bar{h}_1 \boldsymbol{\rho}_1 \quad \bar{h}_2 \boldsymbol{\rho}_2 \quad \dots \quad \bar{h}_{|D|} \boldsymbol{\rho}_{|D|}] . \quad (12)$$

The last rows of \mathbf{A} , \mathbf{b} and \mathbf{s} incorporate the limited amount of installed transceivers $\boldsymbol{\eta} = [\eta_v]_{v \in V}$, $\eta_v \in \{0, 1, \dots, \eta_{\max}\}$ per node v , cf. (4). As such, $\boldsymbol{\varphi} = [\varphi_{v,c}]_{v \in V, c \in C}$ describes whether a transceiver at node v is connected to a circuit c binary.

The slack vector \mathbf{s} has an all-zero block entry in the first block since (2) is already an equality constraint. The remaining blocks $\mathbf{s}_c = [s_c]_{c \in C}$, $s_c \in \{0, 1, \dots, a\}$ and $\mathbf{s}_v = [s_\eta]_{v \in V}$, $s_v \in \{0, 1, \dots, \eta_{\max}\}$ are variables with a limited integer space.

D. Programming the Resource Allocation Problem on the D-Wave Quantum Annealer

The programming of D-Wave's quantum annealer requires our ILP problem to be cast in quadratic unconstrained binary optimization (QUBO) form

$$X^2(\mathbf{q}) = \mathbf{q}^\top \mathbf{Q} \mathbf{q} \quad (13)$$

for bit vectors $\mathbf{q} \in \{0, 1\}^N$. In general, given a QUBO matrix \mathbf{Q} , the embedding of the matrix \mathbf{Q} , i.e. the mapping of the problem matrix to the annealer qubit topology, is also required. Such embeddings are generated heuristically by the D-Wave API. Both \mathbf{Q} and its proper embedding are then submitted to the quantum annealer for optimization, which consists of finding the optimal bit vector \mathbf{q} which minimizes the objective

function X^2 . Further details on how the D-Wave quantum annealer solves the QUBO problem are given in Appx. A.

Our network problem can be expressed in QUBO form using the matrices \mathbf{A} , \mathbf{b} and \mathbf{s} we introduced earlier [11]. We have

$$X^2(\mathbf{q}) = \mathbf{q}^\top \mathbf{Q} \mathbf{q} + C \quad (14)$$

with

$$\begin{aligned} \mathbf{Q} &= p \begin{bmatrix} \mathbf{Q}_{xx} & \mathbf{Q}_{xs} \\ \mathbf{Q}_{sx} & \mathbf{Q}_{ss} \end{bmatrix}, \quad C = p \|\mathbf{b}\|^2 \\ \mathbf{Q}_{xx} &= \mathbf{Z}_x^\top \mathbf{A}^\top \mathbf{A} \mathbf{Z}_x + \text{diag} \left\{ \left(2\mathbf{b}^\top \mathbf{A} + \frac{1}{p} \mathbf{c}^\top \right) \mathbf{Z}_x \right\} \\ \mathbf{Q}_{xs} &= \mathbf{Q}_{sx}^\top = \mathbf{Z}_x^\top \mathbf{A}^\top \mathbf{Z}_s \\ \mathbf{Q}_{ss} &= \mathbf{Z}_s^\top \mathbf{Z}_s + 2\text{diag} \left\{ \mathbf{Z}_s^\top \mathbf{b} \right\}. \end{aligned} \quad (15)$$

The matrices \mathbf{Z}_x and \mathbf{Z}_s are defined for mapping between integer and binary variables

$$\mathbf{x} = \mathbf{Z}_x \mathbf{q}_x, \quad \mathbf{s} = \mathbf{Z}_s \mathbf{q}_s. \quad (16)$$

Since a positive integer variable x_i with an upper limit of $x_{i,\max}$ can be expressed by $R_i = \lceil \log_2(x_{i,\max}) \rceil$ binary values, we can express z_i as

$$\mathbf{z}_i = [2^{R_i-1} \quad 2^{R_i-2} \quad \dots \quad 2^1 \quad 2^0]. \quad (17)$$

The mapping matrix for the integer vector \mathbf{x} with length N_x can then be defined by

$$\mathbf{Z}_x = \begin{bmatrix} z_1 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & z_2 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \dots & \mathbf{0} & z_{N_x} \end{bmatrix}, \quad (18)$$

and a similar construction can be done for \mathbf{s} . The binary variable vectors \mathbf{q} themselves can be decomposed as

$$\mathbf{q} = \begin{bmatrix} \mathbf{q}_x \\ \mathbf{q}_s \end{bmatrix}. \quad (19)$$

Finally, we have introduced a penalty factor p that weights the relation between the ILP's objective function $\mathbf{c}^\top \mathbf{x}$ and the error metric $\|\mathbf{A} \mathbf{x} + \mathbf{b} + \mathbf{s}\|$. In general the penalty factor rates the fulfillment of the ILP's constraints during optimization. Too small a value of p results in more QA solutions that do not satisfy the imposed constraints, whereas too large a value reduces the quality of solutions. Our experience has been that the penalty parameter p should be chosen in the range $1 \leq p \leq 10$.

The exact global minimization of (14) is equivalent to solving our ILP problem. A more formal derivation and discussion of this transformation into QUBO form is given in Appx. B.

E. Post Processing

The optimization of a single QUBO matrix \mathbf{Q} by quantum annealing returns a set with a large number (hundreds or thousands) of solution samples in a second. The post processing starts with the remapping of binary solution samples to integers. Then, the solutions are rated according to the obtained

cost value and their feasibility. Based on the definitions from Sec. III-B and Sec. III-C, the solution’s feasibility is given if variable vectors g and ω fulfill the conditions

$$Gg \stackrel{!}{=} 1, \quad Hg - \omega \stackrel{!}{\leq} 0, \quad \varphi\omega - \eta \stackrel{!}{\leq} 0. \quad (20)$$

Finally, a feasible solution sample with the lowest cost value of a solution sample set will be used as the optimized network configuration.

IV. EVALUATION OF THE SOLUTION APPROACH

A. Procedure of Evaluation

We evaluate our solution approach described in Sec. III as follows:

- 1) For a three-node network with normal-distributed demand values $h_d \in \mathcal{N}(75, 10)$ in Gbit/s, we generate QUBOs with different penalty terms $p \in \{1, 2, 4, 8, 16, 1000\}$ and number of slack accuracy digits $a \in \{1, 2, 3, 4, 5\}$. Both parameters can be seen as a knob to tune the numerical accuracy and will have an effect on the QUBO’s entries. Parameter a will also affect the shape and structure of the QUBO.
- 2) QUBOs with the same penalty, shape and structure are categorized within the same QUBO class.
- 3) We generate a few different embeddings per QUBO class with the heuristic approach *minorminer* provided by the D-Wave’s SDK. In principle, an embedding can be used for different QUBO matrices of the same class.
- 4) We submit the QUBO and its proper embedding with varying annealing parameters like chain strength, annealing schedules, and others. The annealing schedules summarizing the annealing duration, annealing pause duration and annealing fraction of pausing where the annealing pause is applied. We vary the annealing duration per sample within $t_{ps} \in \{1, 50, 100, 500, 1000\}$ μ s and the annealing fraction of pausing between $s_p = 0.35$ and 0.5. The pause duration of $t_p = 20$ μ s allows a longer interaction duration for the transverse Ising process (see. Appx. A). We denote the annealing schedule in short by $t_{ps}@s_p + t_p$.
- 5) For post-processing, obtained binary solutions are translated to integer values. They are classified as feasible if they fulfill the ILP’s constraints. We further evaluate the QUBO’s energy (value of $q^T Qq + C$), the ILP’s objective function (value of $c^T x$) and the number of occurrences per time to solution.

In total, we gathered 5.1 million solution samples distributed over 3300 submissions for 40 different embeddings, resulting in over 2000 distinct parameter combinations. We observed that the parameters of the annealing schedule and penalty have a larger impact on solvability. Table I provides the number of solution samples for a selection of parameter configurations. The cells highlighted with a green color are parameter combinations which returned feasible solutions.

TABLE I
NUMBER OF 1000 SAMPLES PER PENALTY TERM AND ANNEALING SCHEDULE. GREEN ENTRIES FOUND FEASIBLE SOLUTIONS.

Annealing Time or Annealing Schedule	QUBO Penalty p					
	1	2	4	8	16	1000
1 μ s	600		170			48
50 μ s			117			
100 μ s			99			
100 μ s@ 0.35 + 20 μ s	50	50	50	50	50	
100 μ s@ 0.50 + 20 μ s	50	50	60	50	50	
500 μ s			30			
1000 μ s			17			

The complete set of used run parameters, specifications and obtained solutions are logged to a database and is openly accessible¹. The database is realized with EspressoDB [14].

B. Interpretation of the Obtained Solutions

The aforementioned evaluation procedure is a baseline to study the solvability and dependencies on annealing parameters, numerical accuracy and various embeddings of the considered network problem for a three-node network. Fig. 3 shows the energy distribution of the QUBO’s objective function according to (14). It displays, on the one hand, solution vectors obtained by the annealing process for various parameter settings (orange). On the other hand, it displays solutions obtained by a (uniform) random bit guessing approach (blue). We observe that quantum annealing provides solution vectors with significantly lower energy values that follow a different distribution.

A more detailed view is seen by analyzing the feasibility of solution samples, determined by calculating (20) post anneal, and evaluating the cost value $c^T x$. Fig. 4 shows the amount (white numbers) of obtained feasible solutions per integer cost values (x-axis). **The best solution we have found, corresponding to a cost value of 8, is one cost step apart from the best reference solution with a cost of 7.** The results can be further grouped according to the applied annealing times. Here we see that results within the category of a 100 μ s annealing duration are better than for the 1 μ s case. **Overall, feasible solutions are very rare, at the order of 13 per million if compared with all obtained samples.** This obtained ratio determines the values of the y-axis of Fig. 4. At this point it is not possible for us to determine if the choice of the penalty term, the annealing schedule, other parameters, or their interplay are responsible for the quality and frequency of the optimal solution within a suitable region. We are actively investigating this point.

C. Sources of Errors and Statistics

In theory, optimal solutions of the QUBO problem are automatically optimal solutions to the network ILP problem. However, solutions returned by the QA may be non-optimal because of errors during the annealing process. For example, depending on the specification of the annealing schedule, non-adiabatic transitions from the ground state to an excited state

¹Open Data Access: <https://jugit.fz-juelich.de/qnet-public/home/>

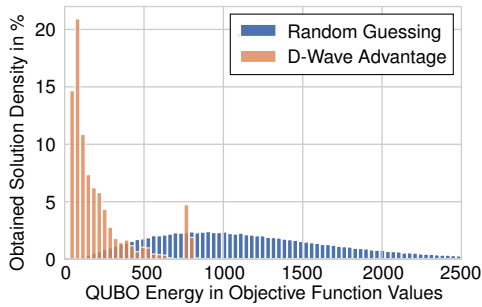


Fig. 3. Solution density with respect to QUBO objective function values according to (14) for a three-node network. Density values correspond to all obtained D-Wave samples independent of the parameter setup and feasibility of the solution (orange) and to randomly sampled integers (blue) for 10^6 samples. The density is defined as the number of samples within a bin divided by the number of all obtained samples; for each category respectively.

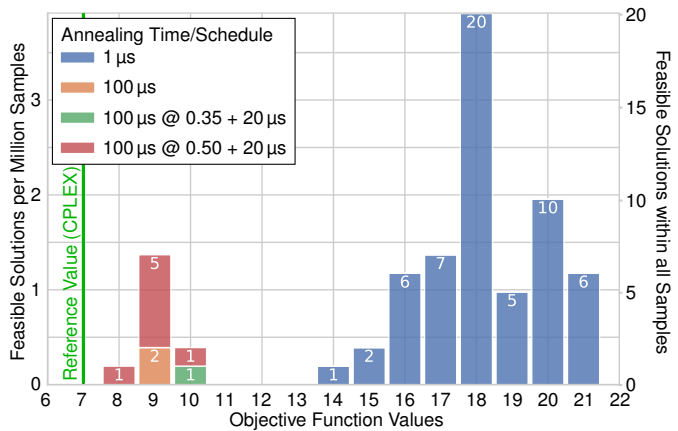


Fig. 4. Solution density with respect to the objective function values (6) of the ILP for a three-node network obtained by the quantum annealer. Solutions are classified as feasible solutions in post-processing by evaluating (20). The density is defined as the number of feasible samples for an objective value divided by the number of all obtained samples over all submits (including parameter configurations which did not find feasible solutions). The samples are categorized by different annealing schedules, specified by the annealing time per sample (without overhead), and the number within the bars is the count for a given category. The green line is a known reference solution obtained by the classical ILP-solver CPLEX.

may occur if the annealing happens too fast, and especially if the energy gaps between excited and ground states are small.

If the annealing happens too slowly, temporal or thermal decorrelation of qubits may occur so that qubits may freeze in position independent of the problem Hamiltonian. On the other hand, it may not be possible to set up the problem on the hardware—even if a valid embedding was found. For example, if qubit chains are too long, they potentially break such that one solves a different problem. Or, since the magnetic fields generated by the hardware are only manipulated with finite precision, the resolution of the QUBO entries may exceed the magnetic field precision. It is *a priori* not known if these problems may occur during an annealing schedule. For a given problem setup, one must therefore perform multiple annealing schedules and post-process the results to identify valid solutions. This procedure provides distributions of solutions with varying degrees of quality.

TABLE II
SELECTED PARAMETERS FOR VARIOUS CONSIDERED NETWORK SIZES.

Network Size (# Nodes)	$ V $	3–4	5–7	8–11	12	13–16
Installed Transceivers per Node	η_{\max}	15	31	63	63	127
Max. Capacity of Circuit Paths	ω_{\max}	3	3	3	7	7

V. SCALABILITY STUDY

A. Embedding of QUBO on the D-Wave Advantage™

The process of determining a valid mapping of the QUBO graph to the hardware graph is an essential constraint for being able to solve a problem. The possibility of finding such a mapping, also called an embedding, is thus limited by the number of logical qubits present in the formulation, i.e., the dimension of the QUBO matrix. Furthermore, even if the number of available qubits is in principle sufficient, the density of the QUBO matrix, defined as ratio of non-zero entries divided by its dimension squared, cannot exceed the total connectivity of the hardware graph. Because the hardware topology may have a different connectivity as required by the QUBO, several hardware qubits must be chained together to form a single logical qubit (nodes of the QUBO graph) — which places stronger constraints on the problem sizes.

This pre-processing procedure, e.g. finding a valid hardware embedding for the given problem, must be performed before submitting problems to the annealer. Depending on the problem size, finding a valid embedding may take several hours on a single CPU. However, since an embedding is a map from QUBO nodes to hardware topology nodes, which does not depend on the relative coupling strengths, it is possible to export valid embeddings which are reusable for a larger set of problems. Thus the expensive part of finding an embedding must be performed only once for a class of QUBO problems. Generally, embeddings which have small chains while using the entire hardware are found to be more optimal [15].

B. Scaling of Network Sizes

For studying the scalability in terms of the communication network size, a growing network topology is used as depicted in Fig. 5. The nodes of this topology are added in ascending order starting with a three-node network (N_1, N_2 and N_3) in the lower left corner. The edges are added according to the shown topology, such that added nodes are connected with the already existing network. Extending to a four-node network N_4 adds the edges N_2-N_4 and N_3-N_4 , for example. The vertical and horizontal distance between neighboring nodes is set to 300 km.

Finally, a set of topologies with $|V| \in \{3, 4, \dots, 16\}$ nodes was studied to test the feasibility of finding an embedding. The network loads are defined by a demand matrix with values h_d for traffic demands between the pairs of disjunct nodes $d \in D$. The values for h_d are sampled from a normal distribution $\mathcal{N}(\mu, \sigma)$ with $\mu = 75$ and $\sigma = 20$ Gbit/s. We used the values in Table II for the parametrization of our ILP for the QUBO mapping procedure.

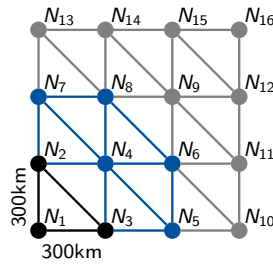


Fig. 5. Topology of a growing network. (●): smallest network with 3 nodes, (●): exemplary extension to a 8 node network, (●) maximal considered extension.

C. Hardware Limitations of the used Quantum Annealer

The optimization problem size, i.e. the dimension of the QUBO, can be expressed by the number of required logical qubits. As logical qubits are realized by a coupled chain of physical qubits, the average chain length can be seen as a metric for the efficiency of an embedding. Fig. 6 shows this metric for 3 to 6 node networks. The varied number of digits a changes the numerical accuracy of the ILP's constraint (3), and changes therewith the problem size. We observe that the average chain length follows a linear law of $2.13 \cdot |V|$. Using the default embedding algorithm (minorminer), we were only able to find embeddings for networks with six or less nodes for the D-Wave Advantage™ 5.2/5.3. Assuming the observed chain-length scaling is valid for future QA hardware, it can be used to extrapolate to larger network sizes with more than six nodes.

Fig. 7 shows the resulting relative hardware utilization of the quantum annealer for network problems with 3 to 16 nodes. The D-Wave Advantage 5.2™ provides in total 5600 physical qubits and roughly 40100 coupling elements. This marks the 100% line of hardware utilization. The amount of required logical qubits increases with the network size in a range from 66 to 3822, i.e. 1.1% to 66.4%. Further, the QUBOs of the network problems require qubit connecting couplers on the logical level in the range of 684 to 182110, i.e. these are the non-zero elements of Q . The amount of maximal available couplers is already exceeded for a network of 12 nodes. As chains of physical qubits must be built by coupling to realize a single logical qubit, the amount of physical available qubits and/or coupling element is already exceeded for networks with 7 or more nodes. The required amount of physical qubits can be obtained by multiplication of the average chain length and amount of logical qubits, e.g. the 6 node network with a 5 digit accuracy a has $532 \cdot 8.8 = 4682$ physical qubits.

We summarize that the procedure of finding an embedding for network problems on the D-Wave's quantum annealer forms a stronger constraint than the constraints imposed by the naive estimations related to the QUBO size and density. Specifically, for the largest problem we were able to embed, the amount of physical qubits needed was roughly an order of magnitude larger than the naive estimate based on the size of the QUBO. Increasing the problem accuracy increases the problem size and therefore the required quantum processing unit (QPU) hardware. This effect is not as strong as adding further nodes to a network. Thus, **the largest impact on the**

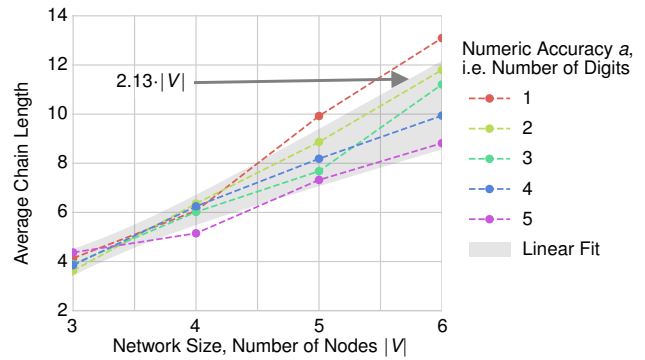


Fig. 6. Average chain length (physical qubits needed to form a logical qubit) depending on the number of nodes in the network and digits required to represent demands. The gray band represents a one-parameter fit with propagated uncertainties.

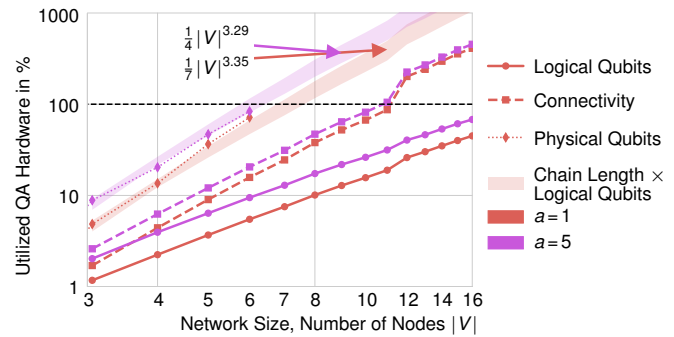


Fig. 7. Hardware utilization of problem depending on problem size. The hardware utilization is expressed by the logical qubits (filled circles) corresponding to the dimension of the QUBO (at most 5.6k qubits for a perfect embedding), the connectivity of the QUBO (squares) corresponding to non-zero entries in the QUBO (at most 40.1k), and physical qubits (diamonds) corresponding to logical qubits times the average chain length (at most 5.6k qubits). Results are displayed for the best scaling using a demand float precision of one bit (red) and the most precise scaling with five bit precision (purple).

problem complexity comes from the number of network nodes. Extrapolation of the physical qubit curve in Fig. 7 allows us to estimate how large a QPU in terms of physical qubits for real network problems should be. We find that **at least 10 times more physical qubits are required to operate a wide-area network with a reliable size of 12 to 16 nodes by quantum annealing.** An adjustment of the topology, i.e., an increase of individual qubit's connectivity would also enable access to larger networks, which is however more difficult to quantify.

VI. CONCLUSION AND OUTLOOK

We proposed an algorithmic approach relying on the cutting-edge technology of quantum computing for the optimization of resource utilization in SDN-controlled optical wide-area networks. The optimization problem is modeled by an ILP, translated in several steps to a QUBO, which can be solved on a quantum annealer like the D-Wave Advantage™ 5.2/5.3. The algorithm can be used for traffic engineering, resource allocation and restoration.

We studied the quantum-based solvability of a three node network. Compared to a random guessing method, solutions obtained by the quantum annealer show a significant lower

energy on average. Solution vectors are checked for feasibility and compared with a reference solution obtained by CPLEX (classical ILP-solver). We showed that feasible solutions with cost values close (one unit-step less optimal) to the reference solution are obtainable. At the current stage, feasible reconfigurations close to the optimal value for a three-node problem can be obtained every minute. This number is defined by the number of samples needed to find a feasible solution and the total run time per sample for an individual anneal. For a given effective anneal time $t_{\text{eff}} \sim t_{\text{ps}} + t_{\text{p}} [\mu\text{s}]$, we estimate the total run time per sample by $0.58 + 5.75 \cdot t_{\text{eff}} [\text{ms}]$, which is based on experience values. In case of $t_{\text{eff}} \sim 120 \mu\text{s}$, we require $N \sim 5 \cdot 10^4$ samples with a total run time of 1.27 ms each, such that a feasible solution can be found in 63.5 s.

Further, we discovered that a setting with penalty of 4 and annealing time of $100 \mu\text{s}$ achieved the best results. We see indications that some annealing parameter configurations have a higher chance to return feasible solutions—potentially allowing one to find feasible solutions within less than a minute of run time. While decreasing the length of an annealing schedule to a few μs in principle allows one to obtain solutions more frequently, at some point the problem setup time dominates the effective annealing time. As such, there will be an optimal value of annealing parameters which minimize the time to solution. The parameter variations for the annealing process has to be studied further as the current solution set represents only an empirical sample set which does not allow one to formulate strong statement on statistical correlations.

Our feasibility study shows that the proposed network problem for up to 6 nodes can be embedded on the D-Wave QPU. We estimate that the amount of physical qubits, assuming the same qubit connectivity, should be in the range of 50000 or above to optimize networks in reasonable sizes (12 to 16 nodes). Finally, as larger Ising model based solvers with up to 100000 qubits [16] are in the reach, embedding 12 to 16 node networks seems realistic in the near future.

For realistic networks, classical heuristics find solutions within several minutes. Compared to classical heuristics, the length of individual annealing schedules are independent of the problem size. However, the probability of finding a feasible and optimal solution will be smaller than the probability we have obtained for the three-node problem due to the increased sample space. Thus, the focus of future works should concentrate on both, finding more optimal ways to embed problems to scale to larger networks and finding optimal setup parameters to decrease the time to solution.

As quantum computing can be superior against classical computing, demonstrated in [9], a further speed-up might be achievable by newer generations of quantum computers or a problem specific QPU, such that network reconfiguration within seconds or below might be possible. It remains to be shown that obtaining feasible solutions within a reasonable time is possible for such larger networks. If successful, the proposed approach for network optimization has the potential to have a large impact on how networks are operated in the future and ultimately enable real-time network automation.

APPENDIX A

BASIC PRINCIPLES BEHIND THE USED D-WAVE MACHINE

Here we provide a short discussion on the D-Wave architecture and the basic principles underlying the quantum annealing process. For further details we refer to [17].

A. The Ising Model

Central to the annealing process is the ability to manipulate a 2-D array of 2-state spins that form the so-called Ising model, as depicted in Fig. 8.

In physics, the Hamiltonian function that describes this Ising system is given by

$$\mathcal{H}_{\text{Ising}}(\mathbf{s}) = \sum_i h_i s_i + \sum_{i>j} J_{i,j} s_i s_j, \quad (21)$$

where $s_i = \pm 1$ are spin projections in the z direction and h_i is an external magnetic field at site i . The coupling between spins at sites i and j is given by $J_{i,j}$. This expression represents the total energy of the system and can be used to derive its equations of motions. As the model represents an array of spins in 2-D, it provides a simple representation of ferromagnetism that exhibits a second order phase transition. From a QA perspective, it serves as an objective function to be minimized.

B. Quantum Annealing as Hardware Process

The quantum mechanical equivalent of a classical system can be obtained by replacing canonical coordinates and momenta with operators (canonical quantization). As a consequence, previously commuting expressions may now not be commutable. In case of the Ising model, the spin variables s_i are replaced by spin operators: the Pauli matrices $\hat{\sigma}$. The D-Wave Advantage™ system starts initially with a *transverse* Ising model. By the application of external magnetic fields, the annealer adjusts the the time-dependent amplitudes of the Hamiltonian operator in both the x , or transverse, direction and z direction, respectively,

$$\hat{\mathcal{H}}_{\text{QPU}} = \frac{A(s)}{2} \hat{\mathcal{H}}_{\text{initial}} + \frac{B(s)}{2} \hat{\mathcal{H}}_{\text{problem}} \quad (22)$$

$$\hat{\mathcal{H}}_{\text{initial}} = \sum_i \hat{\sigma}_x^{(i)} \quad (23)$$

$$\hat{\mathcal{H}}_{\text{problem}} = \sum_i h_i \hat{\sigma}_z^{(i)} + \sum_{i>j} J_{i,j} \hat{\sigma}_z^{(i)} \hat{\sigma}_z^{(j)}. \quad (24)$$

Here $\sigma_{x,z}$ represent Pauli sigma matrices. Because the spin projections in the x and z directions do not commute, this system must be solved quantum mechanically and is therefore also known as the *quantum* Ising model. The external fields depend on time. Initially $A(s) \gg B(s)$ at time $t = 0$ corresponding to an annealing fraction $s = 0$, but adiabatically changes to $B(s) \gg A(s)$ at some anneal time $t = t_f$ corresponding to an annealing fraction $s = 1$, as shown in Fig. 9. After time t_f the system has annealed to purely $\hat{\mathcal{H}}_{\text{problem}}$, and since $\hat{\sigma}_z$ has either ± 1 expectation values, this is identical to the classical Ising model in (21).

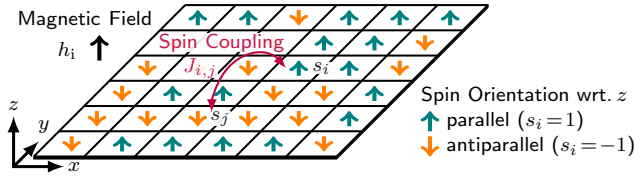


Fig. 8. Visualization of spin configuration in a 2D-lattice (flat material) which is determined by the Ising model (21).

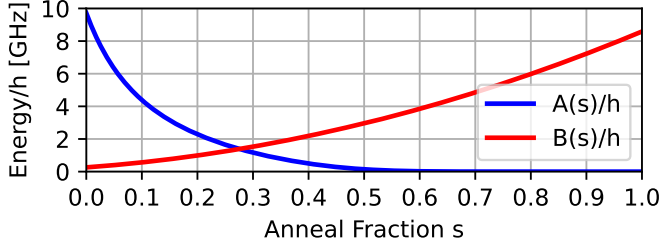


Fig. 9. Energy variation of $A(s)$ and $B(s)$ to perform the annealing within the D-Wave Advantage™ 5.2 acc to (22). Energy values are normalized by the Planck constant $h = 6.62 \cdot 10^{-34}$ Joules based on the relation $E = hf$. The anneal fraction s will be varied over time within the annealing duration to perform either a standard, paused, quenched, reverse or individual annealing schedule.

C. The QUBO and Ising Model as Interface to Program the Problem to be Solved on the D-Wave Quantum Annealer

The programming of D-Wave's quantum annealer requires the representation of the problem as a QUBO problem for bit vectors $\mathbf{q} \in \{0, 1\}^N$

$$X^2(\mathbf{q}) = \mathbf{q}^\top \mathbf{Q} \mathbf{q}. \quad (25)$$

The annealer aims at finding the optimal bit vector \mathbf{q} which minimizes the objective function X^2 .

The Ising model wave function components σ_i are related to components of the QUBO bit vectors \mathbf{q} by a linear shift $\sigma_i = 2q_i - 1$. Equating respective objective functions allows to identify the mapping from QUBO matrix to the Ising Hamiltonian in matrix form

$$\mathbf{q}^\top \mathbf{Q} \mathbf{q} = \boldsymbol{\sigma}^\top \mathbf{J} \boldsymbol{\sigma} + \mathbf{h}^\top \boldsymbol{\sigma} + g, \quad (26)$$

which is valid for

$$\begin{aligned} \mathbf{J} &= \frac{1}{4} \mathbf{Q}_{\text{TL}} \\ \mathbf{h} &= \frac{1}{2} \mathbf{q}_{\text{T}} + \frac{1}{2} \mathbf{Q}_{\text{TL}} \mathbf{1} \\ g &= \frac{1}{4} \mathbf{1}^\top \mathbf{Q}_{\text{TL}} \mathbf{1} + \frac{1}{2} \mathbf{1}^\top \mathbf{q}_{\text{T}}. \end{aligned} \quad (27)$$

These equations were obtained by expressing QUBO vectors \mathbf{q} as Ising vectors $\boldsymbol{\sigma}$, defining the traceless part of the QUBO $\mathbf{Q}_{\text{TL}} = \mathbf{Q} - \text{diag}\{\mathbf{q}_{\text{T}}\}$ with the trace vector $\mathbf{q}_{\text{T}} = \text{diag}^{-1}\{\mathbf{Q}\}$ of \mathbf{Q} (since \mathbf{J} is not allowed to have diagonal components), and using that $q_i^2 = q_i$. Here $\text{diag}\{\cdot\}$ transforms a vector into a diagonal matrix and $\text{diag}^{-1}\{\cdot\}$ extracts the diagonal part of the matrix as a vector. $\mathbf{1}$ is a vector where each component is one.

Typically, the QUBO matrix \mathbf{Q} is symmetric ($\mathbf{Q}^\top = \mathbf{Q}$) and is used only in the form $\mathbf{q}^\top \mathbf{Q} \mathbf{q}$ to determine a energy value. This allows one to transform \mathbf{Q} to a triangular matrix $\mathbf{Q}_{\text{Triang}} = \text{tril}\{\mathbf{Q}\}^\top + \text{triu}\{\mathbf{Q}\}$, such that the problem can be also given in a reduced form according to

$$X^2(\mathbf{q}) = \mathbf{q}^\top \mathbf{Q}_{\text{Triang}} \mathbf{q}. \quad (28)$$

Function $\text{tril}\{\cdot\}$ selects the triangular part below the main diagonal of a matrix, and $\text{triu}\{\cdot\}$ the upper triangular part inclusive the main diagonal.

APPENDIX B

DERIVATION OF THE ILP TO QUBO PROBLEM MAPPING

For a given ILP with introduced slack variables s , which refractors inequality constraints as equality constraints,

$$\begin{aligned} \mathbf{c}^\top \mathbf{x} &\rightarrow \min \\ \mathbf{A} \mathbf{x} + \mathbf{b} + \mathbf{s} &= \mathbf{0}, \end{aligned} \quad (29)$$

we can get rid of the constraints by further introducing a sufficiently large penalty factor $p \geq 1$. Thus, the objective function and constraints of the ILP can be combined to a quadratic optimization of objective and penalty

$$X^2(\mathbf{x}, \mathbf{s}) = \mathbf{c}^\top \mathbf{x} + p \|\mathbf{A} \mathbf{x} + \mathbf{b} + \mathbf{s}\|^2 \rightarrow \min. \quad (30)$$

Given that the penalty term p is large enough, the combined minimization of X^2 in terms of \mathbf{x}, \mathbf{s} returns the optimal vector \mathbf{x}_0 which minimizes X^2 under the constraints.

Integer variables in \mathbf{x} and \mathbf{s} have to be expressed in binary form as \mathbf{q}_x and \mathbf{q}_s according to

$$\mathbf{x} = \mathbf{Z}_x \mathbf{q}_x, \quad \mathbf{s} = \mathbf{Z}_s \mathbf{q}_s, \quad (31)$$

see also (17), (18) and (19). Thus, we define the QUBO as

$$X^2(\mathbf{q}) = \mathbf{c}^\top \mathbf{Z}_x \mathbf{q}_x + p \|\mathbf{A} \mathbf{Z}_x \mathbf{q}_x + \mathbf{b} + \mathbf{Z}_s \mathbf{q}_s\|^2 \rightarrow \min. \quad (32)$$

By applying the rule $\|\epsilon\|^2 = \epsilon^\top \epsilon$, we obtain

$$\begin{aligned} X^2(\mathbf{q}) &= \mathbf{c}^\top \mathbf{Z}_x \mathbf{q}_x + p \mathbf{q}_x^\top \|\mathbf{A} \mathbf{Z}_x\|^2 \mathbf{q}_x + p \mathbf{q}_x^\top \mathbf{Z}_x^\top \mathbf{A}^\top \mathbf{b} \\ &+ p \mathbf{b}^\top \mathbf{A} \mathbf{Z}_x \mathbf{q}_x + p \mathbf{q}_s^\top \|\mathbf{Z}_s\|^2 \mathbf{q}_s + p \mathbf{q}_s^\top \mathbf{Z}_s^\top \mathbf{b} + p \|\mathbf{b}\|^2 \\ &+ p \mathbf{b}^\top \mathbf{Z}_s \mathbf{q}_s + p \mathbf{q}_s^\top \mathbf{Z}_s^\top \mathbf{A} \mathbf{Z}_x \mathbf{q}_x + p \mathbf{q}_x^\top \mathbf{Z}_x^\top \mathbf{A}^\top \mathbf{Z}_s \mathbf{q}_s. \end{aligned}$$

Some expressions like $\mathbf{q}_x^\top \mathbf{Z}_x^\top \mathbf{A}^\top \mathbf{b}$ can be equivalently expressed by $\mathbf{q}_x^\top \text{diag}\{\mathbf{Z}_x^\top \mathbf{A}^\top \mathbf{b}\} \mathbf{q}_x$ as $\mathbf{Z}_x^\top \mathbf{A}^\top \mathbf{b}$ is a vector and the binary vectors \mathbf{q}_x and \mathbf{q}_s will only selectively combine elements of the diagonal matrix during multiplication.

$$\begin{aligned} X^2(\mathbf{q}) &= p \mathbf{q}_s^\top \mathbf{Z}_s^\top \mathbf{A} \mathbf{Z}_x \mathbf{q}_x + p \mathbf{q}_x^\top \mathbf{Z}_x^\top \mathbf{A}^\top \mathbf{Z}_s \mathbf{q}_s \\ &+ p \mathbf{q}_x^\top \left[\frac{1}{p} \text{diag}\{\mathbf{c}^\top \mathbf{Z}_x\} + \text{diag}\{2\mathbf{b}^\top \mathbf{A} \mathbf{Z}_x\} + \|\mathbf{A} \mathbf{Z}_x\|^2 \right] \mathbf{q}_x \\ &+ p \mathbf{q}_s^\top \left[\|\mathbf{Z}_s\|^2 + \text{diag}\{2\mathbf{Z}_s^\top \mathbf{b}\} \right] \mathbf{q}_s + p \|\mathbf{b}\|^2 \end{aligned}$$

Further refraction of the \mathbf{q}_x - and \mathbf{q}_s -dependent terms and their mapping to the following matrix form of the QUBO,

$$X^2(\mathbf{q}) = \mathbf{q}^\top \mathbf{Q} \mathbf{q} + C = p \begin{bmatrix} \mathbf{q}_x \\ \mathbf{q}_s \end{bmatrix}^\top \begin{bmatrix} \mathbf{Q}_{xx} & \mathbf{Q}_{xs} \\ \mathbf{Q}_{sx} & \mathbf{Q}_{ss} \end{bmatrix} \begin{bmatrix} \mathbf{q}_x \\ \mathbf{q}_s \end{bmatrix} + C,$$

leads to the ILP formulation as QUBO problem given in (15).

REFERENCES

- [1] T. Enderle, U. Bauknecht. "Modeling Dynamic Traffic Demand Behavior in Telecommunication Networks". *Proceedings of the 19th ITG-Symposium in Photonic Networks 2018*, pp. 18–25, 2018
- [2] R. Protzmann, et al. "Large-scale Modeling of Future Automotive Data Traffic towards the Edge Cloud" *Proceedings of the 20th ITG-Symposium in Photonic Networks 2019*, 2019
- [3] C. Fraleigh, F. Tobagi, and C. Diot. "Provisioning IP backbone networks to support latency sensitive traffic." *IEEE INFOCOM 2003. Twenty-second Annual Joint Conference of the IEEE Computer and Communications Societies (IEEE Cat. No.03CH37428)*. Vol. 1. 2003, pp. 375–385.
- [4] M. Alasmar, R. Clegg and N. Zakhleniuk, and G. Parisi. "Internet Traffic Volumes Are Not Gaussian—They Are Log-Normal: An 18-Year Longitudinal Study With Implications for Modelling and Prediction", *IEEE/ACM Transactions on Networking*, vol. 29, no. 3, pp. 1266–1279, June 2021. [doi:10.1109/TNET.2021.3059542], [arXiv:2007.10150].
- [5] M. Tomatore, G. Maier, and A. Pattavina. "WDM network optimization by ILP based on source formulation". *Proceedings. Twenty-First Annual Joint Conference of the IEEE Computer and Communications Societies*. vol. 3. pp. 1813–1821, 2002. 10.1109/INFCOM.2002. 1019435.
- [6] F. Feller. "An Optimization-Heuristic Approach to Dynamic Optical Bypassing". *13th ITG Fachtagung on Photonic Networks*. May 2012.
- [7] U. Bauknecht. "A genetic algorithm approach to virtual topology design for multi-layer communication networks". *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO '21)*. ACM, pp. 928–936. [doi:10.1145/3449639.3459379]
- [8] Catherine C. McGeoch. "Principles and Guidelines for Quantum Performance Analysis", *First International Workshop on Quantum Technology and Optimization Problems (QTOP) 2019, Proceedings, Lecture Notes in Computer Science*, vol. 11413, pp. 36–48, Springer, 2019. [doi:10.1007/978-3-030-14082-3_4]
- [9] F. Arute, K. Arya, R. Babbush, et al. "Quantum supremacy using a programmable superconducting processor", *Nature* 574, pp. 505–510 , 2019. [doi:10.1038/s41586-019-1666-5]
- [10] M.O. van Deventer. "Fundamentals of bidirectional transmission over a single optical fibre", PHD Thesis, Technische Universiteit Eindhoven, 1994. [doi:10.6100/IR427807]
- [11] C. C. Chang, C.-C. Chen, C. Körber, T. S. Humble, J. Ostrowski. "Integer Programming from Quantum Annealing and Open Quantum Systems". [arXiv:2009.11970]
- [12] T. Enderle, A. Witt, and F. Christou. "Delay-Differentiated Routing in Meshed Backbone Networks". *Proceedings of the 21st ITG-Symposium in Photonic Networks 2020*, pp. 20–27, Nov. 2020.
- [13] U. Bauknecht, T. Enderle, and A. Witt. "Reduction of Delay Overfulfillment in IP-over-DWDM Transport Networks". *Proceedings of the 23rd Conference on Optical Network Design and Modeling (ONDM 2019)*, LNCS 11616, pp. 528–539, Feb. 2020. [doi:10.1007/978-3-030-38085-4_45]
- [14] C. C. Chang, C. Körber, A. Walker-Loud. "EspressoDB: A scientific database for managing high-performance computing workflow". *J. Open Source Softw.* 5 (2020) 46, 2007, [arXiv:1912.03580]
- [15] D. Willsch, M. Willsch, C. D. G. Calaza, F. Jin, H. De Raedt, M. Svensson and K. Michielsen, "Benchmarking Advantage and D-Wave 2000Q quantum annealers with exact cover problems," *Quant. Inf. Comput.* 21, 2022, 141, [arXiv:2105.02208].
- [16] T. Honjo et al. "100,000-spin coherent Ising machine". *Science Advances* 7.40, 2021. [doi: 10.1126/sciadv.abh0952]
- [17] D-Wave. "QPU Solver Datasheet: Annealing Implementation and Controls". *D-Wave Online Manual*, (visited on 01/12/2022).