

Integrated photonic Ising machine with full connectivity for ultra-large-scale combinatorial optimization

Guanyu Chen

gychen@cqu.edu.cn

Chongqing University <https://orcid.org/0000-0001-7388-7149>

Ziyao Zhang

Chongqing University

Yuan Gao

National University of Singapore

Jiayi Gao

Sun Yat-sen University

Anil Prabhakar

Indian Institute of Technology Madras

Jie Liu

Sun Yat-sen University

Tao Zhu

Chongqing University

Aaron J. Danner

National University of Singapore

Research Article

Keywords:

Posted Date: September 9th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10959649/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Integrated photonic Ising machine with full connectivity for ultra-large-scale combinatorial optimization

Ziyao Zhang^{1,†}, Guanyu Chen^{1,2,†,*}, Yuan Gao^{3†}, Jiayi Gao⁴, Anil Prabhakar⁵, Jie Liu⁴, Tao Zhu¹ and Aaron J. Danner^{3,#}.

¹*Key Laboratory of Optoelectronic Technology & Systems (Ministry of Education) and College of Optoelectronic Engineering, Chongqing University, Chongqing 400044, China*

²*Institute of Advanced Interdisciplinary Studies, Chongqing University, Chongqing 400044, China*

³*Department of Electrical and Computer Engineering, National University of Singapore, 4 Engineering Drive 3, 117583, Singapore*

⁴*State Key Laboratory of Optoelectronic Materials and Technologies, School of Electronics and Information Technology, Sun Yat-Sen University, Guangzhou 510006, China*

⁵*Department of Electrical Engineering, Indian Institute of Technology Madras, Chennai 600036, India*

†Contributed equally

**gychen@cqu.edu.cn ; #adanner@nus.edu.sg*

Abstract: Ising machines can efficiently solve real-world combinatorial optimization problems. Among them, photonic Ising machines (PIMs) exploit the parallelism and broadband bandwidth of optical systems to outperform digital computers in certain metrics. However, the bulky footprint of conventional discrete components imposes significant limitations on the miniaturization of current PIMs. Meanwhile, the scalability of problem size and spin connectivity density is constrained by limited computational resources. To address these challenges, we proposed and demonstrated an integrated PIM (IPIM) based on a hybrid architecture on a silicon-based photonic platform with an optical component footprint of 0.065 mm², achieved success rates exceeding 90% in finding global optimal solutions across all three benchmark tests. To tackle the common challenge in PIMs, where large-scale and high-connectivity-density problems naturally incur high computational complexity, we developed a joint algorithm with sparsity, inverse sparsity, and decomposition operations (JA-SID). Using this approach, we successfully solved Ising problems with sparse and dense spin connectivity, involving 102,400 spins. In addition, we introduced an optimized parameter configuration scheme (OPCS), which reduced the iteration count by over 50% for slowly converging problems, thereby lowering time complexity and power consumption. In practice, we used the IPIM to successfully partition a Facebook social network with over 60,000 users, demonstrating its capability to handle ultra-large-scale real-world problems. Our research substantially extended the applicability of IPIM, offering a promising solution to the increasing global demand for computational resources.

Introduction

A wide range of real world problems can be categorized as combinatorial optimization problems [1]. By translating practical constraints into mathematical formulations and defining appropriate objective functions, such problems can be effectively modeled and solved in various areas, including computer science [2, 3], operations research [1, 4-6], healthcare [7, 8], communications, and circuit design [9]. Due to their NP-hardness and exponential complexity [1, 6], it is difficult to efficiently solve combinatorial optimization problems on classical von Neumann-based computers. Despite the use of heuristic algorithms, inherent limitations in algorithmic principles or hardware often prevent these systems from efficiently solving ultra-large-scale problems [10, 11]. Contrasting with these computational bottlenecks is the exponential growth in demand for large-scale computing, exemplified by million-member community analysis in social network and billion-node placement challenges in advanced chip design. Accordingly, narrowing the gap between current computational capabilities and increasing real world demands is now a key priority.

By mapping combinatorial optimization problems into Ising models, Ising machines (IMs) enable rapid solution finding through efficient exploration of the solution landscape [5, 10, 12-16], positioning them as a powerful approach to tackle NP-hard and NP-complete problems that are typically intractable for classical computers [1, 6]. IMs can be broadly classified into photonic [10, 13, 14, 17-24], magnetic [5], and electronic [15, 16] implementations, among others. Photonic IMs (PIMs) benefit from the intrinsic properties of optical systems, such as parallel processing capability, low power consumption and high speed. The exceptional computational capabilities of optical platforms, especially their performance in both linear and nonlinear regimes [25-29], offer promise for efficient PIMs [30].

Currently, PIMs are implemented through several approaches, including fiber-optic based architectures [10, 13, 19], quantum annealing methods [12], and spatial optical systems [31-34]. Although these techniques have demonstrated considerable potential, they commonly face challenges in terms of scalability, system simplicity, and integrability. Accordingly, attentions have been directed toward integrated PIMs (IPIMs) [17, 22, 30, 35]. Nonetheless, existing implementations are often hindered by fabrication complexities associated with heterogeneous integration [17], or by intrinsic limitations in scaling the number of spins [35]. On the other hand, most PIMs rely on matrix-vector multiplication (MVM) [2, 10, 13, 14, 17, 22-24, 30-34], leading to a quadratic scaling of computational complexity with system size. Some studies have cleverly exploited the sparsity of the coupling matrices in certain Ising problems to reduce computation through sparse algorithms [17]. However, these methods become ineffective for

densely connected spin systems. Although those PIMs may inherently support full spin connectivity, the enormous computational demand results in high time complexity and memory bottlenecks, limiting solving efficiency [10, 17]. Therefore, there remains a critical gap in the development of IPIMs that can concurrently realize low integration complexity and ultra-large scale computational performance.

In this work, we therefore propose and demonstrate an IPIM based on silicon photonics. The IPIM is a hybrid optoelectronic solver, comprising a nonlinear update chip and an electronic feedback module. The IPIM chip integrates a silicon (Si) modulator and a germanium-silicon (Ge-Si) photodetector, providing a solution for stable on-chip Ising sampling in large-scale scenarios. In benchmark tests on graph partitioning problems with 100+-spins across three distinct topologies, our IPIM achieves convergence to global optimal solutions with over 90% probability within 100 iterations. As shown in Figure 1(a), PIMs generally face two major challenges in large-scale computations. The first is the quadratic increase in linear computational cost with problem scale, and the second is the efficiency degradation of sparse matrix algorithms caused by a sharp rise in the number of non-zero elements under large-scale densely connected scenarios. To overcome these limitation, we developed a joint algorithm with sparsity, inverse sparsity, and decomposition operations (JA-SID). The JA-SID mitigates the cache bottleneck induced by diverse large-scale coupling matrices while enhancing linear computational speed and energy efficiency for such problems. With stable optical hardware performance and the JA-SID assistance, we successfully solved the MAX-CUT problem for both a sparse graph and a complete graph with 102,400 vertices, while reducing the computational resource consumption for linear operation by up to four orders of magnitude. To further enhance the computational efficiency of architecture of this type, we introduce an optimized parameter configuration scheme (OPCS). This approach achieves a remarkable reduction of over 50% in required iterations, effectively mitigates the domain freezing problem that plagues such PIM systems, and has been successfully applied to ultra-large scale graph partitioning problems. Finally, we return to real world problem solving by addressing a Facebook social network partitioning problem with over 60,000 netizens. Compared with conventional optimization algorithms, our solution demonstrates superior convergence performance, validating IPIM's potential for rapidly solving ultra-large-scale real-world problems. We believe that our IPIM based on silicon photonics presents a technologically viable solution to meet the increasing computational demands of ultra-large scale and fully connected combinatorial optimization in the future. Moreover, its low

integration complexity and compatibility with existing manufacturing processes make it highly suitable for deployment in real world applications.

Results

Principle of the IPIM

The architecture of the proposed IPIM chip is depicted in Fig. 1(b). Its core components comprise an integrated Si Mach-Zehnder interferometer modulator (MZM) and a Ge-Si photodetector, which together enable the nonlinear dynamics required for spin bistability. In the electronic linear processing stage, all spins are processed in parallel by executing the synchronous buffering rule for feedback vectors directly through MVM. It can be expressed as Eq. 1, illustrated in Fig. 1(c). Here, $X_{(k)}$ represents the spin vector composed of all spin amplitudes at the k -th iteration, and $F_{(k+1)}$ represents the feedback signals vector used for the nonlinear operation in the $(k+1)$ -th iteration. The matrix J is obtained through a linear transformation $J = \alpha I + \beta J_{\text{couple}}$, where J_{couple} is the coupling matrix that describes how spins interact [17]. The parameter α represents the system gain and the parameter β is mutual-coupling coefficient, which are used to adjust the gain strength of every spin and the interaction strength between different spins, respectively. The nonlinear operation can be expressed as Eq. 2, where ζ represents Gaussian noise. During the iterative process, the chip enables nonlinear dynamics that approximate those described by Eq. 2, facilitating spin bifurcation. This results in the spin amplitude x stabilizing into one of two discrete states, corresponding to spin-up or spin-down [17, 20]. The detailed formulation of the Hamiltonian and its associated coupling matrix for the MAX-CUT problem is provided in Supplementary Information Section S1.

$$F_{k+1} = J \cdot X_k \quad (1)$$

$$X_{(k+1)} = \cos^2 \left(F_{(k)} + \zeta - \frac{\pi}{4} \right) - \frac{1}{2} \quad (2)$$

In the JA-SID architecture shown in Fig. 1(c), an algorithmic module is implemented comprising three parts: the sparse matrix algorithm (SMA), inverse sparse matrix algorithm (ISMA), and matrix decomposition algorithm (MDA). Depending on the structure of the coupling matrix, these algorithms are applied individually or in combination to alleviate cache bottlenecks and restrain the growth of computational time for large-scale problems. For SMA, we adopt selective storage of non-zero elements for computation. While for ISMA, it optimizes dense or locally dense matrices by storing zero-element positions to indirectly compute the feedback voltage vector F . For MDA, we exploit the decomposable nature of this PIM computing paradigm by partitioning the matrix J into submatrices $J_{s,11}, J_{s,12}, \dots, J_{s,mn}$ and the

spin vector X into subvectors $X_{s,1}, X_{s,2}, \dots, X_{s,n}$ to align with hardware cache sizes for streaming operation. Alternatively, as an auxiliary step, MDA decomposes the original matrix and distributes the computational tasks of sparse or dense submatrices to SMA and ISMA respectively for parallel processing.

The mathematical process described above achieves convergence through iterative operations within the coherent oscillation system depicted in Fig. 1(d). This system employs a hybrid architecture integrating signal conversion modules, digital signal processing cores, and optoelectronic computing chip. Optoelectronic signal transduction is implemented using the IPIM chip, whose cross-section is shown in Fig. 1(e). The chip integrates a 250- μm -long modulator (Fig. 1(f)) and photodetectors with dimensions of $60\text{ }\mu\text{m} \times 40\text{ }\mu\text{m}$ (Fig. 1(g)). The full electro-optic-electronic response of the chip is characterized in Fig. 1(h). Analog-to-digital (ADC) and digital-to-analog (DAC) conversions are precisely managed by a data acquisition card and a programmable voltage source, respectively. More details about characterization of other integrated photonic devices can be found in Supplementary Information Section S3.

Benchmark tests of the IPIM

We first performed benchmark tests to evaluate the computational capability of the IPIM and to verify that the nonlinear dynamics of the on-chip Si modulators can reliably achieve spin state convergence. The evaluation employs multiple benchmark topologies commonly used in IMs [17, 20, 36]. We selected a square lattice graph (100 spins), a triangular lattice graph (144 spins), and a Möbius ladder graph (100 spins) to test the IPIM's performance on MAX-CUT problems. The experimental results presented in Fig. 2 demonstrate that, across 50 independent trials, the success rates of reaching the ground state within 100 iterations are 100%, 90%, and 94%, respectively. Furthermore, under suboptimal solution criteria (96%, 92%, and 76% of the ground state energy), the system demonstrates stable convergence characteristics. Different from classical algorithms like simulated annealing (SA) that typically require tuning over five parameters [11], the IPIM utilizes only two parameters, α and β . This simplicity enabled efficient optimization of parameter values across different problems: square lattice ($\alpha = 0.28$, $\beta = 0.15$), triangular lattice ($\alpha = 0.29$, $\beta = 0.15$), and Möbius ladder graphs ($\alpha = 0.26$, $\beta = 0.15$). Compared to other schemes [20, 24], this architecture achieves a significant improvement in solving success rates. We attribute the observed benchmark improvements to the integrated device's advantages of reduced alignment sensitivity and enhanced environmental robustness.

JA-SID

For an Ising network with N spins, the overhead of nonlinear operations in each IPIM iteration primarily arises from the responses of modulators and photodetectors, as well as the ADC/DAC. These costs scale linearly with the number of spins. In practical implementations, both time-division and space-division multiplexing can effectively scale the solving capacity of IPIMs or PIMs, with the time and space complexity of nonlinear operations increasing only linearly with system size. However, in the case of linear operations, the coupling matrix J_{couple} for N variables has $N \times N$ dimensions. When spin count increases by a factor of r with variable number, the required floating-point operations (FLOPs) scale as r^2 , imposing quadratically growing consumption on linear computation resources, consequently leading to increases in computation time, energy consumption, space occupation, and device count compared to nonlinear operations.

To address these limitations, we introduce JA-SID as a practical adaptation for this IM architecture, as shown in Fig. 3(a). It is designed with an optimization strategy based on the intrinsic characteristics of Ising problems. First, Ising problems can be classified as dense or sparse according to the abundance of zero elements in their coupling matrices[37]. For most graph problems (e.g., G-set benchmarks), the Ising coupling matrix directly derives from the negative of the adjacency matrix, in which the elements $\{0,1\}$ represent non-connection and connection, respectively.

Classical MVM operations, as defined in Eq. 3, exhibit computational redundancy, resulting in $O(N^2)$ FLOPs. Even when using sparse formats like compressed sparse row (CSR) (as formalized in Eq. 4), non-zero elements "1" must still be stored and computed, resulting in unnecessary multiplications or ineffective handling of dense matrices. To address this bottleneck, we introduce an implicit scarcity principle for elements, in which only the positions of scarce matrix elements (e.g., "-1" in sparse matrices or "0" in dense matrices) are stored, allowing additions or subtractions to replace multiplications. The SMA algorithm targets sparse matrices by directly summing the elements of $-X$ corresponding to the "-1" positions to generate the computed vector Y , as formalized in Eq. 5. Conversely, the ISMA algorithm handles dense matrices by storing the indices of "0" elements and computing Y as the accumulation at these indices minus the total sum of X , as described in Eq. 6. In Eqs. 3 to 6, U_{-1} denotes the index set of "-1" elements in the sparse J_{couple} , U_0 represents the index set of "0" elements in the dense J_{couple} , the vector Y represents the result of the MVM between the coupling matrix and the spin vector. The feedback voltage vector F can be obtained by a linear combination:

$F=\alpha X+\beta Y$. Since the number of operations required for this linear combination is only $3N$, its computational complexity becomes negligible compared to the $O(N^2)$ complexity of MVM in large-scale problems. Moreover, because the coefficients α and β are fixed, the computation time of this operation in hardware is considered insignificant relative to that of MVM.

$$Y_i = \sum_{j=-1}^N J_{couple}^{[i,j]} \times X_j \quad (3)$$

$$Y_i = \sum_{j \in U_{-1}} J_{couple}^{[i,j]} \times X_j \quad (4)$$

$$Y_i = - \sum_{j \in U_{-1}} X_j \quad (5)$$

$$Y_i = \sum_{j \in U_0} X_j - X_{sum} \quad (6)$$

The MDA performs two critical matrix decomposition functions. First, it partitions binary (0-1) matrices into submatrices, distributing subblocks of varying densities to the SMA and ISMA for parallel sparse processing, thereby enabling global acceleration. Second, when dealing with extremely large real-valued matrices that may cause cache bottlenecks, the MDA collaborates with the SMA to implement efficient streaming processing.

The ISMA is a method for IM linear operations that was thoroughly validated to address dense graphs. According to Eq. 6, given the nonzero element density D of the coupling matrix J_{couple} (The definition of D and its relationship with graph density are provided in Supplementary Information Section S2.), the classical MVM requires storing N^2+N data elements for multiplication, while the ISMA only needs $(1-D)N^2+N$. For FLOPs, the counts are $N^2+N(N-1)$ and $2N-1+(1-D)N^2$ respectively. This results in reductions in both cache usage and operation count. Moreover, FLOPs in ISMA are lower-latency addition operations. For instance, when processing random graphs with $D=0.99$ (Fig. 3(b)), ISMA achieves 100× and 200× reductions in cache data volume and FLOPs (as shown in Figs. 3(d) and (e)). For denser graphs with $D=0.999$ (Fig. 3(c)), the improvements reach 1000× and 2000× according to Eq. 6 (as shown in Figs. 3(f) and (g)). As a result, with ISMA support, the randomly generated graphs at these two densities achieved speedup factors of over 60× and 400×, respectively (Fig. 3(h)).

Finally, to investigate the relationship between nonzero element density D of the coupling matrix and algorithmic performance, we scanned random matrices with D ranging from $D=0.1$ to 0.9 in 0.1 increments, spanning graphs of 10,000 to 100,000 nodes (Fig. 3(i)). SMA was employed for $D=0.1-0.5$ and ISMA for $D=0.6-0.9$. The experiment results reveal superior acceleration performance of ISMA at higher sparsity (for example, $D=0.1$) and higher density

($D=0.9$), while the weakest performance occurs at $D=0.5$, consistent with theoretical expectations. It is worth mentioning that these results were obtained in non-parallelized execution environments. Thus, even at $D=0.5$, the cache data volume and FLOPs demonstrate a 50% reduction (for 100,000-node graphs), suggesting that additional parallel operations could be implemented within constrained processor runtime resources to further enhance computational speed.

Performance of ultra-large scale solution

To address the domain freezing problem in large-scale spin systems, we introduce an optimized parameter configuration scheme (OPCS) inspired by the pump intensity scheduling technique used in DOPO system for convergence acceleration [10]. Our scheme operates on parameter α , which schedules the gain, and parameters β , which schedules the coupling strength. While the gain parameter α promotes spin amplitudes reaching steady states, an excessively high α can interfere with spin interactions in the Ising network, causing premature convergence and loss of convergence ability. Consequently, after a brief iteration, most spins stabilize, while remaining spins form resistant domains [10, 19]. Our OPCS mitigates this by moderately strengthening coupling constraints on unconverged spins while reducing the influence of gain on mutual coupling. Specifically, after several iterations, α is decreased and β increased, preventing parameter drift caused by converged spins and avoiding mismatch between required and actual parameters during iteration, thereby accelerating Hamiltonian convergence.

Before evaluating the effectiveness of OPCS in reducing the number of iterations, it is essential to verify the deterministic nature of the system. As shown in Fig. 4(a), when starting from the same initial condition and keeping all parameters fixed, the algorithm consistently converges to the same result in an identical or nearly identical iteration steps, even when small noise perturbations are introduced during the iteration process. This confirms that the convergence process is deterministic, which justifies the use of a fixed initial state to enable a fair and controlled comparison between configurations with and without OPCS.

To evaluate the feasibility of OPCS, we designed three parameter configuration schemes under identical initial conditions, and their corresponding simulation results are presented in Fig. 4(b). The simulation methodology can be found in Supplementary Information Section S4. Here, Param1 adopts the typical fixed parameters ($\alpha=0.25$, $\beta=0.2$ throughout); Param2 follows the OPCS by maintaining $\alpha=0.25$, $\beta=0.2$ for the first 100 iterations before switching to $\alpha=0.015$, $\beta=0.31$; while Param3 serves as a control with $\alpha=0.015$, $\beta=0.31$ applied throughout the entire process. Experimental results demonstrate that the OPCS (implemented with Param2)

significantly reduces iteration steps compared to Param1. And to verify that this optimization effect is not solely attributable to the later-stage parameters ($\alpha=0.015$, $\beta=0.31$), we introduced Param3, whose simulation results reveal that prematurely decreasing α and increasing β causes delayed energy reduction and even temporary energy increase during early iterations. Although eventual convergence is achieved, its overall performance remains inferior to Param2, thereby confirming the necessity of phased dynamic parameter regulation in OPCS. Finally, as conclusively demonstrated in Fig. 4(c), we validated the effectiveness of OPCS (Param2) through five independent simulations with uniform initial values, all achieving over 50% iteration reduction compared to the traditional scheme (Param1). This indicates that the OPCS can also effectively achieves the same proportional reductions in both computational time and energy expenditure. It is worth mentioning that the above simulations are designed not limited to the specific physical system presented in this work.

To assess the computational capability of the chip for large-scale and long-duration problem solving, as well as the effectiveness of OPCS, we employed a graph partitioning task on a square lattice with 102,400 spins—representing an ultra-large scale combinatorial optimization problem. Figures 4(d) and (e) show the experimentally measured spin bifurcation and Hamiltonian evolution. The parameters during experiment can be found in Supplementary Information Section S5. Notably, the Hamiltonian drops significantly within the first 200 iterations, indicating that most spins have reached appropriate configurations. Figure 4(g) shows that spin configurations exhibit rapid energy convergence at early stages, followed by domain formation and a subsequent slowdown in the convergence rate. The proposed OPCS effectively alleviates this stagnation, achieving the optimal solution in only 9,330 iterations—substantially fewer than those required by digital simulations under the same conditions. For comparison, the results obtained without OPCS are presented in Supplementary Information Section S5 and are significantly inferior to those achieved with OPCS.

It is worth noting that a relatively low-performance programmable voltage source was used in this study, together with a digital computer for circuit control. Initially, the limited computational speed and resources of the IPIM system prevented it from completing a full iteration at this scale. To address this, we implemented the JA-SID scheme shown in Fig. 3. Experimental results demonstrate that the IPIM with JA-SID effectively solves large-scale problems above 100,000 spins, achieving an optimal balance between resource use and cost. Additionally, OPCS further improves the IPIM’s computational speed and energy efficiency during problem solving. This approach may also inspire other PIM architectures leveraging bifurcation principles.

Additionally, we also evaluated the capability of our IPIM system to handle fully connected problems. For the complete graph shown in Fig. 4(h), zero elements appear exclusively along the diagonal positions in the coupling matrix. Leveraging this regularity, we further optimized ISMA to store only the distribution pattern of zero elements, eliminating the need to retain the full coupling matrix. This led to a four-order-of-magnitude speedup in linear computation compared to conventional algorithms. It also overcomes the computational bottleneck observed in square lattice graph problems, enabling convergence within fewer than seven iterations (Figs. 4(i) and (j)) to a Hamiltonian of -98,800, which corresponds to 96.48% of the global minimum energy. The reduced number of iterations can be attributed to the gradual strengthening of mutual coupling in the Ising network. Initially, spins are near the phase transition point with near-zero amplitudes, leading to weak interactions. As some spins stabilize and reach maximum amplitude, their influence on coupled spins increases. In a fully connected problem, where every spin interacts with all others, this amplification is especially pronounced, rapidly accelerating system convergence. On the other hand, the rapid convergence of the complete graph reflects the competitive relationship between parameters α and β , which is consistent with the theory underlying the OPCS.

Solving of Facebook network problem

To further investigate IPIM’s capability in solving large-scale real-world problems and assess its competitiveness relative to other solvers, we utilized a publicly accessible Facebook user dataset for evaluation[38]. This dataset contains information on 63,392 users and their mutual follow relationships. In the corresponding undirected graph, each user is represented as a vertex, and mutual follow connections between users correspond to 816,886 edges. Considering the practical relevance of the problem, we apply graph partitioning techniques for community detection. For example, the minimum cut problem can separate the user base into two distinct communities, effectively identifying smaller social subgroups. However, since the minimum cut problem is not NP-hard, we focus on solving the MAX-CUT problem to partition users into two groups that maximize the number of mutual follow connections between them, as illustrated in Fig. 5(a). The practical value of this approach lies in exploiting the semantic significance of the edges to inform strategic planning. Specifically, edges represent mutual follow relationships, which means that targeted information or recommendations sent to one community (partition A) are more likely to spread efficiently to the other community (partition B) via inter-community links, thereby greatly enhancing the effectiveness of information diffusion or marketing campaigns.

When employing the IPIM to solve this problem, we observed that the system could not achieve convergence for all spins within a small number of iterations. However, consistent with IPIM's performance on other problems, the solver can rapidly find a suboptimal solution during the initial phase with explosive speed. As illustrated in Fig. 5(b), the energy decreases in the Hamiltonian during the first 50 iterations accounts for over 90% of the total energy reduction. While this does not guarantee a global optimum, it significantly enhances the response speed for large-scale, time-sensitive real-world tasks.

To validate this advantage, we conducted a comparative analysis between SA and IPIM under identical digital processing environments. The SA algorithm employs an exponential cooling schedule with initial temperature $T_0=90,000$ cooling rate 0.9998 per iteration, and stopping temperature $T_1=0.001$. For large scale networks, it utilizes incremental computation of the objective function's change, avoiding costly global recomputation in each iteration. To enable a fair runtime comparison with the IPIM at the same cut value (454,411), we further configured SA without an inner loop, allowing SA to quickly reach the same cut value rather than carefully searching for the global optimum. Digital simulation results demonstrate IPIM's superior performance on large-scale tasks: at the same cut value (Cut value=454411), SA required approximately 200 times longer computation time than IPIM as illustrated in Fig. 5(c). This superiority stems from the inefficiency of SA in local neighborhood sampling, whereas our bifurcation-based IPIM achieves global updates of the variable vector through MVM in each iteration [39]. Although MVM involves a heavier computational load than neighborhood sampling, which updates only a small number of variables per operation, this limitation can be largely alleviated through MVM optimization, thereby accelerating the overall solution process.

Discussion

Compared with other approaches [17, 18, 35, 40], our proposed IPIM achieves a smaller optical component footprint (0.065 mm^2) while maintaining large problem-solving scale and high connectivity, owing to the compactness of silicon photonic devices and the optimization method for PIM linear operation. Our work also presents insights on optimizing linear operations to enhance large-scale problem-solving capabilities. Among integrated photonics schemes, our approach demonstrates a large problem-solving capacity (102,400 spins), along with comprehensive validation on fully connected problems. Although current discrete implementations can support larger spin counts (up to 360,000 spins) [41], chip-scale architectures could potentially attain comparable scalability via massively parallel multi-channel multiplexing. In terms of stability, some recent discrete approaches have achieved long time stable operation using polarization encoding [42], femtosecond lasers [43] or microwave

coupling [44]. In contrast, our IPIM, benefiting from integration and intensity encoding, can stably run for more than 25 hours on square lattice problems and reach the optimal solution. Table 1 shows the comparison of our proposed IPIM with literature [10, 17, 18, 20, 31, 34, 35, 40-45].

Notably, the computational speed of current systems faces two primary constraints. First, the response time of thermo-optic modulators (ranging from microseconds to milliseconds) forms an optical bottleneck in system speed. Second, although JA-SID can reduce computation time for large-scale MVM through operation count reduction and cache optimization, the ADC/DAC conversion remains strongly limited by the sampling rates of programmable voltage source and data acquisition card. However, on a positive note, many integrated photonic devices are capable of performing nonlinear computations within several picoseconds [46, 47]. Moreover, prior works have demonstrated high-speed processing capabilities using high-performance electronic hardware [10, 17], which alleviates this bottleneck to some extent and provides reference for the electronic design. These high-performance electronic/photonic hardware schemes enable a more thorough exploration of the computing potential of JA-SID and IPIM. In Supplementary Information Section S6, we present two representative high-speed electronic hardware implementations. We analyze a projected IPIM system that incorporates JA-SID and high-performance electronics. Due to the optimization by JA-SID, multiplication operations are eliminated, enabling the use of abundant look-up tables (LUTs) in the field-programmable gate array (FPGA), instead of scarce digital signal processing (DSP) slices. By combining the projected high-performance electronics with future high-speed modulators and photodetectors, our analysis predicts that the overall system can achieve a processing speed of 2.40×10^9 spins per second for solving fully connected problems with 100,000 spins. This represents an improvement compared to solving sparse problems within the same series of schemes [17]. Although this predicted speed is somewhat slower than that of discrete systems solving fully connected problems of the same scale [10], our integrated approach uses nearly a hundred times fewer electronic hardware resources (see Table S5 in Section 6 of the Supplementary Information). These saved linear computing resources will help to further scale up the number of spins and increase the system speed. Moreover, this moderate speed disadvantage can be overcome by increasing the ADC/DAC sampling rate. Notably, this JA-SID method can also apply to FPGA-cluster-based digital simulated bifurcation machines [48], easing their linear computation load in large-scale, dense scenarios. To compare this result with other works, we present a radar chart in Fig. 6. Supplementary Information Section S6 provides additional details regarding this result.

Additionally, by simplifying MVM from multiply-accumulate to accumulation, JA-SID becomes inherently suitable for integrated photonic platforms [49-50]. Its implementation requires only a spin encoder (e.g., an intensity modulator) and an accumulator (e.g., a micro-ring integrator) per channel. This minimal component count directly reduces the number of modulators, lowers system power consumption, and enhances parallelism. In spatial optical computing, JA-SID enables accumulation by encoding only sparse spin-amplitude information onto incoherent light, allowing simple incoherent detection at the receiver end. This approach improves pixel utilization and reduces system complexity, as it relies solely on lens-based optical power superposition without the need for complex phase modulation or precise optical alignment. Consequently, this scheme is potentially compatible with high-refresh-rate LED or VCSEL arrays [51], paving the way for higher throughput and lower latency. Meanwhile, the key operating principle of the proposed IPIM is intrinsically wavelength-independent, enabling its extension to a broad wavelength range.

Most PIMs based on nonlinear dynamics, like ours, excel at solving graph partitioning problems, particularly the MAX-CUT problem [10, 13, 17, 19-21, 24, 35, 37], where the coupling matrix elements are typically discrete values. This makes them well-suited for our proposed JA-SID framework. However, such PIMs are rarely applied to other types of combinatorial optimization problems, such as the traveling salesman problem (TSP) or Boolean satisfiability problem (SAT). These problems involve coupling matrices with continuous elements in the range from 0 to 1, leading to non-uniform interactions that increase the risk of getting trapped in local optima for bifurcation-based solvers, and can only be effectively addressed in JA-SID by introducing low-bit quantization methods that decompose weighted matrices into 0-1 matrices (the processing details are provided in Supplementary Information Section S7). Moreover, the effectiveness of JA-SID diminishes under such coupling matrices, and only through explicit sparsity and decomposition-parallelism techniques can the quadratic growth bottleneck in linear operations be mitigated. This limitation presents a major challenge for current PIM research. Therefore, an important next step may lie in expanding the range of solvable problem types and further optimizing their corresponding linear operations.

Materials and Methods

Silicon Photonic Chip

Silicon photonic devices offer significant advantages due to their high compatibility with CMOS processes, enabling simplified integration and allowing the avoidance of the cost and performance penalties typically associated with complex design and fabrication in IPIMs. The

proposed IPIM chip is fabricated on a 220 nm silicon-on-insulator (SOI) wafer with a 2 μm thick buried oxide (BOX) layer, as shown in Fig. 1(e). The detailed process can be found in Supplementary Information Section S8. For light coupling, an inverse taper edge coupler is utilized. A Ge absorption region is selectively grown on a 130 nm etched Si ridge waveguide. The top of Ge region is heavily doped with phosphorus and the corresponding bottom Si is doped with boron to form a vertical PIN junction. For the Si modulator, the titanium nitride (TiN) heater is deposited above the Si waveguide for phase accumulation. Figures 1(f) to (g) present the microscopic images of our proposed IPIM chip. Figure 1(h) illustrates the electro-optic response curve of the cascaded Si modulator and Ge photodetector, which aligns well with the nonlinear dynamical Eq. described in Eq. 2. This indicates that both the modulator and photodetector can function effectively on the integrated chip, thereby enabling the subsequent solving process to proceed smoothly.

Experiment Setup

Before testing, the silicon photonic chip was optoelectronically packaged. During the experiment, light from a laser source was coupled into the chip via an optical fiber. The optical signal passed through an MZM, where modulation was achieved through the thermo-optic effect. The modulated optical signal was then detected by an on-chip germanium photodetector and converted into an analog electrical signal. This signal was transmitted to a computer via an ADC for linear computation. The computed result was subsequently fed back to the modulator through a DAC, forming a closed opto-electronic loop. Time-division multiplexing was employed to sequentially update all spin values using a single modulator-detector pair.

Data availability

The data that support the findings of this study are available from the corresponding authors on reasonable request.

References

1. Mohseni, N., McMahon, P. L. & Byrnes, T. Ising machines as hardware solvers of combinatorial optimization problems. *Nature Reviews Physics* **4**, 363-379 (2022).
2. Böhm, F. et al. Noise-injected analog Ising machines enable ultrafast statistical sampling and machine learning. *Nature Communications* **13**, 5847 (2022).
3. Bybee, C. et al. Efficient optimization with higher-order Ising machines. *Nature Communications* **14**, 6033 (2023).
4. Dan, A. et al. Clustering approach for solving traveling salesman problems via Ising model based solver. 2020 57th ACM/IEEE Design Automation Conference (DAC). San Francisco, CA, USA: IEEE, 2020, 1-6.

5. Si, J. et al. Energy-efficient superparamagnetic Ising machine and its application to traveling salesman problems. *Nature Communications* **15**, 3457 (2024).
6. Gao, Y. et al. Photonic Ising machines for combinatorial optimization problems. *Applied Physics Reviews* **11**, 041307 (2024).
7. Brusco, M. J. et al. On Ising models and algorithms for the construction of symptom networks in psychopathological research. *Psychological Methods* **24**, 735-753 (2019).
8. Lobanov, M. Y. & Galzitskaya, O. V. The Ising model for prediction of disordered residues from protein sequence alone. *Physical Biology* **8**, 035004 (2011).
9. Yue, W. S. et al. A scalable universal Ising machine based on interaction-centric storage and compute-in-memory. *Nature Electronics* **7**, 904-913 (2024).
10. Honjo, T. et al. 100,000-spin coherent Ising machine. *Science Advances* **7**, eabh0952 (2021).
11. Kirkpatrick, S., Gelatt, C. D. & Vecchi, M. P. Optimization by simulated annealing. *Science* **220**, 671-680 (1983).
12. King, A. D. et al. Coherent quantum annealing in a programmable 2,000 qubit Ising chain. *Nature Physics* **18**, 1324-1328 (2022).
13. Inagaki, T. et al. A coherent Ising machine for 2000-node optimization problems. *Science* **354**, 603-606 (2016).
14. Pierangeli, D., Marcucci, G. & Conti, C. Large-scale photonic Ising machine by spatial light modulation. *Physical Review Letters* **122**, 213902 (2019).
15. Moy, W. et al. A 1,968-node coupled ring oscillator circuit for combinatorial optimization problem solving. *Nature Electronics* **5**, 310-317 (2022).
16. Cai, F. X. et al. Power-efficient combinatorial optimization using intrinsic noise in memristor Hopfield neural networks. *Nature Electronics* **3**, 409-418 (2020).
17. Li, Z. H. et al. Scalable on-chip optoelectronic Ising machine utilizing thin-film lithium niobate photonics. *ACS Photonics* **11**, 1703-1714 (2024).
18. Zhu, H. H. et al. A dynamically programmable quantum photonic microprocessor for graph computation. *Laser & Photonics Reviews* **18**, 2300304 (2024).
19. Cen, Q. Z. et al. Large-scale coherent Ising machine based on optoelectronic parametric oscillator. *Light: Science & Applications* **11**, 333 (2022).
20. Böhm, F., Verschaffelt, G. & Van der Sande, G. A poor man's coherent Ising machine based on opto-electronic feedback systems for solving optimization problems. *Nature Communications* **10**, 3538 (2019).

21. Marandi, A. et al. Network of time-multiplexed optical parametric oscillators as a coherent Ising machine. *Nature Photonics* **8**, 937-942 (2014).
22. Fan, Z. Y. et al. Photonic Hopfield neural network for the Ising problem. *Optics Express* **31**, 21340-21350 (2023).
23. Yamashita, H. et al. Low-rank combinatorial optimization and statistical learning by spatial photonic Ising machine. *Physical Review Letters* **131**, 063801 (2023).
24. McMahon, P. L. et al. A fully programmable 100-spin coherent Ising machine with all-to-all connections. *Science* **354**, 614-617 (2016).
25. Clements, W. R. et al. Optimal design for universal multiport interferometers. *Optica* **3**, 1460-1465 (2016).
26. Harris, N. C. et al. Linear programmable nanophotonic processors. *Optica* **5**, 1623-1631 (2018).
27. Huang, C. R. et al. A silicon photonic–electronic neural network for fibre nonlinearity compensation. *Nature Electronics* **4**, 837-844 (2021).
28. Zhang, J. H. et al. Compact, efficient, and scalable nanobeam core for photonic matrix-vector multiplication. *Optica* **11**, 190-196 (2024).
29. Miller, D. A. B. Perfect optics with imperfect components. *Optica* **2**, 747-750 (2015).
30. Prabhu, M. et al. Accelerating recurrent Ising machines in photonic integrated circuits. *Optica* **7**, 551-558 (2020).
31. Huang, J. Y., Fang, Y. S. & Ruan, Z. C. Antiferromagnetic spatial photonic Ising machine through optoelectronic correlation computing. *Communications Physics* **4**, 242 (2021).
32. Pierangeli, D., Marcucci, G. & Conti, C. Adiabatic evolution on a spatial-photonic Ising machine. *Optica* **7**, 1535-1543 (2020).
33. Fang, Y. S., Huang, J. Y. & Ruan, Z. C. Experimental observation of phase transitions in spatial photonic Ising machine. *Physical Review Letters* **127**, 043902 (2021).
34. Veraldi, D. et al. Fully programmable spatial photonic Ising machine by focal plane division. *Physical Review Letters* **134**, 063802 (2025).
35. Okawachi, Y. et al. Demonstration of chip-based coupled degenerate optical parametric oscillators for realizing a nanophotonic spin-glass. *Nature Communications* **11**, 4119 (2020).
36. Heim, B. et al. Quantum versus classical annealing of Ising spin glasses. *Science* **348**, 215-217 (2015).
37. Hamerly, R. et al. Experimental investigation of performance differences between coherent Ising machines and a quantum annealer. *Science Advances* **5**, eaau0823 (2019).

38. Rossi, R. A. & Ahmed, N. K. The network data repository with interactive graph analytics and visualization. Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence. Austin, TX, USA: AAAI Press, 2015, 4292-4293.
39. Goto, H., Tatsumura, K. & Dixon, A. R. Combinatorial optimization by simulating adiabatic bifurcations in nonlinear Hamiltonian systems. *Science Advances* **5**, eaav2372 (2019).
40. Wu, B. et al. A monolithically integrated optical Ising machine. *Nature Communications* **16**, 4296 (2025).
41. Yao, J. Z., Zhu, R. W. & Yu, J. J. Enhanced, fully connected 360 000-spin spatial photonic Ising machine. *Physical Review Applied* **24**, 044087 (2025).
42. Quinn, L. et al. Coherent Ising machine based on polarization symmetry breaking in a driven Kerr resonator. *Nature Communications* **17**, 2100 (2026).
43. Wei, H. et al. A versatile coherent Ising computing platform. *Light: Science & Applications* **15**, 74 (2026).
44. Hu, Z. W. et al. Programmable optoelectronic Ising machine for optimization of real-world problems. *Light: Science & Applications* **15**, 6 (2026).
45. Al-Kayed, N. et al. Programmable 200 GOPS Hopfield-inspired photonic Ising machine. *Nature* **648**, 576-584 (2025).
46. Lischke, S. et al. Ultra-fast germanium photodiode with 3-dB bandwidth of 265 GHz. *Nature Photonics* **15**, 925-931 (2021).
47. Xu, M. Y. et al. High-performance coherent optical modulators based on thin-film lithium niobate platform. *Nature Communications* **11**, 3911 (2020).
48. Tatsumura, K., Yamasaki, M. & Goto, H. Scaling out Ising machines using a multi-chip architecture for simulated bifurcation. *Nature Electronics* **4**, 208-217 (2021).
49. Ferrera, M. et al. On-chip CMOS-compatible all-optical integrator. *Nature Communications* **1**, 29 (2010).
50. Tang, R. et al. Waveguide-multiplexed photonic matrix–vector multiplication processor using multiport photodetectors. *Optica* **12**, 812-820 (2025).
51. Song, A. et al. Low-power scalable multilayer optoelectronic neural networks enabled with incoherent light. *Nature Communications* **15**, 10692 (2024).

Acknowledgements

This work was supported by the National Key Research and Development Program of China (2024YFB2808600); National Natural Science Foundation of China (62405035, W2521113); Fundamental Research Funds for the Central Universities (2024IAIS-QN018); Natural Science

Foundation of Chongqing Municipality (CSTB2025NSCQ-GPX0737, CX2025019); New Chongqing YC Project (CSTB2024YCJH-KYXM0105); Singapore National Research Foundation under Competitive Research Program (No. NRF CRP24-2020-0003) and funding from a Ministry of Education Academic Research Fund Tier 1 grant. G. C. thanks Guy Verschaffelt and Guy Van Der Sande for enlightening discussions.

Author contributions

G. C. conceived and proposed the project. G. C. and Z. Z. designed the experiments. G. C., A. P. and Y. G. designed this chip and its packaging module. Z. Z. and G. C. performed the experiments and analyzed the data. Z. Z., G. C., J. G. and J. L. performed the simulation and calculation. Z. Z. and G. C. wrote the manuscript with inputs from all authors. G. C. and A. J. D. supervised this project.

Competing interests

The authors declare no competing interests.

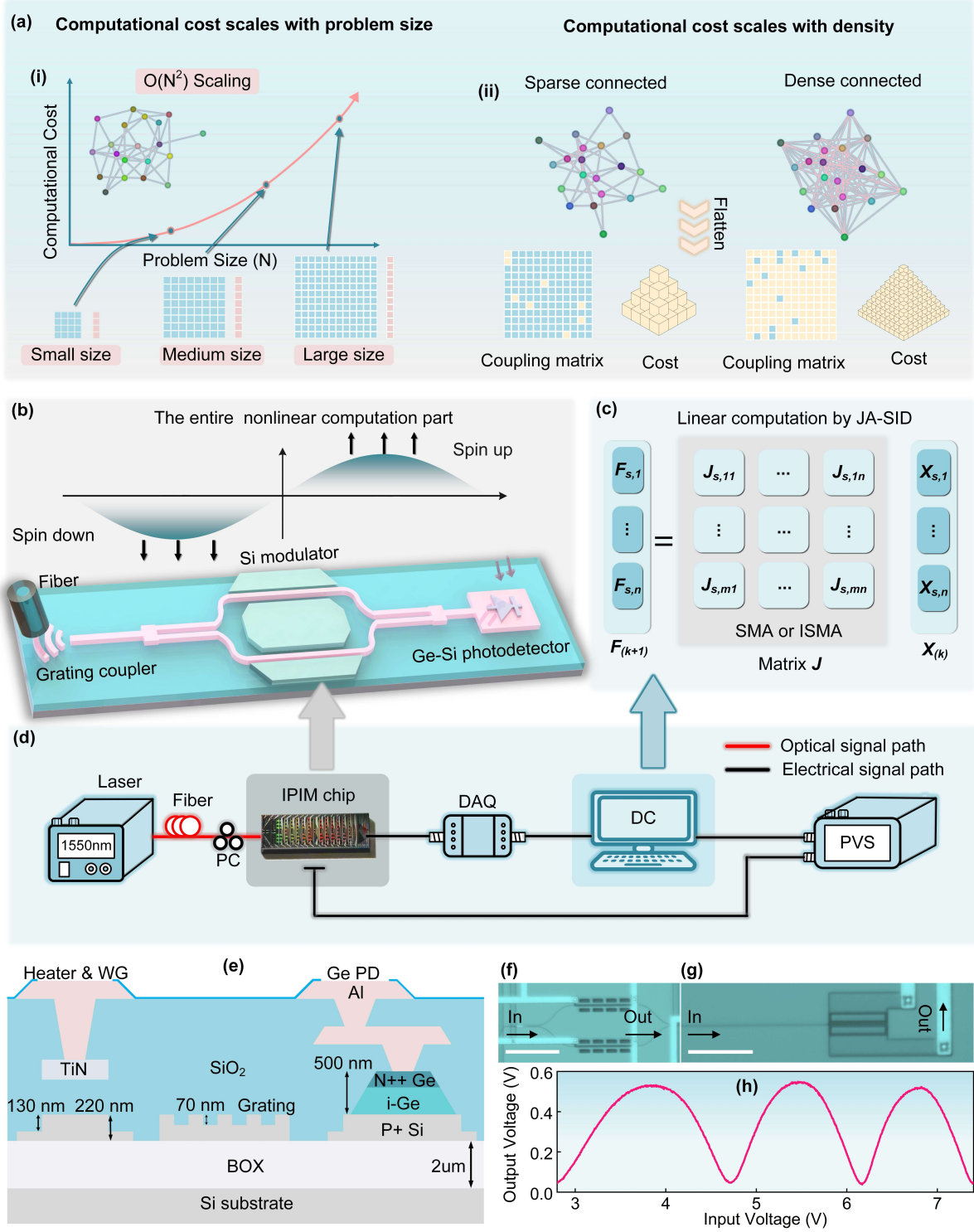


Fig. 1. Working principle of our proposed integrated photonic Ising machine (IPIM). (a) Two challenges in solving large-scale problems. (b) The architecture of the IPIM chip. (c) Schematic of the linear computation by joint algorithm with sparsity, inverse sparsity, and decomposition operations (JA-SID). (d) The framework of the IPIM. (e) Schematic cross-section of the IPIM chip. (f) and (g) show microscope images of the silicon (Si) modulator and germanium (Ge)-Si photodetector, respectively, with scale bars of 100 and 40 μm . (h) shows the optoelectronic nonlinear response curves of both modulator and detector. SMA: sparse matrix algorithm; ISMA: inverse sparse matrix algorithm; PC: Polarization Controller; DAQ: Data Acquisition card; PVS: Programmable Voltage Source; DC: Digital Computer.

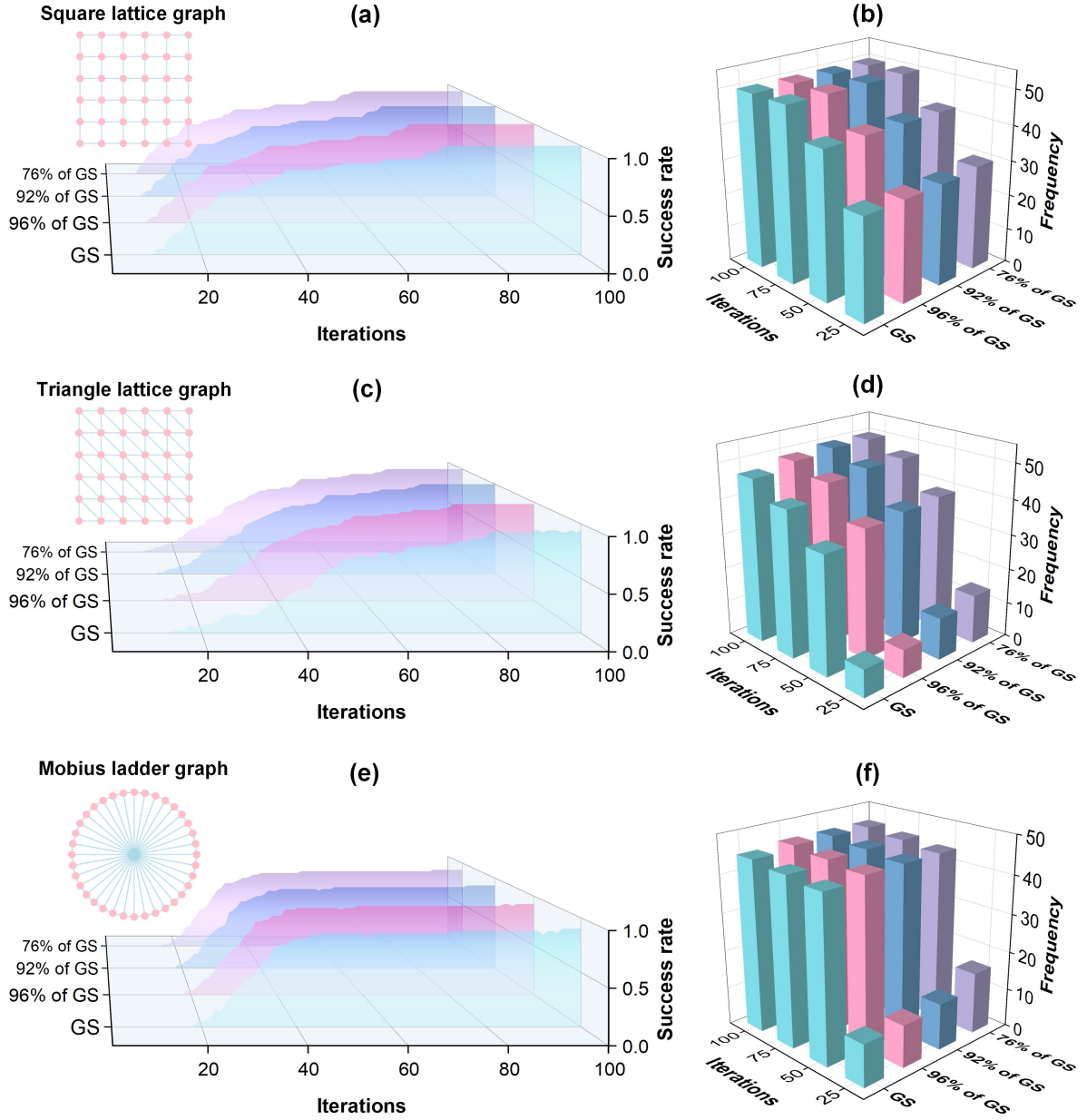


Fig. 2. Demonstration of solving the MAX-CUT problem for three different topologies. (a) and (b) Success rates and frequencies of optimal solutions within 100 iterations for the MAX-CUT problem on a 100-spin square lattice graph under four conditions: Hamiltonian greater than or equal to the ground state, 96% of the ground state, 92% of the ground state, and 76% of the ground state. (c) and (d) Success rates and frequencies of optimal solutions within 100 iterations for the MAX-CUT problem on a 144-spin triangular lattice graph under the same four conditions. (e) and (f) Success rates and frequencies of optimal solutions within 100 iterations for the MAX-CUT problem on a 100-spin Möbius ladder graph under the same four conditions.

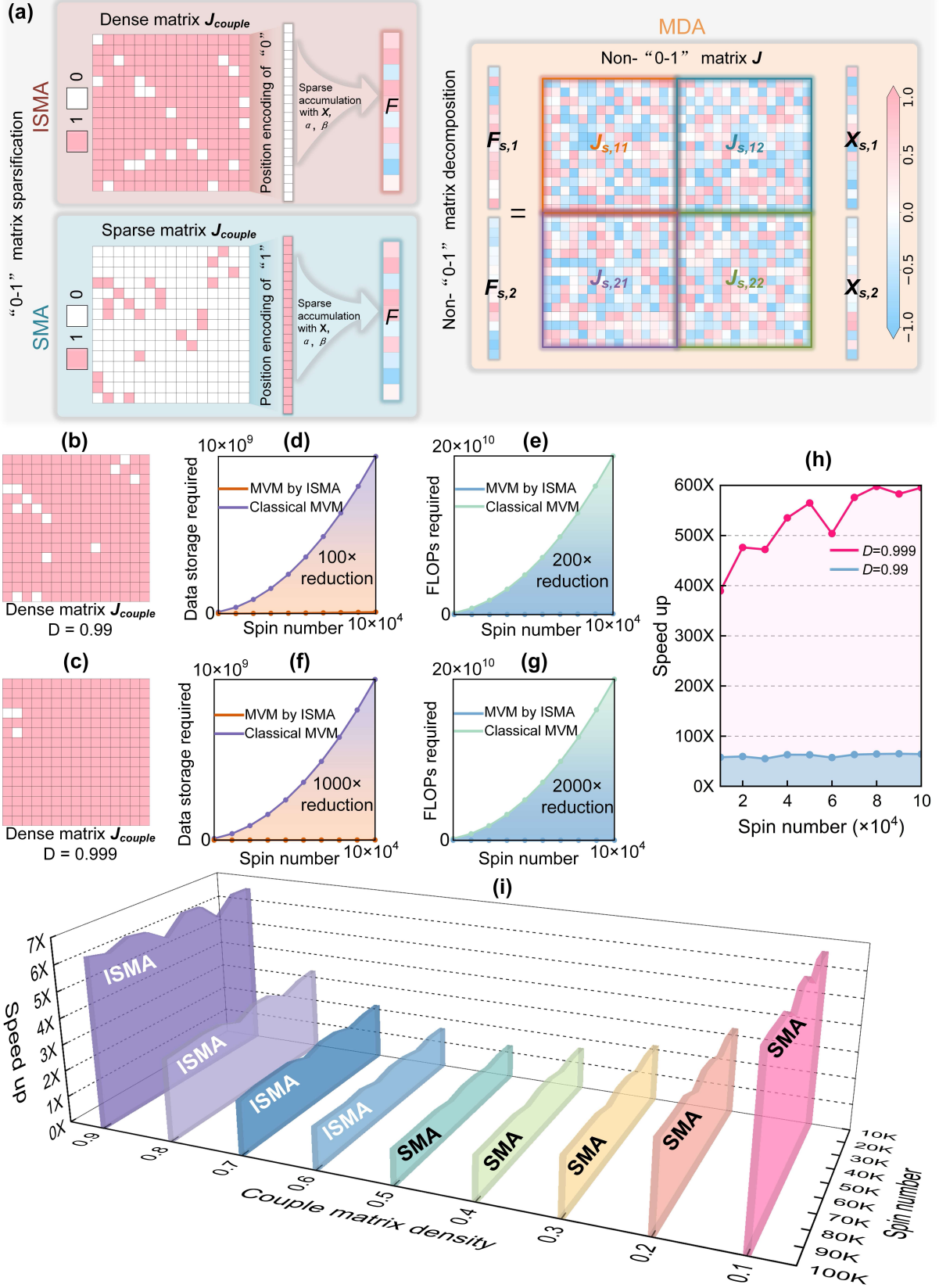


Fig. 3. Performance of the joint algorithm with sparsity, inverse sparsity, and decomposition operations (JA-SID). (a) Schematic of the JA-SID. (b) and (c) are the schematics of dense matrix with $D=0.99$ and 0.999 , respectively. (d) and (e) compare the cache data volume and floating-point operations required by classical MVM and ISMA when executing random graph problems with density $D=0.99$. (f) and (g) compare the cache data

volume and floating-point operations required by classical MVM and ISMA when executing random graph problems with density $D=0.999$. **(h)** Demonstration of the computational speedup ratio achieved by ISMA compared to classical MVM for random graph at $D=0.99$ and $D=0.999$. **(i)** Demonstration of the speedup ratio of ISMA and SMA compared to classical MVM when scanned over the range $D=0.1$ to 0.9 with a step size of 0.1 .

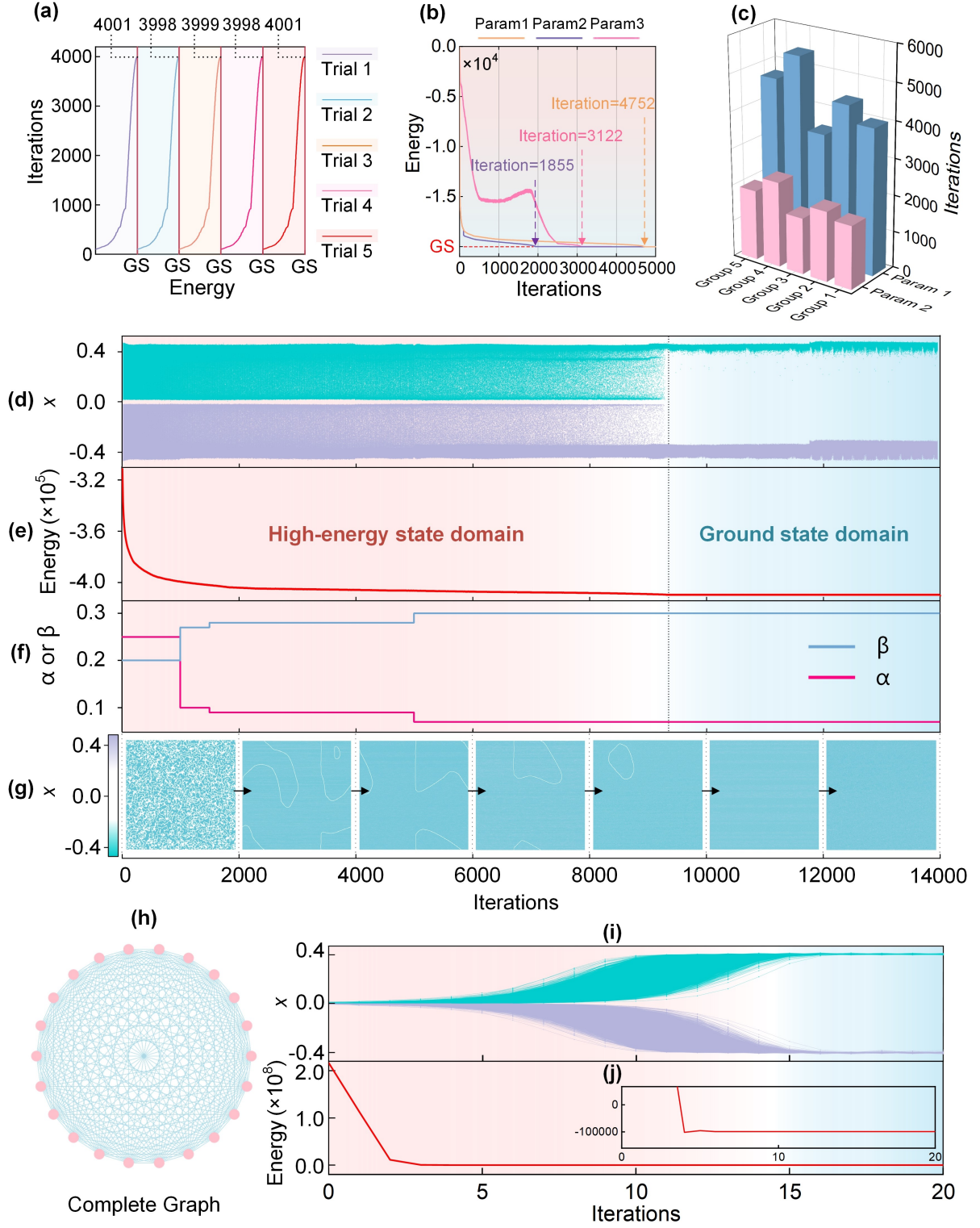


Fig. 4. Performance of the optimized parameter configuration scheme (OPCS) and solving of 102,400 spins MAX-CUT problem. (a) Demonstration of the high consistency of the PIM's iterative behavior under identical initial conditions and fixed parameters. In each trial, the noise amplitude is set to 0.2% of the maximum spin amplitude (0.5), and a different random seed for noise is used. (b) Comparison of iteration counts under varying parameter configurations with fixed initial conditions. (c) Comparison of five experimental sets using traditional and optimized parameter configurations, with average iteration counts calculated from five independent runs under fixed initial conditions. (d) Bifurcation phenomenon observed in the 102,400-spin system during the solving

process. **(e)** System energy converges to the ground state energy upon completion of bifurcation, and **(f)** its corresponding optimization parameter configuration scheme (OPCS) during this process. **(g)** Snapshots of the spin configurations at iteration counts of 10, 2000, 4000, 6000, 8000, 10000, and 12000, respectively. **(h)** Schematic of the complete graph. **(i)** Spin bifurcation diagram for the MAX-CUT problem on a complete graph. **(j)** Hamiltonian descent diagram of the MAX-CUT problem for the complete graph, with the inset showing the Hamiltonian of the converged state to be -98800.

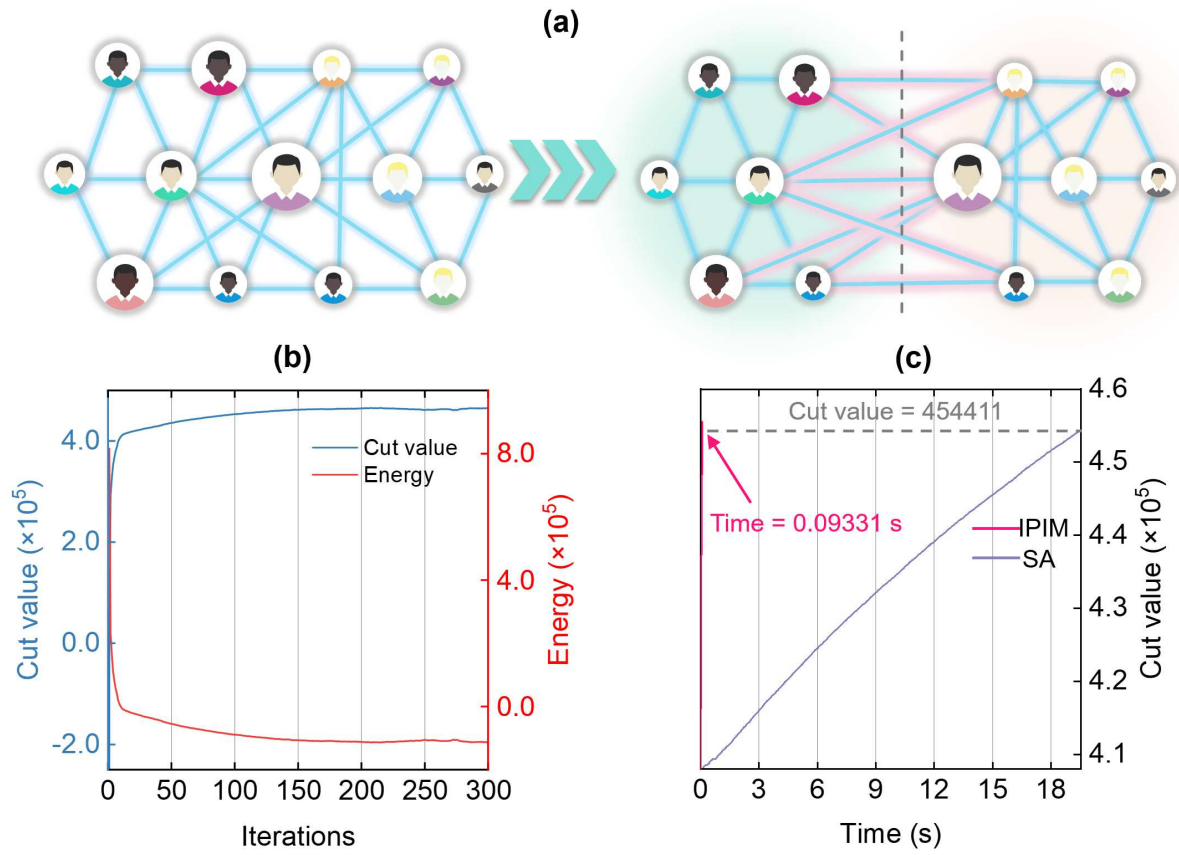


Fig. 5. Results of Facebook community partition and comparison with simulated annealing (SA). (a) Illustration of the MAX-CUT partitioning process applied to a social network, resulting in two maximally interconnected communities. (b) Results for the Facebook network with 63,392 users show that over 300 iterations, the system energy steadily decreases while the cut value increases. (c) Comparison of the computational time required for IPIM and SA in the digital environment to reach the target cut value of 454,411, respectively.

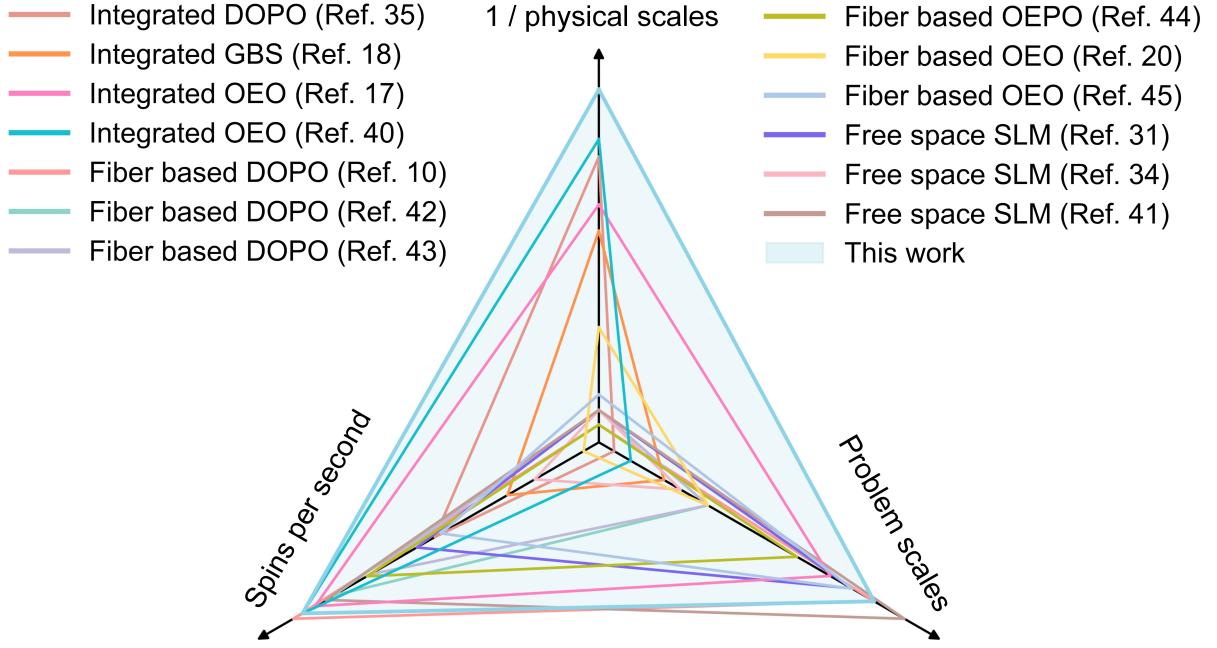


Fig. 6. Performance comparison between our proposed IPIM and state-of-the-art works. For clarity, logarithmic scaling and normalization were applied, and reciprocal axes were used to ensure that higher values consistently indicate better performance.

Table 1 Comparison of our proposed silicon photonics based IPIM with literature. DOPO: degenerate optical parametric oscillator; GBS: Gaussian boson sampling; SLM: spatial light modulator; OEO: optoelectronic oscillator; TFLN: thin film lithium niobate; SPS: spins per second; MVM: matrix-vector multiplication.

Principle	Problem scale	Connection type	Optical component Footprint	Function of optical components	System speed	Success rate	Platform	Ref.
DOPO	2	Neighbor	1.56 mm ² ^a	Nonlinear computation	800 KSPS ^a	/	Si ₃ N ₄ ^c	[35]
DOPO	100,000	Fully connected	0.5 m ² ^a	Nonlinear computation	4.16 GSPS ^a	/	Fiber system	[10]
DOPO	100	Neighbor	0.25 m ² ^a	Nonlinear computation	366 MSPS ^a	/	Fiber system	[42]
DOPO	100	Densely connected	0.25 m ² ^a	Nonlinear computation	40 MSPS ^a	/	Fiber system	[43]
GBS	16	Sparsely connected	52 mm ² ^a	MVM & Probability sampling	10 KSPS ^a	/	CMOS	[18]
SLM	40,000	Sparsely connected	0.25 m ² ^a	Hamiltonian calculation	2.4 MSPS ^a	/	Free space	[31]
SLM	32	Sparsely connected	0.25 m ² ^a	Hamiltonian calculation	1.92 KSPS ^a	/	Free space	[34]
SLM	360,000	Fully connected	0.25 m ² ^a	Hamiltonian calculation	581 MSPS ^a	/	Free space	[41]
OEPO	4,096	Fully connected	0.5 m ² ^a	Nonlinear computation	48.8 MSPS ^a	/	Fiber system	[44]
OEO	16,384	Neighbor	15 mm ² ^a	Nonlinear computation	1.12 GSPS ^a	/	TFLN ^c	[17]
OEO	100	Neighbor	0.005 m ² ^a	Nonlinear computation	100 SPS ^a	90% ^d 34% ^e 52% ^f	Fiber system	[20]
OEO	41,209	Neighbor	0.12 m ² ^a	Nonlinear computation & MVM	550 KSPS ^a 21.3 GSPS ^b	/	Fiber system	[45]
OEO	4	Densely connected	>0.68 mm ² ^a	Linear & Nonlinear computation	2.34 GSPS ^a	/	CMOS	[40]
OEO	102,400	Fully connected	0.065 mm ²	Nonlinear computation	2.40 GSPS ^b	100% ^d 94% ^e 90% ^f	CMOS	Ours

^a Calculated value

^b Predicted value. The predicted values in our work originate from a prediction (see Supplementary Information Section S6) scheme based on carrier-dispersion modulators and high-performance FPGAs.

^c Heterogeneously integrated scheme

^d 100-spin square lattice graph success rate

^e 100-spin Möbius ladder graph success rate

^f 144-spin triangular lattice graph success rate