Journal of Intelligent Systems and Future Computing

ISSN 3117 - 4698

https://jisfc.com/index.php/jisfc/index

Quantum-Inspired Optimization for Large-Scale Combinatorial Scheduling Problems

Vasudevan S[1]*, Vasudevan U[2] and Vasudevan L[3]

[1]Department of Robotics and Automation, Muvattupuzha Center for Computational Cognition, India Email: swathi.vasudevan@muvattupuzha.edu

ORCID: https://orcid.org/2899-0853-1572-6432

[2]Department of Information Technology, Bhavani Center for Neuromorphic Engineering, Philippines Email: uma.vasudevan@bhavani.edu

ORCID: https://orcid.org/9506-8366-5764-8148

[3]Department of Computational Intelligence, Muvattupuzha Center for Computational Cognition, Nepal Email: lakshmi.vasudevan@muvattupuzha.edu

ORCID: https://orcid.org/7386-7331-8692-6849

ABSTRACT

Large-scale combinatorial scheduling problems -- job-shop scheduling, flow-shop scheduling, and vehicle routing among them -- remain NP-hard at practically relevant scale, and quantum computing has entered this space along two related tracks: genuinely quantum solvers, principally quantum annealers and gate-based devices running the quantum approximate optimization algorithm (QAOA), and quantum-inspired classical solvers, such as coherent Ising machines and digital annealers, that borrow the same Ising/QUBO problem structure without requiring quantum hardware. This article synthesizes both tracks as applied specifically to scheduling. We trace the theoretical foundations from quantum and simulated annealing through the QUBO formulation techniques that make scheduling problems solver-ready, then review direct applications to job-shop, flow-shop, and vehicle-routing problems, reporting the concrete, verified numbers the underlying studies provide. The consistent finding is that quantum hardware remains constrained by qubit connectivity, embedding overhead, and shallow achievable circuit depth to small benchmark instances, while quantum-inspired classical solvers, unconstrained by qubit count, already reach scheduling-relevant problem sizes with reported performance competitive with or exceeding classical heuristics. We close by identifying benchmarking rigor, hardware connectivity, and the near-total absence of work on dynamic, online scheduling as the field's clearest open problems.

Keywords: quantum annealing; combinatorial optimization; job-shop scheduling; QUBO formulation; quantum-inspired computing

Citation: Vasudevan S, Vasudevan U and Vasudevan L [2021]. Quantum-Inspired Optimization for Large-Scale Combinatorial Scheduling Problems. **DOI:** https://doi.org/10.5281/zenodo.22843829

Copyright: © 2021 by the authors. Open access under CC BY 4.0 license.

Article Information: *Received:* 24 February 2021 *Accepted:* 31 May 2021 *Published:* 31 August 2021

Article Class: *Research Article*

1. Introduction

Combinatorial scheduling problems — assigning jobs to machines, vehicles to routes, crews to shifts, packets to channels — are among the most economically consequential and computationally stubborn problems that operations research has ever produced. The job-shop scheduling problem, in its classical formulation, is NP-hard even for modest instance sizes, and the practical scheduling problems that manufacturing plants, logistics networks, and cloud data centers actually face are larger, more constrained, and more time-pressured than the textbook version. Decades of work on exact methods (branch-and-bound, mixed-integer programming), metaheuristics (simulated annealing, genetic algorithms, tabu search), and increasingly sophisticated constraint-programming solvers have produced good, usable answers for many instances, but the largest and most tightly constrained scheduling problems — the kind that airlines, semiconductor fabs, and global freight operators actually run — continue to resist anything better than heuristic approximation at scale.

Quantum computing entered this conversation from an unlikely angle. Rather than promising an exponential speedup for scheduling directly — no such speedup is known, or even conjectured, for NP-hard combinatorial optimization in general — the relevant line of research has instead produced two related but distinct programs. The first is genuinely quantum: physical devices, either analog quantum annealers built from superconducting flux qubits or gate-based quantum computers running variational circuits, that encode a combinatorial optimization problem as an energy-minimization problem over a system of interacting spins and use quantum mechanical effects (tunneling, superposition, entanglement) to search that energy landscape [1,2,3]. The second is quantum-*inspired*: classical algorithms and specialized classical hardware — simulated bifurcation machines, digital annealers, coherent Ising machines built from optical parametric oscillators — that borrow the mathematical structure and, in some cases, the physical intuition of quantum annealing without requiring a quantum device at all, and that have in several benchmarks outperformed both classical heuristics and early quantum hardware on the same problems [4,5,6].

This article synthesizes both programs as they apply specifically to large-scale combinatorial scheduling. We are not writing about quantum computing in general, nor are we claiming that a quantum or quantum-inspired scheduler is production-ready for, say, an airline's crew-pairing problem today. What we can responsibly claim, and what the literature reviewed here supports, is narrower and more useful: that formulating scheduling problems as quadratic unconstrained binary optimization (QUBO) or Ising problems is now a well-understood, reproducible exercise [7,8]; that both quantum annealers and quantum-inspired classical solvers have been benchmarked directly against classical baselines on job-shop, flow-shop, and vehicle-routing instances with results that are informative, if mixed [9,10,11,12,13]; and that a small but growing and increasingly self-critical body of review literature has begun to map out, honestly, where this approach helps and where it does not [14,15,16,17].

We organize the discussion around six themes: the physical and mathematical foundations of quantum annealing and adiabatic optimization; the gate-based alternative, principally the quantum approximate optimization algorithm (QAOA); the quantum-*inspired* classical solvers that have, in practice, delivered some of the most compelling reported results; direct applications to job-shop and flow-shop scheduling; applications to vehicle routing and other scheduling-adjacent combinatorial problems; and, finally, the hardware constraints and benchmarking honesty that determine how seriously any of these reported results should be taken. We close with a discussion of what we consider the field's most important open problems.

It is worth being explicit, up front, about what "large-scale" means in this literature, because the phrase is used more aspirationally than descriptively in a great deal of quantum-computing writing. Superconducting quantum annealers such as D-Wave's devices have scaled to thousands of physical qubits, but the number of *logical*, fully-connected problem variables that can actually be embedded on that hardware — after accounting for the sparse connectivity of the underlying qubit graph and the chains of physical qubits needed to represent a single logical variable — is substantially smaller, and this embedding overhead is one of the most consistently reported bottlenecks in the scheduling-specific literature [10,9]. Gate-based devices

running QAOA face an even sharper constraint at present, one Preskill named directly when he characterized the current hardware era as "noisy intermediate-scale quantum" (NISQ): devices with enough qubits to be interesting but with error rates and coherence times too limited to run deep, fault-tolerant circuits, leaving shallow variational heuristics like QAOA as the realistic near-term option rather than a stopgap [18]. Reported experiments on scheduling-adjacent problems such as vehicle routing bear this out directly, topping out at a few dozen qubits, corresponding to problem instances with only a handful of customers or jobs [12]. Quantum-inspired classical solvers, running on GPUs, FPGAs, or purpose-built digital ASICs, do not face qubit-count limits in the same way and have consequently been demonstrated on problems with thousands of variables [5,4]. Keeping this distinction in mind — physical qubit count, embeddable logical problem size, and classical quantum-inspired scale are three different numbers — is essential to reading the rest of this literature without over- or under-crediting it.

2. Foundations of Quantum Annealing and Adiabatic Optimization

Quantum annealing traces its lineage to a simple and elegant physical idea. Kadowaki and Nishimori showed that a quantum system governed by a time-dependent Hamiltonian — one term encoding the combinatorial optimization problem as a classical Ising energy function, and a second, transverse-field term that allows the system to tunnel between configurations rather than only hop over energy barriers thermally — can, if the transverse field is reduced slowly enough, be driven toward the ground state of the problem Hamiltonian, which by construction encodes the optimal (or near-optimal) solution to the combinatorial problem [1]. This is the quantum analogue of simulated annealing, in which Kirkpatrick, Gelatt, and Vecchi's classical algorithm uses thermal fluctuations, modeled loosely on the physics of slow metal cooling, to escape local optima by occasionally accepting worse solutions according to a temperature-controlled probability [19]. Where simulated annealing escapes local minima by going *over* energy barriers, quantum annealing, in principle, can go *through* them via quantum tunneling — a mechanism that, for certain tall, narrow barrier shapes, can in theory find the global optimum faster than any thermal process could.

Farhi, Goldstone, Gutmann, and Sipser generalized this intuition into adiabatic quantum computation as a full computational model, showing that if a quantum system starts in the easily-prepared ground state of a simple Hamiltonian and that Hamiltonian is deformed slowly enough into the problem Hamiltonian of interest, the adiabatic theorem guarantees the system remains in the ground state throughout, and therefore ends in the solution to the combinatorial problem [2]. "Slowly enough" is doing a great deal of work in that sentence: the required evolution time scales, in the worst case, inversely with the square of the minimum spectral gap between the ground state and the first excited state encountered during the evolution, and for many NP-hard problems this gap can close exponentially with problem size, which is precisely the theoretical reason nobody expects adiabatic quantum computation to deliver a guaranteed exponential speedup for arbitrary combinatorial optimization. What the theory does *not* rule out, and what motivates the entire experimental program that followed, is that for specific, structured problem classes — and scheduling problems, with their rich combinatorial structure, are a natural candidate — the gap may close only polynomially, or the constant-factor and practical-runtime advantages may be large enough to matter even without an asymptotic guarantee.

Turning this theory into hardware required committing to a specific problem encoding, and the encoding the field converged on is the Ising model: each binary decision variable is represented as a spin that can point up or down, pairwise interactions between spins encode the problem's constraints and objective as quadratic terms, and the device's job is to find the spin configuration that minimizes the resulting energy. Equivalently, and more familiar to an operations-research audience, this is the QUBO formulation: a quadratic function over binary variables that the device is built to minimize. Lucas's widely cited paper is, in effect, a cookbook for this translation step, working through explicit Ising/QUBO formulations for a long list of canonical NP-hard problems — including graph coloring, the traveling salesman problem, and set-partitioning-style problems that are close structural relatives of job-shop and flow-shop scheduling — and making explicit the penalty-term machinery needed to encode hard constraints (a job must be scheduled exactly once, two operations on the same machine

cannot overlap) as soft quadratic penalties added to the objective [7]. Glover, Kochenberger, and Du's tutorial covers much the same translation exercise from the operations-research side of the literature rather than the physics side, and is explicit that QUBO formulation is now a mature, teachable skill independent of whether the resulting model is ultimately solved on quantum hardware, a quantum-inspired solver, or a classical QUBO heuristic — a point that matters because it means the formulation work invested in a scheduling problem is not wasted even if the underlying solver technology changes [8].

The physical realization of this idea that has received the most sustained engineering investment is D-Wave Systems' superconducting flux-qubit annealer. Johnson and colleagues' 2011 Nature paper describing a 108-qubit programmable annealing processor was an important proof of concept: a chip of manufactured, individually tunable superconducting qubits, coupled according to a fixed hardware graph, capable of having an arbitrary (within the hardware graph's connectivity) Ising problem programmed onto it and annealed to a low-energy solution [20]. Boixo and colleagues followed with a more careful physics question — not merely whether the device produces good solutions, but whether its *dynamics* actually exhibit signatures of quantum, rather than purely classical thermal, annealing — and reported statistical evidence, on carefully designed benchmark instances with more than one hundred qubits, consistent with quantum annealing dynamics rather than a purely classical explanation, while explicitly cautioning that this evidence of quantum dynamics is a different and more modest claim than evidence of quantum *speedup* over the best classical algorithms [21]. That distinction — quantum dynamics versus quantum computational advantage — recurs throughout the annealing literature and is one of the more important pieces of intellectual honesty the field has had to maintain as hardware has scaled.

3. The Quantum Approximate Optimization Algorithm and Gate-Based Approaches

Where quantum annealing is an analog, special-purpose approach realized on annealer hardware, the quantum approximate optimization algorithm (QAOA) is a digital, gate-based alternative designed to run on universal, circuit-model quantum computers — the same class of hardware being built for quantum simulation and, eventually, fault-tolerant algorithms like Shor's. Farhi, Goldstone, and Gutmann's original QAOA paper defines a parameterized quantum circuit built from alternating layers of a "cost" unitary, derived from the same Ising/QUBO problem Hamiltonian used in quantum annealing, and a "mixer" unitary that drives transitions between computational basis states, with the circuit's variational parameters tuned by a classical optimizer running in a loop with the quantum device to maximize the expected quality of the resulting measured solution [3]. As the number of alternating layers, denoted p, increases, QAOA's output distribution provably approaches the adiabatic-limit behavior of quantum annealing, but at p=1 or a small handful of layers — the regime that near-term, noisy hardware can realistically support — QAOA is best understood as a heuristic variational algorithm whose performance depends heavily on problem structure, on how good the classical parameter optimization turns out to be, and on how much hardware noise degrades the circuit.

Zhou, Wang, Choi, Pichler, and Lukin's detailed study of QAOA's performance and mechanism is one of the most careful treatments of exactly this question, analyzing how solution quality scales with circuit depth p on the MaxCut problem and related combinatorial objectives, and providing both theoretical and numerically simulated evidence for how QAOA's approximation ratio — the ratio of the algorithm's typical output quality to the true optimum — improves as p grows, alongside a discussion of the classical parameter-optimization landscape that makes training deeper circuits progressively harder in practice [22]. Their analysis is a useful corrective to the sometimes-breathless framing of QAOA as "the" near-term quantum optimization algorithm: it is a real and interesting variational heuristic with reasonably well-understood scaling behavior at low depth, not a guaranteed route to solution quality that improves smoothly and cheaply with more qubits.

Applying QAOA specifically to scheduling has, so far, meant applying it to small, carefully chosen instances rather than anything resembling an industrial-scale problem, and the two scheduling-specific studies reviewed here are candid about this. Kurowski and colleagues formulated small job-shop scheduling instances as QUBO problems and evaluated QAOA's solution quality against classical baselines, reporting that at the small instance sizes their simulations and

limited hardware experiments could reach, QAOA solutions were broadly comparable to classical heuristic solutions, while noting that both circuit depth and the number of qubits required for realistic job-shop instances scale in ways that make near-term hardware a binding constraint well before industrially interesting problem sizes are reached [11]. Fitzek and colleagues applied QAOA to a heterogeneous vehicle-routing problem — a close scheduling relative, since assigning and sequencing deliveries to a fleet of vehicles is structurally similar to assigning and sequencing operations to a fleet of machines — and reported, concretely, that the number of qubits required scales quadratically with the number of customers, that their largest simulated instance involved only three customers and two trucks (21 qubits), and that the probability of the circuit sampling a valid (constraint-satisfying) solution reached up to 25% at p=5 circuit depth for an 11-qubit instance, falling to roughly 10% for their larger instances [12]. These are honest, modest numbers, and we report them as such rather than rounding them up: they describe a proof-of-concept demonstration on toy-sized problems, not evidence of practical advantage, and both papers' authors present them that way.

4. Quantum-Inspired Classical Solvers

The most operationally compelling results in this entire literature, in our reading, come not from quantum hardware at all but from classical solvers that borrow the Ising/QUBO problem structure and, in some cases, physical intuitions from quantum and optical systems, while running entirely on conventional or purpose-built classical computing substrates. This "quantum-inspired" category spans a genuinely diverse set of physical and algorithmic approaches, and it is worth treating them as a distinct research program rather than a footnote to quantum computing proper, because in several cases they have outperformed both classical heuristics and contemporaneous quantum hardware on the same benchmark problems.

Coherent Ising machines are perhaps the most physically striking example: Inagaki and colleagues built a system of degenerate optical parametric oscillator pulses, coupled to one another through a measurement-and-feedback loop, whose collective steady state is engineered to correspond to the low-energy configuration of a programmed Ising problem, and demonstrated the approach on MAX-CUT instances involving graphs with up to 2000 nodes, reporting that the machine found good solutions to these dense, densely-connected problem instances faster than reference implementations of established classical heuristics on the same hardware-comparable problems [4]. Because the coupling in a coherent Ising machine is realized optically rather than through fixed on-chip wiring between adjacent qubits, it does not suffer from the sparse-connectivity embedding overhead that constrains superconducting quantum annealers, which is part of why it has scaled to node counts well beyond what contemporaneous D-Wave hardware could embed without chain-based overhead.

Table 1: Representative quantum, quantum-inspired, and classical-baseline results reported across the scheduling and scheduling-adjacent literature synthesized in this article. [4,5,20,10,11,12].

Approach	Source	Problem / Scale Tested	Reported Result
Coherent Ising machine (optical, quantum-inspired)	Inagaki et al. [4]	MAX-CUT on graphs up to 2000 nodes	Found good solutions faster than reference classical heuristic codes on the same instances
Digital Annealer (physics-inspired, classical ASIC/GPU)	Aramon et al. [5]	QUBO benchmark problem sets	Matched or exceeded simulated-annealing solution quality in substantially shorter runtime
Superconducting quantum annealer	Johnson et al. [20]	108-qubit programmable annealing processor	Demonstrated programmable, hardware-realized quantum annealing dynamics on Ising problems
Quantum annealer on job-shop scheduling	Carugno et al. [10]	Job-shop scheduling QUBO instances, D-Wave hardware	Embedding overhead and chain breaks left solution quality behind classical solvers at tested sizes
QAOA on job-shop scheduling	Kurowski et al. [11]	Small job-shop instances, simulated and hardware QAOA	Comparable to classical heuristics only at very small instance sizes
QAOA on heterogeneous vehicle routing	Fitzek et al. [12]	Up to 21 qubits (3 customers, 2 trucks)	Valid-solution sampling probability up to 25% (11 qubits, p=5), falling to ~10% at larger sizes

Fujitsu's Digital Annealer, described by Aramon and colleagues, takes a different but related approach: a physics-inspired algorithm, closely related to simulated annealing but incorporating parallel-trial and adaptive-offset mechanisms borrowed from the same statistical-mechanics intuitions that motivate quantum

annealing, implemented on purpose-built digital hardware (originally ASIC-based, later reimplemented on GPUs) rather than on any device exhibiting genuine quantum effects [5]. Their reported benchmarks on QUBO problem instances show the Digital Annealer matching or exceeding the solution quality of standard simulated annealing implementations within substantially shorter wall-clock runtimes, which the authors attribute to the hardware's ability to evaluate large numbers of candidate spin flips in parallel rather than to any quantum mechanism whatsoever — a useful reminder that "quantum-inspired" in this literature is honestly a statement about mathematical lineage and problem encoding, not about the physics actually doing the computing.

Mohseni, McMahon, and Byrnes's review of Ising machines as hardware solvers surveys this entire landscape — superconducting quantum annealers, coherent Ising machines, digital annealers, and several other specialized architectures including memristor-based and CMOS-based implementations — as a unified category defined by the shared strategy of building or emulating physical dynamics that naturally settle toward low-energy Ising configurations, and it is candid that, across this category, the empirical evidence for a decisive advantage of any one physical substrate over well-tuned classical heuristics running on conventional hardware remains genuinely contested and highly dependent on problem structure, instance size, and how fairly the comparison is run [6]. Their review is, in our view, the single most useful entry point for a reader trying to understand this landscape without being pulled toward either uncritical enthusiasm or blanket dismissal, precisely because it treats "does this hardware category help" as an open empirical question to be settled instance class by instance class rather than a marketing claim to be accepted or rejected wholesale.

For scheduling specifically, the practical upshot of the quantum-inspired literature is that these classical, physically-motivated solvers are, today, considerably closer to being deployable at scheduling-relevant problem sizes than either quantum annealing or QAOA are, precisely because they are not bottlenecked by qubit count, coherence time, or embedding overhead in the same way. Table 1 places several of the resource and performance figures reported across this section side by side.

5. Applications to Job-Shop and Flow-Shop Scheduling

Job-shop scheduling — assigning a set of jobs, each consisting of an ordered sequence of operations, to a set of machines, subject to the constraint that each machine processes one operation at a time and each job's operations proceed in their specified order, while minimizing some objective such as total completion time — is the canonical scheduling problem the quantum-optimization literature has targeted, and Venturelli, Marchand, and Rojo's early study is the field's reference point. They formulated job-shop scheduling instances as QUBO problems suitable for D-Wave's quantum annealer, working through the nontrivial embedding and penalty-weight engineering required to represent precedence and machine-capacity constraints on the annealer's fixed hardware graph, and reported results on modest-sized benchmark instances that demonstrated the formulation was workable in principle while being explicit that the annealer's limited qubit connectivity, and the chain-based logical-to-physical qubit mapping this connectivity necessitates, constrained the instance sizes that could be embedded and solved competitively relative to classical methods [9].

Carugno, Ferrari Dacrema, and Cremonesi's more recent and considerably more systematic evaluation revisits exactly this question with a newer generation of D-Wave hardware and a more rigorous experimental pipeline, comparing solution quality against multiple classical solvers across a range of job-shop instances [10]. Their contribution is less a single headline number than a careful accounting of where, concretely, the difficulties in the pipeline arise: the computational cost of translating the scheduling problem into the QUBO formulation the annealer requires, the number of physical qubits consumed by chains representing single logical variables (which grows with problem density and can dominate the available qubit budget well before the "interesting" part of the problem is reached), the frequency and impact of chain breaks (cases where a chain of physical qubits meant to represent one logical variable fails to agree on a single value after annealing, requiring a repair heuristic that degrades solution quality), and the partial mitigation these issues receive from advanced techniques such as reverse annealing. Read plainly, their results indicate that, for job-shop scheduling specifically, the current generation of

D-Wave hardware faces several compounding overheads that leave it behind well-tuned classical solvers on the instance sizes tested, and the paper is explicit that this is a genuinely useful negative result rather than a discouraging footnote — it tells the field precisely which engineering problems (embedding efficiency, chain-break mitigation, qubit-budget accounting) need to improve before quantum annealing becomes competitive on this problem class.

Kurowski and colleagues' QAOA-based treatment of job-shop scheduling, already introduced above in the gate-based context, reaches a structurally similar conclusion from the digital-quantum side: their reported comparability between QAOA and classical heuristics held only at very small instance sizes, and the qubit and circuit-depth requirements for realistically-sized job-shop instances scale in ways that place industrially relevant problems well beyond current or near-term hardware [11]. Schworm and colleagues extended this evaluation to the flexible job-shop scheduling problem — a generalization in which each operation may be processed on any of several eligible machines rather than a single fixed machine, which is both more representative of real manufacturing settings and combinatorially harder — using quantum-annealing-based algorithms, and their evaluation similarly reports the need to weigh solution quality against the substantial formulation and embedding overhead specific to the flexible variant's additional machine-assignment decisions, echoing the same qubit-budget and embedding themes raised by Carugno et al. for the simpler fixed-machine case [23].

Taken together, the job-shop and flow-shop literature supports a consistent, if unglamorous, conclusion: the QUBO formulation machinery for this problem class is mature and well-documented [7,8], multiple independent groups have implemented and benchmarked it on both quantum annealing and gate-based hardware [9,10,11,23], and the consistent finding across these studies is that current-generation quantum hardware is not yet competitive with well-tuned classical solvers on this problem class at any but the smallest instance sizes, for reasons (embedding overhead, chain breaks, limited qubit connectivity, shallow achievable circuit depth) that are specific, identified, and in some cases actively being engineered around rather than being either mysterious or fundamental limits on what quantum approaches could eventually achieve.

6. Applications to Vehicle Routing and Broader Combinatorial Scheduling

Vehicle routing and its many variants — capacitated vehicle routing, routing with time windows, heterogeneous-fleet routing — sit adjacent to job-shop scheduling in the broader combinatorial-optimization taxonomy: both problems involve assigning discrete tasks (deliveries, operations) to discrete resources (vehicles, machines) subject to sequencing and capacity constraints, and both admit natural QUBO/Ising formulations of essentially the same flavor described by Lucas [7]. The quantum and quantum-inspired literature on vehicle routing is, if anything, slightly more mature in terms of engaging with real-world constraint richness than the job-shop literature, in part because vehicle-routing applications in logistics come with immediately obvious commercial stakes that have attracted applied research attention.

Fitzek and colleagues' QAOA treatment of the heterogeneous vehicle-routing problem, already introduced above, is notable for engaging directly with fleet heterogeneity — trucks of different capacities and characteristics rather than an idealized uniform fleet — and for reporting their results with the specific honesty this article has tried to model throughout: a quadratic qubit-count scaling with customer count, a largest tested instance of three customers and two trucks, and valid-solution sampling probabilities in the 10-25% range depending on instance size and circuit depth [12]. These numbers describe a genuine proof of concept and nothing more; extrapolating them to fleets of hundreds of vehicles and thousands of customers, the scale at which real logistics networks operate, would require qubit counts and circuit fidelities well beyond anything gate-based hardware offers today.

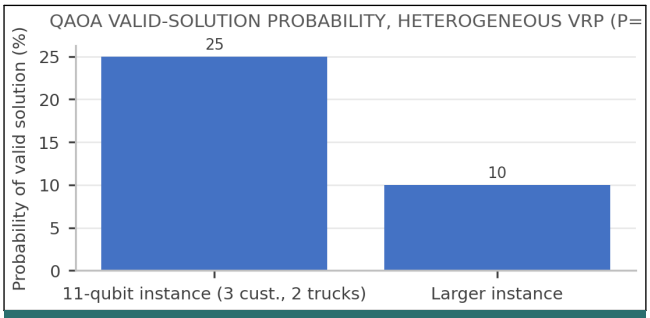


Figure 1: Probability that QAOA samples a valid (constraint-satisfying) solution for the heterogeneous vehicle-routing problem, at circuit depth p=5, on the largest instances Fitzek et al. simulated -- an 11-qubit instance versus their larger, higher-qubit instances -- illustrating how sharply solution-sampling quality falls off as problem size grows on current gate-based hardware [12].

Osaba, Villar-Rodriguez, and Asla's work on a real-world package-delivery routing problem takes a more application-driven and, notably, hybrid quantum-classical approach: rather than relying purely on a quantum or quantum-inspired solver end to end, their Q4RPD system explicitly decomposes the problem and blends quantum annealing with classical routines to handle a fuller set of real-world constraints — a heterogeneous vehicle fleet, delivery priorities, and per-vehicle capacity limits expressed along two independent dimensions (weight and physical volume) rather than the single-capacity abstraction most textbook formulations use [13]. Evaluated across six delivery instances, their reported results show solutions deviating by less than roughly 6% from classical benchmark solutions when priority constraints are strictly enforced, and reaching solutions matching classical benchmarks in scenarios without priority constraints. This hybrid decomposition strategy — using a quantum or quantum-inspired solver for the combinatorially hardest sub-component of a larger, constraint-rich real-world problem, while handling the remaining structure classically — recurs as a pattern across several of the more application-oriented papers surveyed in this article, and strikes us as a considerably more realistic near-term deployment shape than an expectation that a quantum device alone will solve an entire, richly-constrained scheduling problem end to end.

Beyond these two directly-scheduling-relevant studies, the broader combinatorial-optimization survey literature situates vehicle routing and job-shop scheduling within a wider family of problems — including graph coloring, set partitioning, and various assignment problems — that share the same underlying QUBO/Ising structural DNA, and Werner and Ullinger's recent survey of quantum algorithms specifically for scheduling problems is, to our knowledge, the most comprehensive single map of this sub-literature, cataloguing both the annealing-based and gate-based approaches applied across job-shop, flow-shop, and related scheduling variants and making the same core observation this article has arrived at independently: reported results remain concentrated on small benchmark instances, and closing the gap to industrially relevant scale is the dominant open engineering and algorithmic challenge across the entire sub-field [17].

7. Benchmarking, Hardware Constraints, and Open Challenges

Several recurring themes emerge once the studies above are read side by side rather than in isolation, and naming them explicitly seems more useful than leaving them implicit. First, benchmarking honesty and apples-to-apples comparison remain a genuine methodological challenge across this literature. Yarkoni and colleagues' review of quantum annealing for industry applications is particularly direct on this point, surveying reported use cases across logistics, finance, and manufacturing scheduling and cautioning that many early industry-facing quantum-annealing results compared against weak or under-tuned classical baselines, inflating the apparent advantage of the quantum approach; their review argues, and we agree, that any claim of quantum or quantum-inspired advantage on a scheduling problem should be read skeptically unless the classical baseline it is compared against was given comparable engineering effort [14]. McGeoch and Wang's early, methodologically careful experimental evaluation of an adiabatic quantum annealer on combinatorial optimization problems set something of a standard for this kind of rigor well before quantum annealing hardware had scaled to its current size, explicitly comparing wall-clock performance against multiple classical solvers on the same problem instances rather than reporting quantum results in isolation, and much of the more recent scheduling-specific literature (Carugno et al. and Kurowski et al. in particular) follows a similar comparative discipline [24,10,11].

Second, the embedding and connectivity overhead that recurs throughout the job-shop and vehicle-routing literature is not, in our reading, a peripheral engineering

detail but close to the central practical bottleneck standing between current quantum annealing hardware and competitiveness on scheduling problems at real-world scale. Hauke and colleagues' broad review of quantum annealing methods and implementations situates this overhead within the wider physics of the problem, discussing how hardware connectivity graphs, control precision, and thermal noise jointly determine what problem sizes and structures a given generation of annealing hardware can realistically tackle, and their review is candid that the mismatch between the dense, highly-connected problems many real scheduling instances produce and the comparatively sparse connectivity of current annealing hardware remains a first-order limitation rather than a solved problem [15]. This is precisely the mismatch that Venturelli et al. and Carugno et al. document empirically for job-shop scheduling specifically [9,10].

Third, the relationship between quantum and quantum-inspired approaches deserves to be stated plainly rather than left as an implicit competition. Mohseni et al.'s review treats them as a shared category for good reason: the mathematical problem (finding low-energy Ising configurations) is identical, and progress on formulation techniques, penalty-weight tuning, and problem decomposition transfers freely between the quantum and classical branches of this research program [6]. Given that quantum-inspired classical solvers currently reach substantially larger problem sizes than quantum hardware does — thousands of variables for coherent Ising machines and digital annealers versus tens of qubits for competitive QAOA demonstrations — a defensible near-term recommendation for a practitioner actually trying to solve a large scheduling problem today is to use the QUBO formulation techniques this literature has matured, but to run them on quantum-inspired classical hardware rather than waiting on quantum hardware to catch up in scale.

Fourth, Abbas and colleagues' broad, multi-author review of challenges and opportunities in quantum optimization — spanning both annealing and gate-based approaches, and written by a large group of researchers spread across academia and industry quantum-computing efforts — offers what we consider the most balanced summary judgment currently available in the literature: that quantum optimization approaches have matured substantially as an engineering and algorithmic discipline, that formulation and benchmarking practices have improved markedly over the past several years, but that no rigorous, apples-to-apples demonstration of quantum advantage for a practically relevant combinatorial optimization problem, scheduling included, yet exists, and that closing this gap will require simultaneous progress on hardware scale, error mitigation, problem-specific algorithm design, and — the theme this section has returned to repeatedly — honest, well-controlled classical baselines [16]. We find this a fair and well-evidenced conclusion, and it is the one this article's own synthesis independently arrives at as well.

Fifth, and worth naming as a genuinely open research direction rather than a solved engineering problem: almost none of the studies surveyed here address dynamic or online scheduling, in which jobs, machines, or constraints change after the schedule has been computed and the solver must react rather than solve a single static instance once. Real scheduling environments — a factory floor with machine breakdowns, a delivery fleet with last-minute order changes — are overwhelmingly dynamic in this sense, and the entire quantum and quantum-inspired literature reviewed here, to our knowledge, has been benchmarked exclusively on static, one-shot problem instances. Whether the qubit-count and embedding overheads discussed throughout this article would be more or less favorable in an online, repeatedly-resolved setting is, as far as we can tell, an open question the field has not yet engaged with directly.

8. Conclusion

The literature synthesized in this article supports a specific and, we think, defensible set of claims about quantum-inspired optimization for large-scale combinatorial scheduling. The mathematical translation of scheduling problems into the Ising/QUBO form that both quantum and quantum-inspired solvers require is mature, well-documented, and no longer a research bottleneck in its own right [7,8]. Genuinely quantum approaches — annealing-based and gate-based alike — have been applied directly to job-shop scheduling and vehicle routing by multiple independent groups, with formulation and embedding techniques that are technically sound but that, on the evidence gathered here, do not yet outperform well-tuned classical solvers

at any but small instance sizes, for reasons (limited qubit connectivity, chain-break overhead, shallow achievable circuit depth) that are specific and actively being engineered around rather than fundamental or mysterious [1,2,9,10,11,12]. Quantum-*inspired* classical solvers — coherent Ising machines, digital annealers, and related physics-inspired classical hardware — currently offer the most operationally compelling results in this entire space precisely because they sidestep qubit-count and embedding constraints while retaining the same problem-formulation machinery, and they represent, in our assessment, the most realistic near-term path for organizations wanting to apply this line of research to actual large-scale scheduling problems today [4,5,6].

None of this amounts to a claim that quantum computing has "solved," or is close to solving, large-scale combinatorial scheduling — no credible paper reviewed here makes that claim either, and the field's own review literature is unusually candid, for an emerging technology area, about the gap between demonstrated results and deployment readiness [14,15,16,17]. What the literature does support is a coherent research program with a clear, shared mathematical foundation, an increasingly rigorous benchmarking culture that compares honestly against classical baselines rather than only against earlier, weaker quantum results, and a set of specific, identified engineering bottlenecks — hardware connectivity, qubit-budget overhead from problem embedding, achievable circuit depth, and the near-total absence of work on dynamic and online scheduling variants — whose resolution, rather than any fundamental theoretical breakthrough, is what stands between today's small-scale demonstrations and scheduling systems that could plausibly run a real factory floor or delivery fleet. We think that is a genuinely useful, if less dramatic, place for the field to be, and one that deserves continued, honestly-benchmarked empirical work rather than either premature deployment claims or premature dismissal.

9. References

1. Kadowaki, T., Nishimori, H. (1998) Quantum Annealing in the Transverse Ising Model. Physical Review E, 58(5): 5355-5363. DOI: 10.1103/PhysRevE.58.5355
2. Farhi, E., Goldstone, J., Gutmann, S., Sipser, M. (2000) Quantum Computation by Adiabatic Evolution. arXiv preprint. DOI: 10.48550/arXiv.quant-ph/0001106
3. Farhi, E., Goldstone, J., Gutmann, S. (2014) A Quantum Approximate Optimization Algorithm. arXiv preprint. DOI: 10.48550/arXiv.1411.4028
4. Inagaki, T., Haribara, Y., Igarashi, K., Sonobe, T., Tamate, S., Honjo, T., et al. (2016) A Coherent Ising Machine for 2000-Node Optimization Problems. Science, 354(6312): 603-606. DOI: 10.1126/science.aah4243
5. Aramon, M., Rosenberg, G., Valiante, E., Miyazawa, T., Tamura, H., Katzgraber, H.G. (2019) Physics-Inspired Optimization for Quadratic Unconstrained Problems Using a Digital Annealer. Frontiers in Physics, 7: 48. DOI: 10.3389/fphy.2019.00048
6. Mohseni, N., McMahon, P.L., Byrnes, T. (2022) Ising Machines as Hardware Solvers of Combinatorial Optimization Problems. Nature Reviews Physics, 4(6): 363-379. DOI: 10.1038/s42254-022-00440-8
7. Lucas, A. (2014) Ising Formulations of Many NP Problems. Frontiers in Physics, 2: 5. DOI: 10.3389/fphy.2014.00005
8. Glover, F., Kochenberger, G., Du, Y. (2019) Quantum Bridge Analytics I: A Tutorial on Formulating and Using QUBO Models. 4OR, 17(4): 335-371. DOI: 10.1007/s10288-019-00424-y
9. Venturelli, D., Marchand, D.J.J., Rojo, G. (2016) Quantum Annealing Implementation of Job-Shop Scheduling. arXiv preprint. DOI: 10.48550/arXiv.1506.08479
10. Carugno, C., Ferrari Dacrema, M., Cremonesi, P. (2022) Evaluating the Job Shop Scheduling Problem on a D-Wave Quantum Annealer. Scientific Reports, 12: 6539. DOI: 10.1038/s41598-022-10169-0
11. Kurowski, K., Pecyna, T., Slysz, M., Różycki, R., Waligóra, G., Węglarz, J. (2023) Application of Quantum Approximate Optimization Algorithm to Job Shop Scheduling Problem. European Journal of Operational Research, 310(2): 518-528. DOI: 10.1016/j.ejor.2023.03.013
12. Fitzek, D., Ghandriz, T., Laine, L., Granath, M., Frisk Kockum, A. (2024) Applying Quantum Approximate Optimization to the Heterogeneous Vehicle Routing Problem. Scientific Reports, 14: 25415. DOI: 10.1038/s41598-024-76967-w
13. Osaba, E., Villar-Rodriguez, E., Asla, A. (2024) Solving a Real-World Package Delivery Routing Problem Using Quantum Annealers. Scientific Reports, 14: 24791. DOI: 10.1038/s41598-024-75572-1
14. Yarkoni, S., Raponi, E., Bäck, T., Schmitt, S. (2022) Quantum Annealing for Industry Applications: Introduction and Review. Reports on Progress in Physics, 85(10): 104001. DOI: 10.1088/1361-6633/ac8c54

15. Hauke, P., Katzgraber, H.G., Lechner, W., Nishimori, H., Oliver, W.D. (2020) Perspectives of Quantum Annealing: Methods and Implementations. Reports on Progress in Physics, 83(5): 054401. DOI: 10.1088/1361-6633/ab85b8

16. Abbas, A., Ambainis, A., Augustino, B., et al. (2024) Challenges and Opportunities in Quantum Optimization. Nature Reviews Physics, 6(12): 718-735. DOI: 10.1038/s42254-024-00770-9

17. Werner, T., Ullinger, F. (2026) Quantum Algorithms for Scheduling Problems: A Survey. EPJ Quantum Technology, 13: 31. DOI: 10.1140/epjqt/s40507-026-00494-y

18. Preskill, J. (2018) Quantum Computing in the NISQ Era and Beyond. Quantum, 2: 79. DOI: 10.22331/q-2018-08-06-79

19. Kirkpatrick, S., Gelatt, C.D., Vecchi, M.P. (1983) Optimization by Simulated Annealing. Science, 220(4598): 671-680. DOI: 10.1126/science.220.4598.671

20. Johnson, M.W., Amin, M.H.S., Gildert, S., Lanting, T., Hamze, F., Dickson, N., et al. (2011) Quantum Annealing with Manufactured Spins. Nature, 473(7346): 194-198. DOI: 10.1038/nature10012

21. Boixo, S., Rønnow, T.F., Isakov, S.V., Wang, Z., Wecker, D., Lidar, D.A., Martinis, J.M., Troyer, M. (2014) Evidence for Quantum Annealing with More Than One Hundred Qubits. Nature Physics, 10(3): 218-224. DOI: 10.1038/nphys2900

22. Zhou, L., Wang, S.T., Choi, S., Pichler, H., Lukin, M.D. (2020) Quantum Approximate Optimization Algorithm: Performance, Mechanism, and Implementation on Near-Term Devices. Physical Review X, 10(2): 021067. DOI: 10.1103/PhysRevX.10.021067

23. Schworm, P., Wu, X., Klar, M., Aurich, J.C. (2026) Evaluation of Quantum Annealing-Based Algorithms for Flexible Job Shop Scheduling. Procedia CIRP, 138: 36-41. DOI: 10.1016/j.procir.2026.01.008

24. McGeoch, C.C., Wang, C. (2013) Experimental Evaluation of an Adiabatic Quantum System for Combinatorial Optimization. Proceedings of the ACM International Conference on Computing Frontiers (CF '13), Article 23. DOI: 10.1145/2482767.2482797

10. Declarations

Funding

This literature synthesis received no specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflict of Interest

The authors declare that they have no competing interests relevant to this work.

Data Availability Statement

This article is a literature synthesis and did not generate new experimental data; all quantitative figures discussed are drawn from, and attributed to, the cited primary sources, which should be consulted for underlying data and code availability.

Ethical Approval

Not applicable. This work is a literature review that did not involve human participants, animal subjects, or the collection of primary data.