IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 4 February 2024, accepted 22 February 2024, date of publication 26 February 2024, date of current version 7 March 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3370475

RESEARCH ARTICLE

Inverse Reticle Optimization With Quantum Annealing and Hybrid Solvers

PO-HSUN FANG, YAN-SYUN CHEN, JHIH-SHENG WU, AND PEICHEN YU, (Member, IEEE)
Department of Photonics, College of Electrical and Computer Engineering, National Yang Ming Chiao Tung University, Hsinchu 30010, Taiwan

Corresponding authors: Peichen Yu (peichen.yu@nycu.edu.tw) and Jhih-Sheng Wu (jwu@nycu.edu.tw)

This work was supported in part by the National Science and Technology Council, Taiwan, under Grant NSTC 112-2119-M-A49-011 and Grant NSTC 111-2112-M-A49-015-MY3.

ABSTRACT Reticle optimization is a computationally demanding task in optical microlithography for advanced semiconductor fabrication. In this study, we explore the effectiveness of D-Wave's quantum annealing (QA) and hybrid steepest descent (SD) solvers in solving pixelated binary reticle optimization problems. We show that the energy derived from the objective function depends on annealing time and inter-sample correlation. Specifically, longer annealing times and reduced inter-sample correlations result in lower energy. Moreover, introducing efficient pausing strategies in forward annealing could reduce the QA runtime by approximately 100-fold while achieving similar results to long annealing times. Finally, reticles with increased variables lead to widespread irregular values in default sorted QA energies due to quantum chain breakages, which could potentially limit the probability of attaining the optimal solution. A hybrid approach that applies the classical SD algorithm to the QA results increases the probability of locating the global minimum solution and reduces runtime to about one-third compared to the classical SD solver. These findings facilitate our comprehension of quantum computing for accelerating computational lithography in semiconductor manufacturing.

INDEX TERMS Inverse lithography technology, optical proximity correction, quantum annealing, quantum computing, semiconductor.

I. INTRODUCTION

Reticle modification is a resolution enhancement technique for improving pattern fidelity in optical microlithography by altering the size and shape of the physical circuit layout on the photomask. Over the past three decades, reticle modification methods have advanced from rule-based and model-based optical proximity correction (OPC) to free-form photomasks known as inverse lithography technology (ILT) [1]. ILT is a mathematically rigorous inverse method that converts photomask generation into an optimization problem aimed at creating an on-wafer image that closely matches the original design [2], [3], [4]. As a next-generation OPC, ILT is expected to address complications in advanced-node lithography including pushing the resolution limit and increasing the process window for ArF 193 immersion and extreme ultraviolet (EUV) lithography. However, ILT faces significant computational challenges, with computation times often being tens of times longer than conventional OPC at advanced CMOS nodes. To date, the demand for computing resources can only be met by massive parallel computation [5]. Another significant challenge associated with ILT is maintaining data consistency, especially at reticle template boundaries. Although various optimization algorithms, such as pixel flipping [6], [7], level set [8], and gradient descent [9], have shown the capability of correcting the entire chip layout, these ILT solutions are still regarded as local minima. Consequently, extra attention is necessary to address stitching issues due to inconsistent correction data that can arise at template boundaries. An objective of ILT has been finding global minimal solutions that are consistent throughout. However, a pixelated binary reticle with M-by-M elements gives rise to M^2 variables and 2^{M^2} possible pixel configurations. As M grows, it soon becomes impossible to locate the

The associate editor coordinating the review of this manuscript and approving it for publication was Sawyer Duane Campbell.

global minimum using conventional methods, further complicating the computational challenges associated with ILT.

Quantum computing (QC), since its proposal in the 1980s, has promised significant speedup in solving challenging computational problems with large dimensions by leveraging quantum phenomena such as superposition and entanglement. Current quantum computers are based on two main paradigms: gate-based QC and adiabatic quantum computation (AQC), with various physical implementations including superconductors, trapped ions, cold atoms, and photons [10], [11], [12]. Quantum annealing (QA), as a relaxed approach of AQC, offers a heuristic quantum optimization algorithm to find the ground state of Ising models [13], [14], [15]. These problems can be posed in Ising form using the $\{-1, 1\}$ basis and spin variables, or as a quadratic unconstrained binary optimization (QUBO) problem using the $\{0, 1\}$ basis and binary variables. Quantum annealers, particularly D-Wave's quantum adiabatic optimizer machine using superconducting qubits, have received considerable interest due to the number of available qubits and programmability [16], [17]. Various research and industrial applications have been tested in fields including machine learning, scheduling, chemistry, pharmaceutics, etc. [18], [19], [20], [21], [22], [23], [24], [25]. Compared to classical optimization algorithms such as genetic algorithms, particle swarm optimization, and differential evolution, QA algorithms can avoid getting trapped in local minima and have the potential to find better global solutions due to the tunneling and superposition nature of qubits (Fig. 1). Moreover, QA algorithms are faster in solving Ising models and are more robust to noise and other sources of error than quantum-inspired algorithms [26]. With the number of qubits scaling up on near-term quantum devices, QA provides a practical path to harnessing quantum resources for solving complex optimization problems like ILT. It is therefore of practical interest to investigate the applicability and performance of D-Wave's QA solver in computational lithography.

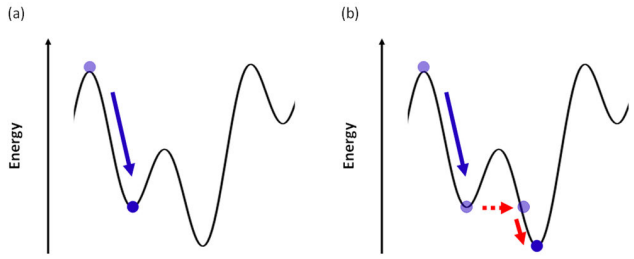

FIGURE 1. Illustration of the solution search process for classical optimization (a) and quantum annealing (b) algorithms.

In this study, we formulate a classic ILT problem, the pixelated binary mask optimization, into a QUBO model [27]. The associated Hamiltonian coefficients are embedded into the physical qubit grids on the D-Wave Advanced QA Hardware Advantage 6 system [28], [29]. We investigate the effects of annealing time, inter-sample correlation (ISC), and pausing strategy of forward annealing on the quality of the solution and the probability of locating the global minimum solution. Since current ILT problems are largely solved by gradient-based algorithms, we further compare the solutions provided by QA vs hybrid QA and classic steepest descent (SD) algorithms to gain insights into how QA might be integrated into or improve upon the industry practice. The SD algorithm iteratively minimizes the cost function by moving along the negative gradient direction to descend to a minimum. The step size is chosen to maximize descent at each iteration. Finally, we increase the reticle size and further characterize the performance of D-Wave's QA and hybrid solvers.

II. METHOD

A. QUBO FORMULATION

Optical microlithography simulations utilize Köhler's illumination model to calculate the optical intensity distribution, that is, the aerial image of a photomask on a wafer. The aerial image in a coherent imaging system can be described with the expression of Eq. (1):

$$I(x) = |E(x)|^2 = \left|\tilde{h}(x) * m(x)\right|^2, \tag{1}$$

where I and E denote the aerial image and the electric field at a specific coordinate x on the wafer, respectively. $\tilde{h}$ represents the impulse response function of the coherent imaging system and m the photomask function, with the "$*$" operator denoting the two-dimensional convolution operation. To model the exposure tool as a partially coherent imaging system with an extended illumination source, the image in Eq. (1) can be restructured to incorporate the contributions from multiple coherent sources as the intensity sum of individual optical kernels. This approximation is known as the Sum of Coherent System (SOCS) approach [1], [30], [31]. In this work, we assume coherent imaging without losing much generality, since the first kernel of a partially coherent imaging system typically holds a much higher influence on the final aerial image than the others.

The standard algorithm for ILT involves finding an optimized photomask that minimizes the criteria of a cost function. A common objective function for ILT can be defined as the squared sum of the difference between the produced aerial image and the target, as expressed in Eq. (2).

$$S = \int |I_t(x) - I(x)|^2 \, d^2x, \tag{2}$$

where $I_t(x)$ is the target aerial image inferred from the drawn layout, $I(x)$ the calculated aerial image, and S is the total cost obtained by summing the squared difference between I_t and I over the entire image plane. Therefore, we can formulate the ILT as an *argmin* optimization problem, as expressed in Eq. (3)

$$\widehat{m} = \underset{m}{\operatorname{argmin}}\, S, \tag{3}$$

The D-Wave QA architecture currently only allows for unconstrained optimization problems, where constraints need

to be incorporated into the objective function as penalty terms. For this initial study evaluating the effectiveness of D-Wave's QA algorithm on an ILT problem, we have not yet added constraints such as mask constraints or other merit terms like photoresist contour and aerial image contrast. Our objective function consists solely of the pixel-matching term. This simplified form allows us to isolate the performance of QA on the core ILT objective before introducing additional complexities.

We next formulate the ILT problem expressed in Eq. (3) into QUBO form by considering a binary photomask of N pixels with element values of either 0 or 1 as expressed in Eq. (4):

$$m(x) = \sum_{i=1}^{N} p(x - x_i)\,\sigma_i, \tag{4}$$

where p(x) denotes a 2D shape function of pixels, N is the total number of pixels, and $\sigma_i \epsilon \{0, 1\}$ represents the transmittance of i^{th} pixel located on position x_i. The electric field of the mask through a coherent imaging system is therefore given by:

$$E(x) = \sum_{i=1}^{N} \phi(x - x_i)\,\sigma_i, \tag{5}$$

where $\phi(x) \equiv \tilde{h} * p(x)$ represents a single-pixel field profile. Thus, the aerial image is the absolute square of the field profile:

$$I(x) = \left| \sum_{i=1}^{N} \phi(x - x_i)\,\sigma_i \right|^2. \tag{6}$$

When attempting to convert Eq. (2) to the QUBO Hamiltonian with Eq. (6), we encounter a fourth-power interaction among binary pixels. Following a similar approach to [26], we redefine the objective function to evaluate the amplitude profile on the wafer, as expressed in Eq. (7):

$$S_{amp} = \int |A_t(x) - A(x)|^2\, d^2x, \tag{7}$$

where $A_t(x) = \sqrt{I_t(x)}$ denotes the target amplitude image and $A(x) = |E(x)|$ calculated from Eq. (5). Equation (7) can be recast into the Hamiltonian of a QUBO problem, H_{qubo} as expressed in Eq. (8):

$$H_{qubo} = \sum_{i,j,i\neq j} J_{ij}\sigma_i\sigma_j + \sum_i h_i\sigma_i + C, \tag{8}$$

where $J_{ij} = J_{ji}$ denotes the symmetric interaction coefficient or the quadratic coefficient between pixels i^{th} and j^{th}, h_i is the linear coefficient of pixel i^{th}, and C is a constant bias. We then expand the cost function in Eq. (7) into Eq. (9):

$$S_{amp} = \int |A_t(x)|^2\, d^2x - 2\int \begin{pmatrix} Re(A(x))\,Re(A_t(x)) \\ +Im(A(x))\,Im(A_t(x)) \end{pmatrix} d^2x + \int |A(x)|^2\, d^2x, \tag{9}$$

where the first, second, and third term correspond to the constant (C), linear (h), and quadratic terms (J) in Eq. (8), respectively, as explicitly expressed in Eq. (10), (11), and (12):

$$J_{ij} = \sum_{i}^{N}\sum_{j}^{N} \begin{bmatrix} Re(\phi_k(x - x_i))\,Re(\phi_k(x - x_j)) \\ +Im(\phi_k(x - x_i))\,Im(\phi_k(x - x_j)) \end{bmatrix} \tag{10}$$

$$h_i = (-2)\sum_{i}^{N} \begin{bmatrix} Re(\phi_k(x - x_i))\,Re(A_t(x)) \\ +Im(\phi_k(x - x_i))\,Im(A_t(x)) \end{bmatrix} \tag{11}$$

$$C = \int |A_t(x)|^2\, d^2x \tag{12}$$

B. QUANTUM ANNEALING ALGORITHM

The QA algorithm is based on the adiabatic theorem, which states that if a quantum system is in its ground state and the Hamiltonian governing the system's dynamics is changed slowly enough, then the system will remain in its ground state throughout the evolution. The theorem can be used for computation by preparing a system in the ground state with an easy-to-solve initial Hamiltonian H_i, and then slowly transitioning to the more complex Hamiltonian of the problem to be solved, denoted as the time-dependent Hamiltonian H_f in Eq. (13).

$$H(s) = A(s)\,H_i + B(s)\,H_f, \tag{13}$$

where s represents a normalized time factor characterizing the annealing fraction, which is defined as the ratio of the current time to the total annealing time. $A(s)$ and $B(s)$ are monotonic weighting functions usually defined by the quantum computer hardware, as shown in Fig. 2(a), such that $A(s = 0) = 1, B(s = 0) = 0$, and $A(s = 1) = 0, B(s = 1) = 1$. As the transition from H_i to H_f occurs gradually, the system undergoes a transformation from its initial (ground) state H_i, to the ground state of the problem Hamiltonian H_f. Throughout this process, H_i gradually diminishes in influence, and the system is increasingly governed by H_f. When s reaches 1, the system becomes purely classical, and the final states of the qubits are measured. This measurement yields the lowest (ground-state) energy of the classical QUBO model with binary variables, following the same structure as described in Eq. (8).

Once the problem Hamiltonian is defined, the simulation process takes on five stages to produce solutions of interest [12], [14], [15]: (1) *Converting the QUBO definition into a logical graph*, where each node represents a variable, and each edge denotes the interaction term between a pair of variables. (2) *Embedding the logical graph into the QA hardware*, where the logical graph is translated onto the physical hardware graph of the QPU by selecting sets of physical qubits to represent a single logical node and to identify the couplings between the physical qubits that realize the correct interactions between the logical variables. Here we use D-wave's default mapping in Advantage system 6, which maps the logical variables to the physical qubits using the Pegasus graph for optimized connectivity and improved scalability.

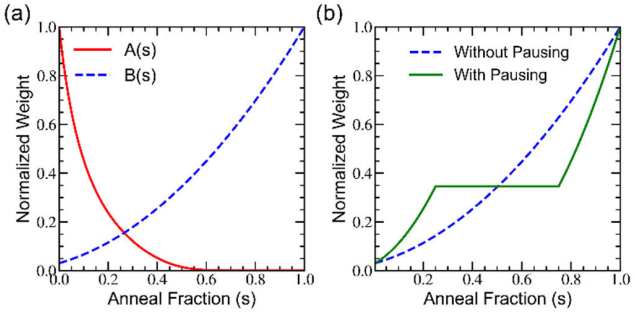

FIGURE 2. (a) schematic annealing schedule which shows the change of A(s) (red solid curve) and B(s) (blue dash curve) in D-wave advantage system 6. (b) Schematic annealing schedules with and without pausing are shown in green solid and blue dash curves respectively.

(3) *Programming and Initialization*, where programming the quantum annealer requires setting the parameters that define the problem to be solved, which is set as the final Hamiltonian. This involves setting the weights for each qubit bias (to control the magnetic field acting on the qubit) and coupler strength (to control the interaction between qubits). After programming, the spin configuration of the QPU is initialized as the lowest energy configuration of an easy-to-implement initial Hamiltonian, where the qubits are placed in an equal superposition of all possible states. (4) *Annealing Process*, which involves solving the Ising/QUBO model by transitioning the system from the initial to the final Hamiltonian using predefined annealing functions to minimize energy. It can also serve as the central component of a hybrid approach, where a quantum processor handles the inner loop of the calculation. And finally, (5) *Readout solutions and Resampling*, in which after the annealing phase, the qubits reach an eigenstate or a superposition of eigenstates in the computational basis, each corresponding to a potential minimum of the final Hamiltonian. However, since QA is a heuristic approach, there is a nonzero probability that the computation yields a ground state of the system. To mitigate the uncertainty, the anneal-readout sequence is repeated multiple times for each input to acquire multiple candidate solutions through resampling.

The calculations are performed using the D-wave Advantage system 6 to explore the effects of global annealing trajectories, also referred to as annealing time and annealing schedule.

A. *Annealing Time*: It can be adjusted by scaling the quadratic growth in $B(s)$, thereby allowing for the entire annealing process to occur at a faster or slower rate (Fig. 2(b), dashed blue curve). This adjustment provides control over the evolving rate of the quantum system.

B. *Annealing Schedule*: This can be manipulated by introducing a pause through the annealing process. A pause involves interrupting the quadratic growth of $B(s)$ and maintaining it at a specific scale for a particular duration (Fig. 2(b), solid green curve). An appropriate pausing event may increase the probability of tunneling between the excited and the ground states.

C. PSEUDO CODES

In this section, we describe the algorithms presented in pseudo codes. Table 1 shows the overall steps for QA using the D-Wave quantum processor. It takes a QUBO matrix as input and outputs the solution vector and energy after running on the quantum annealer. The key steps are selecting the D-Wave sampler, embedding topology, setting annealing parameters, sampling from the quantum state, and reading out the final state. In addition to the quantum annealer, D-Wave's cloud service allows the implementation of classical solvers for binary quadratic problems [32]. The SD algorithm was executed on Amazon Elastic Compute Cloud with Intel(R) Xeon(R) Platinum 8175M CPU. Table 2 outlines the SD algorithm for minimizing a QUBO objective function. It iteratively flips variables to greedy reduce the energy at each step. The gradient or flip energies are calculated to determine the steepest descent direction. It repeats this greedy variable flip process until reaching a minimum. Table 3 demonstrates the hybrid approach combining QA and an SD post processing. It first runs QA to get an initial solution state. Then it uses this state to seed SD as a local search to refine the solution. The hybrid approach leverages the global search capabilities of QA and combines it with the local optimization of SD.

TABLE 1. Quantum annealing algorithm for QUBO optimization.

Algorithm Quantum Annealing

Input: **QUBO** matrix
Output: Solution vector ***S*** and **Energy** value

Select Sampler:
Establish D-wave Sampler \\ *Advantage system 6.3*
Composite the Embedding \\ *Pegasus*

Compile QUBO Model:
parameters ➔ sample number, annealing time,
reduce intersample correlation…
Response = Sampler.sample(**QUBO,** parameters)

Results:
Get ***S*** and **Energy** from Response

III. RESULTS AND DISCUSSIONS

In this section, we evaluate the effectiveness of D-Wave's quantum annealer in minimizing the QUBO Hamiltonian through the identification of an optimal binary mask. The study focuses on the Advantage 6 system, which supports up to 5760 qubits and accommodates up to 64 logic variables using Pegasus embedding. We explore variable configurations of $N = 5 \times 5$, 6×6, 7×7, and 8×8. Starting with an arbitrary binary mask, we compute its amplitude image as the target, serving as the benchmark and global minimum solution sought via the QA algorithm. We present results from the QA solver with and without pauses, a hybrid approach

TABLE 2. Classical steepest descent algorithm for QUBO optimization.

Algorithm Steepest Descent

Input: QUBO matrix
Output: Solution vector ***S*** and **Energy** value

For each Sample:
Start with an initial state \\ *Random or given*
Calculate the initial energy based on the problem definition.
Iterative Step ➔ Energy minimizes or reaches a threshold.
1. Calculate flip energy for each variable.
2. While there is room for energy reduction:
˙Identify the variable that, when flipped, results in the most substantial energy reduction \\ *Steepest Descent*
˙Flip the identified variable.
˙Recompute flip energies for the flipped variable and its neighbors.

Result and Termination:
Store the final solution vector ***S***, and **Energy**.

TABLE 3. Hybrid quantum annealing and steepest descent algorithm for QUBO optimization.

Algorithm Hybrid QA + SD

Input: QUBO matrix
Output: Solution vector ***S*** and **Energy** value

Select Sampler:
Establish D-wave Sampler \\ *Advantage system 6.3*
Composite the Embedding \\ *Pegasus*

Quantum Annealing:
parameters ➔ sample number, annealing time, reduce intersample correlation…
Response = Sampler.sample(**QUBO,** parameters)

SD Post Processing \\ *Local Search*
Use the results from QA as an initial state.
Do the Steepest Descent Process

Results:
Store the final solution vector ***S***, and **Energy**

that applies the classical SD algorithm to the QA results, and insights into solution characteristics as the reticle size grows.

A. QA SOLVER

The first simulation experiment is conducted for a binary mask in a $N = 5 \times 5$ pixel array, where the 2D shape function is designed as a simplified letter "A", as shown in the upper row of Fig. 3(a). The amplitude image is calculated by convolving an impulse response function with the predefined mask and then normalized to the maximal amplitude value, as shown in the bottom row of Fig. 3(a). A simulation cycle is configured as follows: the annealing time is set to 10 μs and resampled 100 times to increase the probability of finding the optimal solution. As a result, the quantum processing unit (QPU) access time is 1 ms in total without delays between samples. Fig. 3(b) to (d) show the three lowest-energy sampled states from individual samples within one simulation cycle; the top row shows the resulting binary masks, and the bottom shows the corresponding amplitude images. As depicted in Fig. 3(b) bottom row, the amplitude image derived from the lowest-energy sampled solution closely resembles the target. However, the optimized binary mask still exhibits minor discrepancies, with two pixels failing to match the original mask.

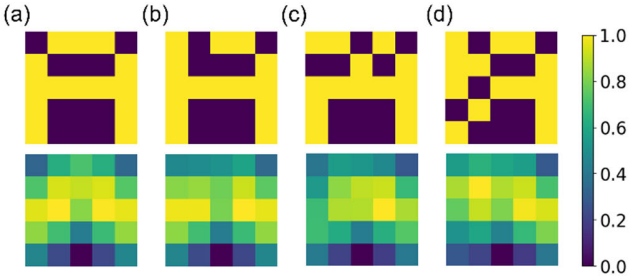

FIGURE 3. (a) The target mask pattern, representing the ground truth for quantum annealing, and the calculated amplitude image (bottom row). (b) to (d) The three lowest-energy-sampled solutions resulting from 100 individual samples within one simulation cycle. The top row shows the binary masks, and the bottom the corresponding amplitude image. All share the common scale on the right.

Since QA is a heuristic algorithm, the lowest energies obtained from individual samples and cycles are different. Therefore, in addition to 100 samples within a QA cycle, we repeat the simulation cycles 100 times to calculate the average lowest energy for various annealing times: 10 μs, 20 μs, 40 μs, 100 μs, and 1000 μs. The results are normalized to the maximal average value of the 10 μs annealing time and plotted in the dashed blue histogram in Fig. 4. It is observed that the normalized energy resulting from the 100 cycles monotonically decreases with annealing time. The longer the annealing time, the lower the average energy.

Moreover, in D-Wave's quantum annealer, ISC refers to the relationship between different samples during the sampling process. Specifically, ISC measures the degree of similarity or dependence between these individual samples. A higher ISC implies that samples are more correlated and more akin to each other, potentially resulting in redundancy or limited exploration of the solution space. Conversely, a lower ISC suggests greater diversity among samples, allowing for a more comprehensive exploration that potentially uncovers better solutions. In previous simulations, the ISC is set to off by default, meaning that there is no delay time between individual samples. Therefore, we investigate the impact of reduced ISC on the outcomes by adding a time delay between different sample reads. According to the D-wave documentation, a time delay mitigates the correlation due to the spin-bath polarization effect [33].

After introducing a delay time to mitigate the ISC, the normalized energy diminishes compared to the scenario without delay, as depicted by the dotted red histogram in Fig. 4. In this instance, the normalized energy is decreased by 2.64% to 7.39% in comparison to the original value. Despite this reduction, the overall average energy also continues to decline

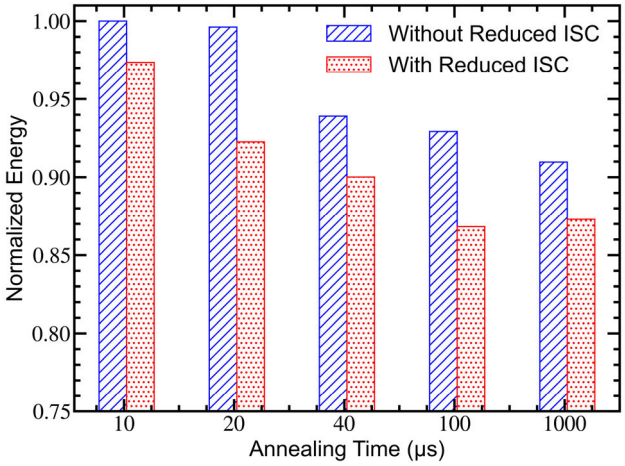

FIGURE 4. Comparison of the averaged lowest energies with and without reduced inter-sample correlation (ISC) across a range of annealing times (10 to 1000 μs). Reduced ISC involves introducing a time delay between individual samples.

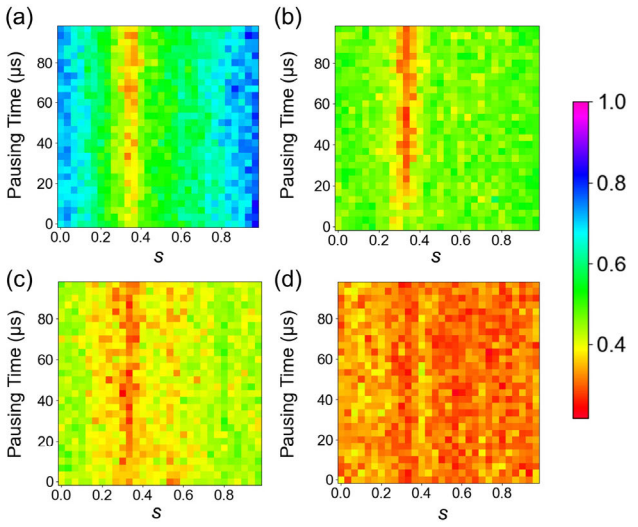

FIGURE 5. Average energy map across pausing at time fraction s vs duration, for different annealing times: (a) 1 μs (b) 5 μs (c) 20 μs, and (d)100 μs.

as annealing time increases, albeit with signs of saturation around the 100 μs annealing time mark, wherein the energy improvement becomes marginal. These findings affirm the efficacy of ISC reduction in broadening the exploration scope and consequently, attaining solutions that better approach the global minimum. Consequently, reduced ISC is applied throughout the rest of the work.

Next, we introduce pausing into the forward annealing [34], [35], [36], [37]. We perform a comprehensive exploration of the optimized annealing schedule with pausing at various annealing fractions and different pausing durations. Fig. 5(a) to 5(d) show the average lowest energy plotted against the annealing fraction for $s = 0 \sim 1$ and a pausing duration from 0 to 1000 μs for annealing times of 1, 5, 20, and 100 μs, respectively. In Fig. 5(a), we observe a clear energy valley near $s = 0.36$, indicating that implementing a pause in the middle of the QA process promotes tunneling into the ground state, thereby yielding a lower energy. The pause duration appears to have much less influence on reducing the energy than the time fraction. As the annealing time is extended to 5 and 20 μs, the discernible presence of an energy valley gradually diminishes and ultimately vanishes for the 100 μs annealing time, as shown in Figs. 5(b) through 5(d). These observations align with previous results in which increasing the annealing time leads to a progressive reduction in the average lowest energy, ultimately reaching a saturation point around the 100 μs annealing time. In other words, when the annealing time is as short as 1 μs, introducing pausing at position $s = 0.36$ may result in similar lowest energies as those obtained with the 100 μs annealing time. However, when the annealing time increases to 100 μs, the benefit of pausing completely vanishes. To summarize, pausing is beneficial for finding optimal solutions beyond the capabilities of short annealing times but becomes ineffective with extended annealing times.

TABLE 4. Comparison of quantum annealing (QA), steepest descent (SD), and hybrid (QA+SD) solvers.

Properties	QA	SD	Hybrid
Ground-truth Probability (%)	0.0075	0.1350	0.3125
Runtime per shot (ms)	0.1	29.0	20.2
Equivalent runtime (s)	1.33	21.46	6.44

B. HYBRID QA AND SD SOLVER

In this section, we compare solutions obtained from the QA, the classical SD, and hybrid (QA+SD) solvers. The hybrid solver applies the SD post-processing to solutions derived from the quantum annealer. We conducted 40,000 simulations using the previous 5×5 binary mask with each of these methods. Both the QA and hybrid solvers used 50 cycles of annealing, with each cycle consisting of 800 samples (shots), resulting in 40,000 shots for each approach. We compare the success probabilities of finding the ground truth solution across these methods, as shown in Table 4. The runtime per shot represents the time required for one simulation for each method. Additionally, we calculate the equivalent runtime required to obtain a global minimum solution, based on the probability of ground-truth occurrence as expressed in Eq. (14).

$$\text{Equivalent runtime} = \frac{\text{Runtime per shot}}{\text{Ground} - \text{truth Probability}}. \quad (14)$$

Table 4 reveals that QA as a heuristic algorithm exhibits modest accuracy in this 25-variable binary reticle example. However, with QA, each simulation only takes 100 μs, which is 290 times faster than the classical SD algorithm. Considering the ground-truth probability, QA still provides

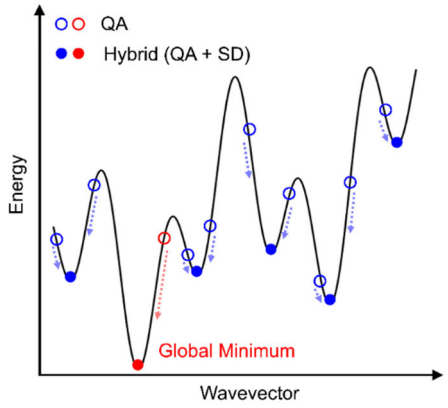

FIGURE 6. Schematic of the energy obtained from QA and hybrid solvers. Initial QA results are represented by transparent dots, while solid dots illustrate the results after the application of the SD post-processing step.

approximately 16 times runtime reduction compared to SD. Furthermore, the hybrid QA+SD solver only requires 30% of the classical runtime. Clearly, both approaches benefit from the significant runtime reduction of the quantum annealer, highlighting the importance of accelerating quantum computing for real-world optimization problems.

QA has the lowest equivalent runtime to locate the global minimum, however, the number of lowest-energy solutions generated from each independent simulation can be massive due to the low probability of accuracy. In practice, the hybrid solver achieves a much higher success rate in locating the ground-truth solution – almost three times higher than the SD solver and more than 41 times higher than the QA solver alone. Therefore, the hybrid QA+SD solver is a more practical approach when solving real-world optimization problems. We elaborate on this argument using the schematic in Fig. 6. The QA algorithm involves tunneling and entanglement effects that can quickly and extensively explore the solution space. However, these solutions are often not perfect and can be very diverse and distributed near local and global valleys. The SD algorithm subsequently applied to the QA solution can bring the QA solution to the nearest valley, thereby improving the probability of finding the global minimum solution. For example, the blue and red dots shown in Fig. 6 represent the eigenvalues of the ground state obtained from individual samples in QA, which may or may not be located near valleys along other wavevector directions. However, after the post-processing of the SD algorithm, they all reach the nearest valley. In this illustration, the red dot that eventually reaches the global valley is not the smallest ground-state solution obtained from the original QA samples. By leveraging the advantages of quantum and classical computing, hybrid QA and SD solvers may provide more efficient solutions to binary reticle optimization problems than QA or the classical method alone.

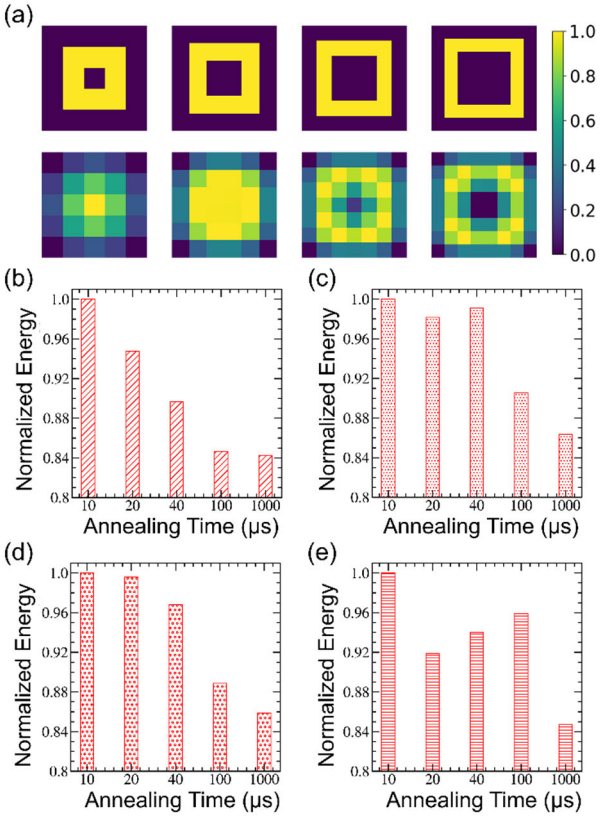

FIGURE 7. (a)The predefined square-donut-shaped mask with varying sizes: 5 × 5 to 8 × 8 pixels and their corresponding amplitude images, sharing the common scale on the right. (b)-(e) The normalized ground state energy for each mask across a range of annealing times (10 to 1000 μs).

C. LARGE RETICLES

Next, we increase the number of variables, expanding the dimensions of the binary mask from 5 × 5 to 6 × 6, 7 × 7, and the final configuration of 8 × 8 pixels. In Fig. 7(a), the predefined mask configurations assume a symmetric, square-donut shape, placed above the corresponding target amplitude image. As evidenced in Fig. 7(b) through Fig. 7(e), the decline in the average energy with increased annealing time is more clearly observable for smaller reticles possessing fewer variables, while fluctuations in energy still occur for larger dimensions as in the reticles with 6×6 and 8×8 pixels. Notably, within the optimized 8 × 8 reticle, the average lowest energy experiences significant fluctuations with varying annealing times. These fluctuations could potentially be attributed to QPU stability, particularly when the number of logical variables approaches the upper limit of physical qubits needed for the Pegasus embedding.

To explore this further, we plotted the lowest energies from 500 samples within a single QA simulation cycle for different reticle sizes. As shown in Fig. 8(a), the default sorted energies generally increase monotonically, with some outliers. The number and spread of these irregular values grow with more variables, eventually blurring the trajectory for the

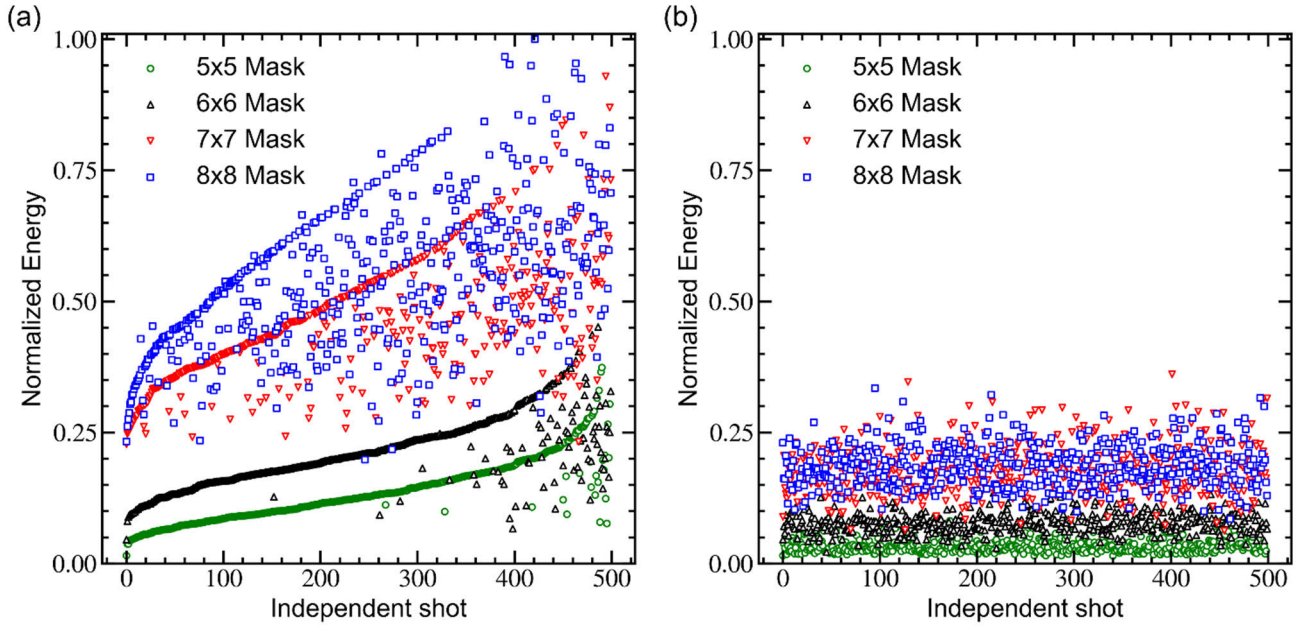

FIGURE 8. Simulated lowest energies in a single annealing cycle with 500 samples for mask sizes of 5 × 5 to 8 × 8 pixels. (a) the QA results, (b) the SD outcomes after QA (hybrid approach).

8 × 8 case. These irregularities stem from chain breakages in the quantum annealer, where longer chains heighten the risk of system instability. Although bolstering chain strength could prevent these breaks, it introduces distortions to problem states. As the implementation of ILT problems requires densely interconnected qubits in Dwave architecture, optimizing chain strength and embedding techniques emerge as crucial factors to the solution quality, warranting deeper exploration in future investigation.

We then applied SD post-processing to the QA results in Fig. 8(a). As shown in Fig. 8(b), the hybrid QA+SD approach substantially reduces the minimum energy for all cases, even from higher starting QA values. As previously explained in Fig. 6, the result demonstrates the power of SD refinement following global QA sampling. Smaller 5×5 and 6×6 reticles converge to lower average energies than larger 7 × 7 and 8 × 8 cases. The probability of finding the true global minimum also drops with the larger size, indicating greater difficulty. For 7 × 7 and 8 × 8, the hybrid solver fails to recover the predefined solutions, limited by quantum hardware. Figure 9 shows the optimized solutions found by the QA, SD, and hybrid solvers for the 7 × 7 and 8 × 8 reticles, along with their corresponding amplitude images. For the 7 × 7 case, the lowest normalized energies are 0.156 (QA), 0.042 (SD), and 0.027 (hybrid). For the 8×8 case, the energies are 0.228 (QA), 0.081 (SD), and 0.078 (hybrid). Although the true global minimum is not achieved, the hybrid solver consistently yields lower energy solutions compared to using QA or SD independently. The substantial energy reductions highlight the value of the hybrid approach in tackling the complex optimization challenges of larger binary reticles. Even without reaching the ground truth, combining QA and SD provides significantly improved solutions over the individual techniques, which underscores the potential of hybrid quantum-classical solvers for advanced ILT optimization problems beyond the scope of current quantum hardware capabilities.

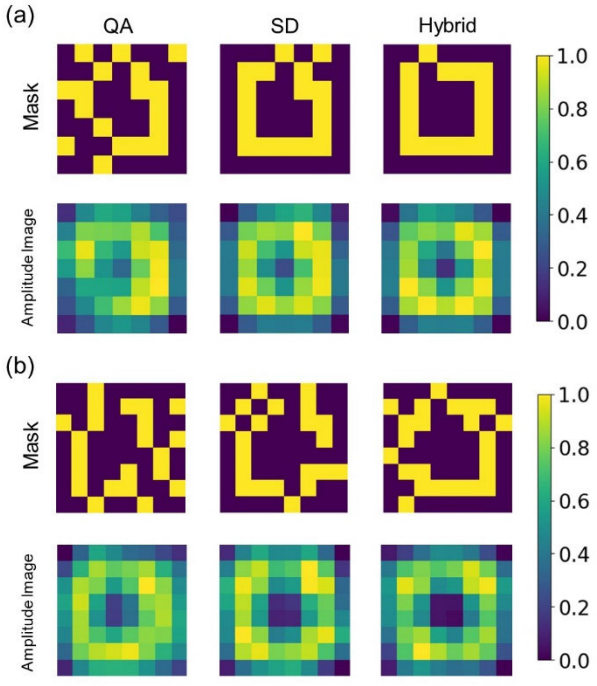

FIGURE 9. The optimal solutions of (a) 7 × 7 (b) 8 × 8 masks calculated by QA, SD, and hybrid QA+SD solvers with corresponding amplitude images.

IV. CONCLUSION

This study delved into the efficacy of using D-Wave's quantum annealer and hybrid quantum-classical solvers for the optimization of pixelated binary reticle problems encompassing 5×5, 6×6, 7×7, and 8×8 variables. Our investigation highlights that extended annealing times, reduced inter-sample correlation, and strategically timed pausing during quantum annealing collectively contribute to improved lowest-energy-sampled solutions, leading to minimized Hamiltonian values. Furthermore, we show that a hybrid solver approach yields an increased probability of accuracy compared to the exclusive use of either pure quantum or classical solvers, while also offering runtime savings of approximately two-thirds to the classical method. With quantum computing being a rapidly advancing field, our findings support that D-Wave's quantum annealer could prove to be crucial for practical applications in computational lithography, particularly as hardware scalability and stability continue to improve.

ACKNOWLEDGMENT

The authors would like to thank the National Center for High-performance Computing (NCHC), National Applied Research Laboratories (NARLabs), Taiwan, for providing computational and storage resources.

REFERENCES

[1] A. K. Wong, *Resolution Enhancement Technique for Optical Lithography*. Bellingham, WA, USA: SPIE Press, 2001.

[2] L. Pang, "Inverse lithography technology: 30 years from concept to practical, full-chip reality," *J. Micro/Nanopatterning, Mater., Metrol.*, vol. 20, no. 3, Aug. 2021, Art. no. 030901.

[3] L. Pang, Y. Liu, and D. Abrams, "Inverse lithography technology (ILT): What is the impact to the photomask industry?" in *SPIE Proc.*, San Jose, CA, USA, May 2006, p. 62831.

[4] D. S. Abrams and L. Pang, "Fast inverse lithography technology," in *SPIE Proc.*, San Jose, CA, USA, Mar. 2006, pp. 61540C-1–61540C-8.

[5] NVIDIA. (2023). *NVIDIA cuLitho—Accelerate Computational Lithography*. [Online]. Available: https://developer.nvidia.com/culitho

[6] Y. Granik, "Fast pixel-based mask optimization for inverse lithography," *J. Micro/Nanolithography, MEMS, MOEMS*, vol. 5, no. 4, Oct. 2006, Art. no. 043002.

[7] A. Poonawala and P. Milanfar, "Mask design for optical microlithography—An inverse imaging problem," *IEEE Trans. Image Process.*, vol. 16, no. 3, pp. 774–788, Mar. 2007.

[8] Y. Shen, N. Wong, and E. Y. Lam, "Level-set-based inverse lithography for photomask synthesis," *Opt. Exp.*, vol. 17, no. 26, pp. 23690–23701, Dec. 2009.

[9] J.-C. Yu and P. Yu, "Impacts of cost functions on inverse lithography patterning," *Opt. Exp.*, vol. 18, no. 22, pp. 23331–23342, Oct. 2010.

[10] J. F. Poyatos, J. I. Cirac, and P. Zoller, "Complete characterization of a quantum process: The two-bit quantum gate," *Phys. Rev. Lett.*, vol. 78, no. 2, pp. 390–393, Jan. 1997.

[11] T. Albash and D. A. Lidar, "Adiabatic quantum computation," *Rev. Mod. Phys.*, vol. 90, no. 1, Jan./Mar. 2018, Art. no. 015002.

[12] S. Yarkoni, E. Raponi, T. Bäck, and S. Schmitt, "Quantum annealing for industry applications: Introduction and review," *Rep. Prog. Phys.*, vol. 85, no. 10, Oct. 2022, Art. no. 104001.

[13] T. Kadowaki and H. Nishimori, "Quantum annealing in the transverse ising model," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 58, no. 5, pp. 5355–5363, Nov. 1998.

[14] P. Hauke, H. G. Katzgraber, W. Lechner, H. Nishimori, and W. D. Oliver, "Perspectives of quantum annealing: Methods and implementations," *Rep. Prog. Phys.*, vol. 83, no. 5, May 2020, Art. no. 054401.

[15] A. B. Finnila, M. A. Gomez, C. Sebenik, C. Stenson, and J. D. Doll, "Quantum annealing: A new method for minimizing multidimensional functions," *Chem. Phys. Lett.*, vol. 219, nos. 5–6, pp. 343–348, Mar. 1994.

[16] E. Gibney, "D-wave upgrade: How scientists are using the world's most controversial quantum computer," *Nature*, vol. 541, no. 7638, pp. 447–448, Jan. 2017.

[17] S. W. Shin, G. Smith, J. A. Smolin, and U. Vazirani, "How 'quantum' is the D-Wave machine?" 2014, *arXiv:1401.7087*.

[18] H. Ushijima-Mwesigwa, C. F. A. Negre, and S. M. Mniszewski, "Graph partitioning using quantum annealing on the D-Wave system," in *Proc. 2nd Int. Workshop Post Moores Era Supercomputing*, Denver, CO, USA, Nov. 2017, pp. 22–29.

[19] R. K. Nath, H. Thapliyal, and T. S. Humble, "A review of machine learning classification using quantum annealing for real-world applications," *Social Netw. Comput. Sci.*, vol. 2, no. 5, pp. 1–11, Sep. 2021.

[20] J. Biamonte, P. Wittek, N. Pancotti, P. Rebentrost, N. Wiebe, and S. Lloyd, "Quantum machine learning," *Nature*, vol. 549, no. 7671, pp. 195–202, Sep. 7671.

[21] K. Ikeda, Y. Nakamura, and T. S. Humble, "Application of quantum annealing to nurse scheduling problem," *Sci. Rep.*, vol. 9, no. 1, p. 12837, Sep. 2019.

[22] D. Venturelli, D. J. J. Marchand, and G. Rojo, "Quantum annealing implementation of job-shop scheduling," 2015, *arXiv:1506.08479*.

[23] Y. Cao, J. Romero, J. P. Olson, M. Degroote, P. D. Johnson, M. Kieferová, I. D. Kivlichan, T. Menke, B. Peropadre, N. P. D. Sawaya, S. Sim, L. Veis, and A. Aspuru-Guzik, "Quantum chemistry in the age of quantum computing," *Chem. Rev.*, vol. 119, no. 19, pp. 10856–10915, Oct. 2019.

[24] A Bayerstadler et al., "Industry quantum computing applications," *EPJ Quantum Technol.*, vol. 8, no. 1, p. 25, Dec. 2021.

[25] C. Ross, G. Gradoni, Q. J. Lim, and Z. Peng, "Engineering reflective metasurfaces with Ising Hamiltonian and quantum annealing," *IEEE Trans. Antennas Propag.*, vol. 70, no. 4, pp. 2841–2854, Apr. 2022.

[26] J. M. Arrazola, A. Delgado, B. R. Bardhan, and S. Lloyd, "Quantum-inspired algorithms in practice," *Quantum*, vol. 4, p. 307, Aug. 2020.

[27] Y. Okudaira and S. Yashiki, "Pixelated mask optimization on quantum computers," *Proc. SPIE*, vol. 11327, Mar. 2020, Art. no. 1132705.

[28] D-Wave Systems Inc. (2022). *Advantage Processor Overview*. [Online]. Available: https://www.dwavesys.com/media/3xvdipcn/14-1058a-a_advantage_processor_overview.pdf

[29] D-Wave Systems Inc. (2023). *QPU-Specific Physical Properties*. [Online]. Available: https://docs.dwavesys.com/docs/latest/_downloads/3e684673f98390d0ba497033b7432e3b/09-1272A-B_QPU_Properties_Advantage_system6_2.pdf

[30] M. Born and E. Wolf, *Principles of Optics*, 7th ed. Cambridge, U.K.: Cambridge Univ. Press, 1999.

[31] N. Cobb, "Sum of coherent systems decomposition by SVD," Dept. Elect. Eng. Comput. Sci., Univ. California, Berkeley, CA, USA, Tech. Rep. 94720, 1995, pp. 1–7.

[32] D-Wave Systems Inc. (2023). *D-Wave Ocean Software Documentation: Dwave–Greedy*. [Online]. Available: https://docs.ocean.dwavesys.com/projects/greedy/en/latest

[33] D-Wave Systems Inc. (2023). *QPU Solver Datasheet: Spin-Bath Polarization Effect*. [Online]. Available: https://docs.dwavesys.com/docs/latest/c_qpu_errors.html

[34] J. Marshall, D. Venturelli, I. Hen, and E. G. Rieffel, "Power of pausing: Advancing understanding of thermalization in experimental quantum annealers," *Phys. Rev. Appl.*, vol. 11, no. 4, Apr. 2019, Art. no. 044083.

[35] D. Venturelli and A. Kondratyev, "Reverse quantum annealing approach to portfolio optimization problems," *Quantum Mach. Intell.*, vol. 1, nos. 1–2, pp. 17–30, May 2019.

[36] M. Zielewski, K. Takahashi, Y. Shimomura, and H. Takizawa, "Efficient pause location prediction using quantum annealing simulations and machine learning," *IEEE Access*, vol. 11, pp. 104285–104294, 2023.

[37] P. X. Fang, Y. S. Chen, J. S. Wu, and P. Yu, "Exploring quantum annealing strategies for reticle optimizations," *Proc. SPIE*, vol. 12495, Apr. 2023, Art. no. 1249521.

PO-HSUN FANG received the B.S. degree in photonics from National Yang Ming Chiao Tung University, in 2022, where he is currently pursuing the M.S. degree with the Department of Photonics. His research interests include inverse lithography technology (ILT) for ArF 193 immersion and extreme ultraviolet (EUV) lithography.

YAN-SYUN CHEN received the master's degree from the Institute of Electronics Engineering, National Tsing Hua University, Hsinchu, Taiwan, in 2017. From 2017 to 2019, he joined the Develop Team, Taiwan Semiconductor Manufacturing Company (TSMC), as a N5&N3 EPI Process Engineer (FinFET, Source and Drain). During his tenure, he also became a member with the N3 One-Team. Since 2019, he has been a Research Assistant with the Department of Photonics, National Yang Ming Chiao Tung University. His research interests include utilizing quantum annealing techniques to optimize semiconductor manufacturing technology, including inverse lithography technology (ILT). Furthermore, he is responsible for collaborating with various research institutions in Taiwan to promote the educational outreach and adoption of quantum annealing technology.

JHIH-SHENG WU received the B.Sc. and M.Sc. degrees in physics from National Taiwan University, in 2006 and 2010, respectively, and the Ph.D. degree in physics from the University of California, San Diego, in 2016. From January 2017 to December 2019, he was a Postdoctoral Fellow with the Center of Nano-Optics, Georgia State University. He was a Postdoctoral Fellow with RCAS, Academia Sinica, from January 2020 to July 2020. Since August 2020, he has been a Faculty Member and is currently an Assistant Professor with the Department of Photonics, National Yang Ming Chiao Tung University. His research interests include theoretical aspects of nano-optics, quantum optics, and quantum application.

PEICHEN YU (Member, IEEE) received the Ph.D. degree in electrical engineering from the University of Michigan, Ann Arbor, MI, USA, in 2004. From 2004 to 2006, she joined the Advanced Design Group, Intel Corporation, Hillsboro, OR, USA, as a Resolution Enhancement Technology (RET) Design Engineer. Since 2006, she switched careers to academia and is currently a Professor with the Department of Photonics, National Yang Ming Chiao Tung University. Her research interests include nanostructures and metasurfaces patterning for optoelectronic applications. She is also actively engaged in the development of RET solutions, including inverse lithography technology (ILT) for ArF 193i and EUV lithography. She has published over 60 refereed technical articles in the above research areas. Her work has been highlighted in various scientific journals, including *Virtual Journal of Nanoscale Science and Technology*, SPIE Newsroom, and *NPG Nature Asia-Material*. She is also a member of the IEEE Photonics Society and SPIE. She received several research and teaching awards in Taiwan.