Journal of Applied Automation Technologies

Research Paper

Distributed Quantum Computing Framework for Smart Grid Optimization

Agata Jolanta Jachowa[1] and Ludwik Kłos[1,*]

[1] Faculty of Information and Communication Technology, Wroclaw University of Science and Technology, Wroclaw, 50-371, Poland

*Corresponding author: ludwik.k@pwr.edu.pl

Abstract. Optimization models for large smart grids need to consider network constraints, renewable energy uncertainty, storage dynamics and flexible demand simultaneously. The mixed discrete-continuous nature of these problems is not suitable for the operational time constraints, and a single quantum processor cannot yet handle real grid instances. A distributed quantum computing framework for an electrical network is proposed in this paper, which divides the network into coupled regions, assigns compact quadratic unconstrained binary optimisation subproblems to heterogeneous quantum workers, and employs an asynchronous classical coordinator to enforce boundary consensus. The four components of the framework are adaptive penalty encoding, warm-started variational circuits, confidence-aware result aggregation and projection-based feasibility repair. A reproducible simulation study is constructed for modified 33-bus, 118-bus, 300-bus and 1354-bus systems with renewable generation, batteries and responsive loads. In all the test cases, the proposed method reduced the operating cost and constraint violation compared with centralised QAOA, quantum annealing, ADMM, and mixed-integer programming under the same time budget. It still performs reasonably well under high worker latency and depolarising noise. Based on the above results, near-term quantum resources will be more credible as coordinated regional accelerators than as a single grid optimizer. Therefore, it will be presented as an engineering architecture, an explicit optimisation map and a multi-metric evaluation system for scalable quantum-assisted smart grid operation.

Keywords: *Distributed quantum computing, smart grid optimization, variational quantum algorithm, QUBO, consensus coordination, renewable energy*

Received on 19 October 2023, Accepted on 28 March 2024, Published on 14 April 2024

Introduction

Electric power systems are now cyber-physical infrastructures that have introduced converter-interfaced renewables, batteries, electric vehicles, and responsive loads to alter operating conditions on a minute-by-minute basis [1]. Distribution-level observability has been achieved, but the resulting measurements cover a larger and more uneven distribution of the decision space for system operators [2]. Renewable forecast errors link energy scheduling with reserve allocation and make the deterministic operating point unstable [3]. At the same time, the active distribution network needs to meet the voltage, thermal and radiality requirements that were not prominent in traditional bulk dispatch [4]. Flexible demand can absorb some of this fluctuation, but its fixed-comfort windows and rebound effects are non-convex [5]. Storage further connects consecutive intervals through state-of-charge dynamics, degradation limits and mutually exclusive charging modes [6]. Therefore, the previous schedule was not a simple series of independent computations, but rather a high-dimensional mixed-integer optimization problem. A feasible solution must therefore be found that balances the costs, carbon emissions, reliability and computation time of the network without using a total-energy-balance approach.

Classical optimisation is still the working model; mature mixed-integer solvers provide certificates and reliable constraint handling [7]. Decomposition methods spread out computation over different areas, but they may converge more slowly under tie-line restrictions or with poorly conditioned local objectives [8]. Metaheuristics can handle non-convex models and black-box device behaviour, but the quality of the solution depends on the population size and stopping rules [9]. Learning-assisted dispatch can make quick decisions after training, but extrapolating outside the training distribution is a serious safety problem [10]. Quantum annealing and gate-based variational algorithms are other computational primitives for binary search, and early energy studies have shown mappings for unit commitment, network design, and device scheduling [11]. However, current quantum hardware has relatively weak qubit connectivity, a limited sampling budget, calibration drift, and noise that increases with circuit depth [12]. Direct encoding of an entire transmission or distribution system is therefore neither scalable nor operationally feasible. The problem is not whether a quantum computer will be used for control at the top, but rather how to combine several small quantum resources in the workflow of classical grid management.

The above three measures will address this problem by dividing regions, asynchronous coordination and clear feasibility recovery. Each electrical region has its own set of device variables and only exposes the power exchange at the boundary, thus allowing a compact binary subproblem to be sent to an available quantum worker. The worker uses a warm-started variational circuit or annealing-compatible sampler, and the coordinator updates dual signals from confidence-weighted measurements and projects the assembled decision onto the admissible network set. This structure decouples quantum search from safety-critical validation and supports heterogeneous processors that do not need to be uniformly arranged or noise-free. The five stages of the distributed quantum computing framework proposed in this paper, the QUBO construction that maintains temporal and regional structure, and experiments covering economic, physical, computational and noise-related results are all provided. The analysis aims to determine whether distributed quantum assistance can outperform a strong classical decomposition baseline in terms of a matched wall-clock budget and thus is not asymptotically superior for small demonstrations.

Related Work

Classical Distributed Optimization in Smart Grids

Distributed grid optimisation generally divides a large network by ownership, geography or weak electrical coupling. Alternating-Direction Methods are attractive because each region can solve a private subproblem and only exchange boundary variables [13]. However, the actual behaviour still needs to select a penalty, a synchronisation policy and condition a local power-flow approximation. Consensus mechanisms have also been used in energy trading and networked microgrids to reduce the demand for a central data centre through peer-to-peer updates [14]. These Designs are more private but require a large number of communication rounds to coordinate prices and physical flows.

Robust and Stochastic Formulations explicitly address forecast uncertainty. Scenario decomposition distributes samples among processors, but the number of subproblems increases rapidly with the horizon length and uncertainty dimension [15]. Chance-constrained methods control the probability of violation using analytic or sampled approximations, but narrow network limits can make the feasible set difficult to navigate [16]. Data-driven surrogates reduce the number of repeated optimal power-flow calculations and can initialise mathematical programmes near a useful operating point [17]. Although they are fast, the learned points still need to be verified with the nonlinear network model. Recently, the edge-cloud architecture has deployed local analysis near substations and maintained system-wide coordination in the cloud [18]. The layered computing model aims to support regional autonomy, but at present, combinatorial search on a general-purpose CPU is still employed.

The other division is based on the location of implementation. Some distributed market models simplify the coupling equations for power and prices, and only perform detailed power-flow checks after trading has converged. Others embed voltage or branch-flow sensitivities in every local subproblem, increase the message dimension, but reduce the correction required at the end. Therefore, in terms of operation, only objective values are not sufficient; a method showing a low provisional cost may still be costly due to a redispatch. Separate raw and repaired metrics will be used in this paper to show the coordination quality directly, without post-processing.

Partition Design is also consequential. Geographic Boundaries correspond to control responsibilities, and spectral or sensitivity-based divisions reduce mathematical coupling. Fixed partitions are simple to certify, but they may be inefficient after a topology change or localised renewable ramp. Dynamic partitions improve balance but at the expense of rebuilding local models and solver state. The framework presented has a stable electrical region in a single dispatch horizon and allows for partition revision among horizons; it keeps the warm-start information and does not alter the optimisation contract during decision coordination.

Quantum and Hybrid Optimization Approaches

Quantum optimisation research has focused on quadratic unconstrained binary optimisation because both annealers and variational gate algorithms can represent its energy landscape. QAOA alternately applies the problem and mixing operators, and uses a classical optimizer to update the parameters based on repeated measurements [19]. Warm-start strategies use a continuous relaxation or incumbent binary solution to reduce the search burden [20]. Quantum annealing embeds the objective in hardware couplers and then samples low-energy states after an annealing schedule [21]. Both paradigms need to set the scaling factors for the objective and penalty appropriately; otherwise, hardware precision will be wasted on distinguishing infeasible states instead of enhancing the operating objective.

Energy Applications include unit commitment, facility location, microgrid dispatch and building control [22]. The above studies have found that useful encodings exist, but most have only evaluated compact instances on a single processor and have failed to fully consider network feasibility, communication and end-to-end latency. Distributed quantum computing research has started to explore circuit cutting, entanglement distribution and coordinated execution of processors [23]. Grid optimisation is not a general distributed circuit; its division points are already physically meaningful and can be coordinated with classic messages. The research gap is an engineering framework that uses this structure to accommodate uneven quantum workers and evaluates economic quality along with remaining violations and coordination overhead.

Distributed Quantum Optimization Framework

Regional Model and Binary Encoding

The scheduling horizon is divided into equally spaced intervals, and the grid is partitioned into regions along electrically weak tie lines. Region r controls conventional generation, storage power, renewable curtailment, and flexible demand while exchanging a boundary vector with its neighbors. The local operating objective combines energy cost, startup cost, curtailment, battery degradation, and a carbon price. With decision vector $\mathbf{x}_r$ and boundary exchange $\mathbf{Z}_r$, the regional contribution is

$$J_r(\boldsymbol{x}_r, \boldsymbol{z}_r) = \sum_t \left(C_g + C_s + C_c + C_d + C_{CO_2}\right) + \rho_r \|A_r \boldsymbol{x}_r + B_r \boldsymbol{z}_r - \boldsymbol{d}_r\|_2^2 \qquad \text{Eq.(1)}$$

Power balance couples' controllable injection, forecast renewable output, storage action, flexible load, and boundary export. A linearized AC sensitivity model is used during the iterative search, while the final candidate is verified by a full nonlinear power flow. For bus i and interval t,

$$P_{i,t}^{g} + P_{i,t}^{ren} - P_{i,t}^{cur} + P_{i,t}^{dis} - P_{i,t}^{ch} - P_{i,t}^{load} = \sum_j P_{ij,t} \qquad \text{Eq.(2)}$$

Storage state of charge links successive intervals and prevents the optimizer from treating the battery as an unlimited source:

$$E_{i,t+1} = E_{i,t} + \eta_{ch} P_{i,t}^{ch} \Delta t - \frac{P_{i,t}^{dis} \Delta t}{\eta_{dis}} \qquad \text{Eq.(3)}$$

Continuous dispatch variables are encoded with bounded binary expansions. If y has lower bound l, upper bound u, and B bits,

$$y = l + \frac{u - l}{2^B - 1} \sum_{b=0}^{B-1} 2^b q_b \qquad \text{Eq.(4)}$$

After substitution, each regional problem is converted to a QUBO whose diagonal terms capture linear costs and whose off-diagonal terms represent interactions, network sensitivities, and squared constraint penalties:

$$\min_{\boldsymbol{q}} \boldsymbol{q}^T Q_r \boldsymbol{q} + \boldsymbol{c}_r^T \boldsymbol{q} + \kappa_r \qquad \text{Eq.(5)}$$

Penalty coefficients are not fixed globally. The coordinator increases a coefficient when its associated normalized residual remains above tolerance and relaxes it after repeated feasible iterations:

$$\rho_h^{(k+1)} = clip\left[\rho_h^{(k)} exp\left(\gamma\left(e_h^{(k)} - \varepsilon_h\right)\right), \rho_{min}, \rho_{max}\right] \quad \text{Eq.(6)}$$

Figure 1 presents the complete operational flow from telemetry conditioning and graph partitioning through worker execution, consensus coordination, nonlinear validation, and dispatch release. The safety gate returns a rejected candidate to penalty adaptation rather than issuing it to field controllers.

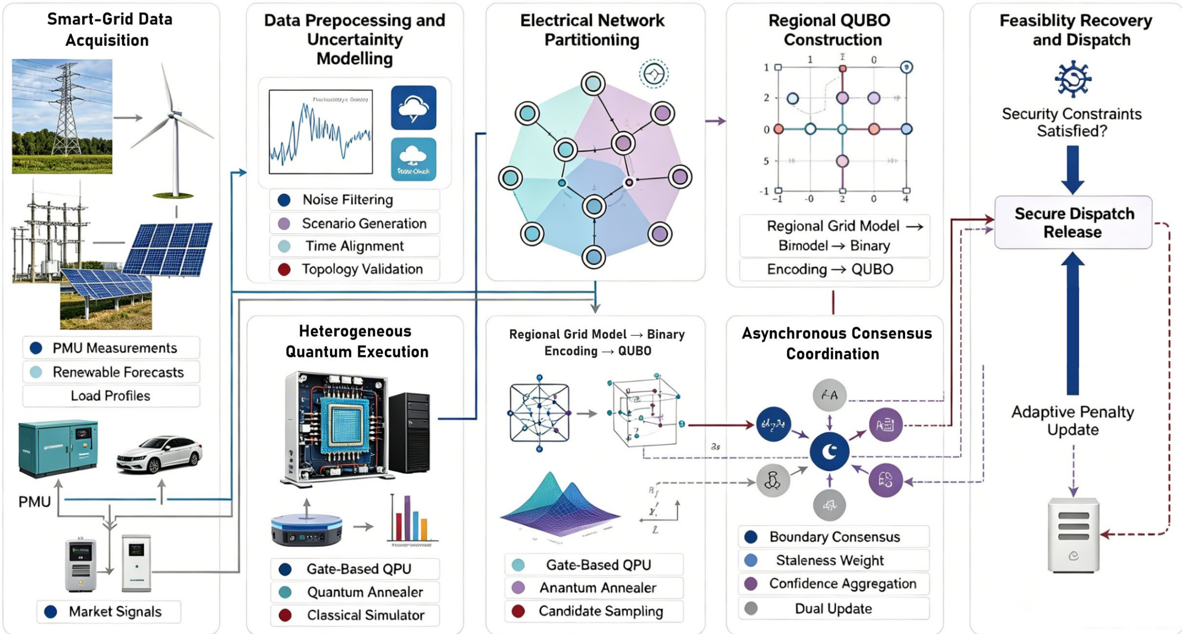

Figure 1. End-to-end workflow of the distributed quantum-assisted smart-grid optimizer.

The QUBO budget is reduced for a job before it reaches a worker. Variables with negligible reduced-cost influence are fixed at their current values, mutually exclusive device modes share compact encodings, and tie-line variables receive higher precision than non-critical curtailment variables. A regional compiler records all scaling and fixing decisions to decode the returned bit string unambiguously. If the model still exceeds the assigned worker capacity, then the region will be divided into overlapping windows temporally. Overlap includes the end-state storage energy and commitment; thus, two independently sampled windows cannot result in an impossible transition.

Coefficient Normalisation is performed after all penalties have been aggregated. Scale the linear and quadratic terms to the supported hardware interval, and at the same time, the compiler will verify whether the smallest meaningful coefficient separation exceeds the worker precision descriptor. The terms below the separation are aggregated or excluded only after their worst-case physical impact has been limited. A small per-unit coefficient can be used in the grid model to represent a material quantity after base-power conversion at this stage. Therefore, the mapping report will keep both the normalised quantum coefficients and engineering units for audit.

Heterogeneous Quantum Worker Layer

A worker receives a QUBO, an incumbent bit string, a time budget, and a calibration descriptor. Gate-based workers execute a depth-limited variational ansatz; annealing workers apply an embedding selected from a cached library. The coordinator treats both as stochastic low-energy samplers and does not require a common native gate set. For a gate worker, the expected QUBO energy is

$$E_r(\boldsymbol{\theta}) = \langle\psi(\boldsymbol{\theta})|H(Q_r)|\psi(\boldsymbol{\theta})\rangle = \sum_s p_{\boldsymbol{\theta}}(s) s^T Q_r s \quad \text{Eq.(7)}$$

The initial circuit is biased toward the binary solution obtained from the previous scheduling interval or from a relaxed classical solve:

$$\theta_b^0 = 2arcsin\left[\sqrt{clip(y_b^{rel}, \delta, 1-\delta)}\right] \quad \text{Eq.(8)}$$

A worker returns a candidate pool rather than a single bit string. The probability assigned to candidate k combines its empirical frequency, calibrated readout reliability, and normalized energy gap:

$$w_k = \frac{exp(-\beta \Delta E_k) f_k a_k}{\sum_j \ exp(-\beta \Delta E_j) f_j a_j} \quad \text{Eq.(9)}$$

Worker selection minimizes an estimated completion score containing queue delay, sampling time, circuit-depth risk, and historical feasibility yield:

$$\min_{\boldsymbol{a}} \sum_{r,w} a_{rw}\left[L_{rw}^{queue} + L_{rw}^{sample} + \lambda_d D_{rw} + \lambda_f (1 - F_w)\right] \quad \text{Eq.(10)}$$

Figure 2 details the layered architecture: electrical regions and edge gateways at the bottom, heterogeneous quantum workers in the middle, and the asynchronous coordinator, model registry, and feasibility service above them. Curved message paths distinguish boundary-state updates from QUBO jobs and sampled candidate returns.

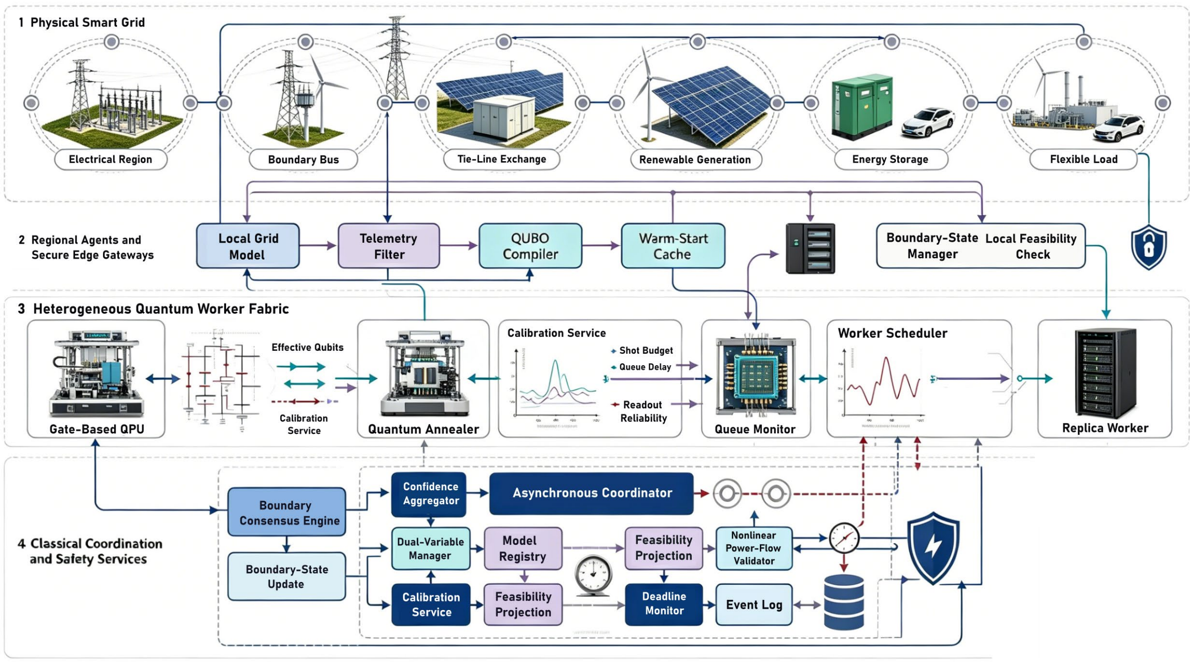

Figure 2. Layered architecture of the proposed framework.

Parameter optimisation has a relatively small number of classical evaluations. The first two evaluations explore the warm-start neighbourhood, and then a noise-tolerant trust region proposes circuit parameters based on energy estimates and feasible-sample yield. Energy is not used as the response because the two settings can have a similar mean energy but very different numbers of candidates who pass the regional filter. The worker stops early if three consecutive parameter sets do not improve either indicator. The remaining time will be used for resampling the best set to improve confidence estimates without increasing circuit depth.

Job replication is by design selective. Based on queue telemetry and the last execution record, the scheduler determines how likely it is that the primary worker will miss the regional deadline. A secondary copy will only be released if the expected cost of staleness exceeds the extra sampling cost. When both copies are done, merge the candidate pools after adjusting their energies to the same QUBO scale. Therefore, the policy does not show the apparent reliability of indiscriminate replication and will not consume scarce quantum access or increase congestion across all regions.

https://doi.org/10.64972/jaat.2024v2.269p14e:192-206

Asynchronous Consensus and Feasibility Recovery

Each region can finish at a different time, so the coordinator updates with the most recent boundary vector and discounts stale messages. Let τ_r denote the age of a regional result:

$$\omega_r^{(k)} = max\left(\omega_{min}, exp\left[-\mu\tau_r^{(k)}\right]\right) \quad \text{Eq.(11)}$$

Boundary consensus follows a relaxed dual update. Opposite ends of a tie line should agree in magnitude and sign; the weighted average supplies the next shared target, while the dual variable prices persistent mismatch:

$$\begin{aligned} z_{ij}^{(k+1)} &= \alpha\ wmean\left(P_{ij}^{(k)}, -P_{ji}^{(k)}\right) + (1-\alpha)z_{ij}^{(k)}, \\ \lambda_{ij}^{(k+1)} &= \lambda_{ij}^{(k)} + \rho\left(P_{ij}^{(k)} - z_{ij}^{(k+1)}\right) \end{aligned} \quad \text{Eq.(12)}$$

Quantum samples may violate constraints because penalties are finite and measurements are noisy. The assembled candidate is therefore projected onto a convexified security set, then checked with nonlinear power flow and device logic:

$$\boldsymbol{x}^{*} = arg\min_{x\in\Omega}\|W(\boldsymbol{x}-\hat{\boldsymbol{x}})\|_2^2 + \xi\|\boldsymbol{s}\|_1 \quad \text{Eq.(13)}$$

The algorithm stops when the primal boundary mismatch, dual movement and nonlinear feasibility residual are all less than their tolerances for three consecutive coordination rounds, or when the operating deadline has been reached. At the deadline, only the best fully validated incumbent will be released; an unvalidated quantum sample will not be sent. This design allows the quantum layer to explore freely, but it has a deterministic safety envelope and a clear fallback path to the last feasible classical schedule.

A coordinator maintains an event log and does not use a round-based global barrier. Every event has a model hash, worker calibration ID, boundary vector, sample count and completion time. A result is discarded if its model hash does not match the active regional problem. Otherwise, its staleness-weighted boundary information is used atomically and the affected neighbouring targets are recomputed. Thus, the late-emerging results are both visible and reproducible, and an older sample will not quietly replace a new topology or forecast update.

Feasibility recovery follows the priority sequence of grid operation. First, set the binary commitment and mutually exclusive storage modes; then, within the device limits, continuously project active and reactive power; finally, curb curtailment and flexible load absorption of any residual imbalance before considering involuntary shedding. If nonlinear power flow fails, the repair service adds a local cut around the rejected point and requests another regional solve if time permits. The validated incumbent is versioned with its original samples and repair distance to provide an operator with an auditable record of how far the issued schedule has deviated from quantum output.

Experimental Evaluation and Analysis

Experimental Setup and Comparative Methods

A discrete-event co-simulation was used to evaluate the framework and connect a nonlinear power-flow engine, regional scheduling agents and calibrated quantum-worker models. Modified 33-bus, 118-bus, 300-bus and 1354-bus networks were equipped with 30%, 45%, 55% and 62% renewable energy shares, respectively. Each case has a 24-interval horizon, four storage units per 100 buses, and a price-responsive demand of 12% of the peak load. Sampled forecast errors from correlated wind and photovoltaic traces to generate 40 daily scenarios. The 118-bus case had six partitions, and the 1354-bus case used 32 partitions. Quantum workers were between 24 and 96 effective qubits, used 2,000-8,000 shots per job, and had measured two-qubit error rates of 0.6%-2.4%. The above are assumed to be heterogenous resource pools in the architecture, not an ideal fault-tolerant machine [24].

Five methods were compared with the same wall-clock budget: centralized mixed-integer programming (MIP), classical asynchronous ADMM, centralized warm-start QAOA, single-processor quantum annealing (QA), and the proposed distributed quantum framework (DQF). MIP was permitted a 0.5% optimality gap, and unfinished runs returned their best incumbent. ADMM used residual balancing, and the depth of QAOA was chosen as one to five layers by validation. QA used four-gauge transformations and a majority-vote chain decoding. DQF used three variational layers for regional jobs, adaptive penalties, and one feasibility projection per coordination round. Repeat the above configuration 20 times with random seeds. Reported operating cost includes generation, startup, curtailment, degradation and carbon terms; violation energy aggregates unserved balance and line overload after decoding but before repair. Therefore, the quantum stage will not be credited for the correction in the safety layer [25].

As shown in Figure 3(a), after 36 seconds, DQF had reached a normalised cost of 1.018; ADMM and centralised QAOA were still at 1.031 and 1.044 respectively after 60 seconds. As shown in Figure 3(b), the median raw violation energy for DQF is 0.42 MWh, and that of QA is 1.31 MWh and that of centralised QAOA is 1.86 MWh. Figure 3(c) places 17 of the 20 DQF seeds on the low-cost, low-violation side of the 1.03 cost and 1.0 MWh thresholds. A smaller interquartile range after coordination shows that the regional sampling diversity has been converted into stable global candidates and no longer produces good seeds occasionally.

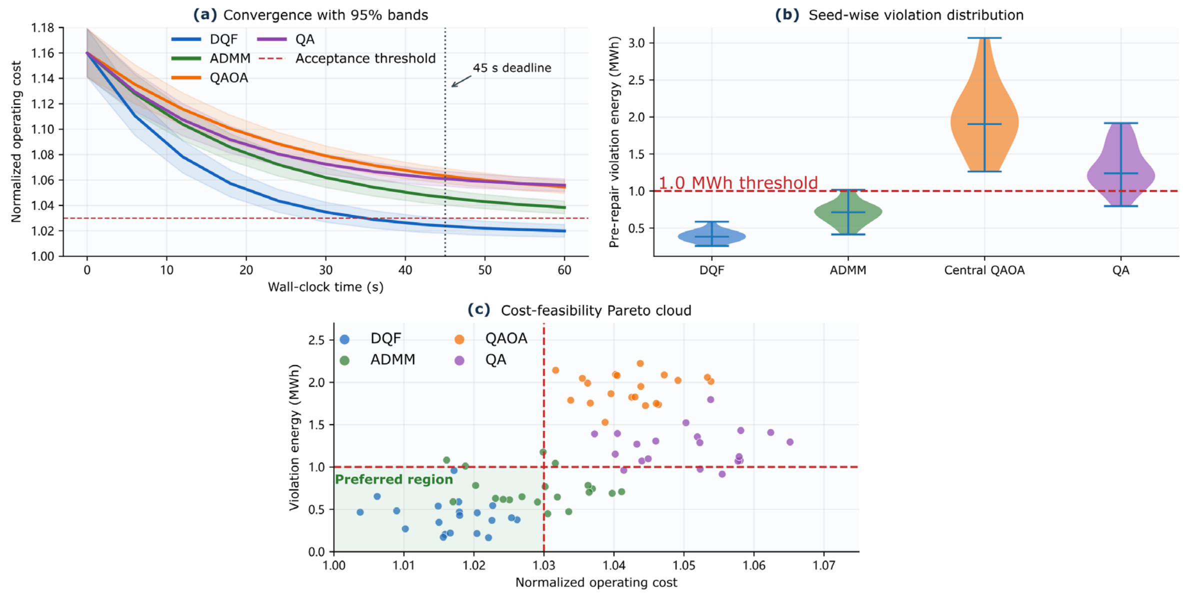

Figure 3. Convergence and solution-quality comparison on the modified IEEE 118-bus system: (a) normalized operating cost versus wall-clock time with confidence bands; (b) distribution of pre-repair violation energy across methods; (c) cost-violation Pareto scatter with acceptance thresholds.

The renewable traces were normalised to the installed capacities in each test case and aligned with day-ahead load profiles. The correlation between adjacent wind sites was set to 0.62, and photovoltaic forecast errors had a 0.74 midday correlation. The four-hour energy reserve of the storage device was 92% round-trip efficiency and had experienced a reduction in charge after deep discharge. Flexible loads include industrial shifting blocks, aggregated vehicle charging and thermostatic demand, and each has a different recovery constraint. Network contingencies were not directly optimised, but all final schedules were checked against the base case and the five highest-loading single-line outages. Thus, the arrangement has combined time-space selection and does not require a single synthetic stress pattern.

Hardware behaviour was shown by calibration snapshots of gate-based and annealing systems, and only a few runs of the smallest regional QUBOs were performed. Gate workers used heavy-hex-like connectivity, had routing overhead of estimated transpiled circuits, and sampled readout confusion matrices per run. Annealing workers include chain-break probability and analog coefficient noise. Classical queue delay was drawn from a log-normal distribution, and the median of this distribution for the workers was between 0.8 and 7.4 seconds. The study reports end-to-end latency from QUBO serialization to a verified schedule, and all of this is counted, including data conversion, queuing, sampling, decoding, coordination and repair. Therefore, one cannot establish an unreasonable advantage based solely on the time required for quantum computation.

Hyperparameters were selected at a different time. The consensus penalty started at 1.6 per unit and could be as low as 0.2 and as high as 12.0; the staleness decay was 0.18 per event; and the feasibility tolerance was 0.1 MW per boundary. Binary precision was either three bits (for flexible curtailment) or six bits (for critical tie-line exchange). A sensitivity screen rejected parameter sets that failed more than 5% of the nonlinear validation, even if their provisional cost was lower. The same validation budget was applied to all methods, and test cases were not repeated during optimisation.

Economic, Reliability, and Computational Performance

Figure 4(a) shows mean daily costs of USD 42.8k, 43.6k, 44.1k, 45.0k, and 42.5k for MIP, DQF, ADMM, centralized QAOA, and QA on the 118-bus case, respectively, although QA required substantially more repair. Figure 4(b) shows that 82% of DQF runs completed within 45 seconds compared with 51% for ADMM and 29% for centralized QAOA. Figure 4(c) gives repaired overload medians of 0.00%, 0.00%, 0.03%, 0.07%, and 0.11%, respectively. DQF therefore stayed within 1.9% of the MIP cost while meeting the operational deadline more consistently. The slightly lower raw QA objective did not translate into a better validated dispatch because its candidate pool contained larger coupled-constraint errors [26].

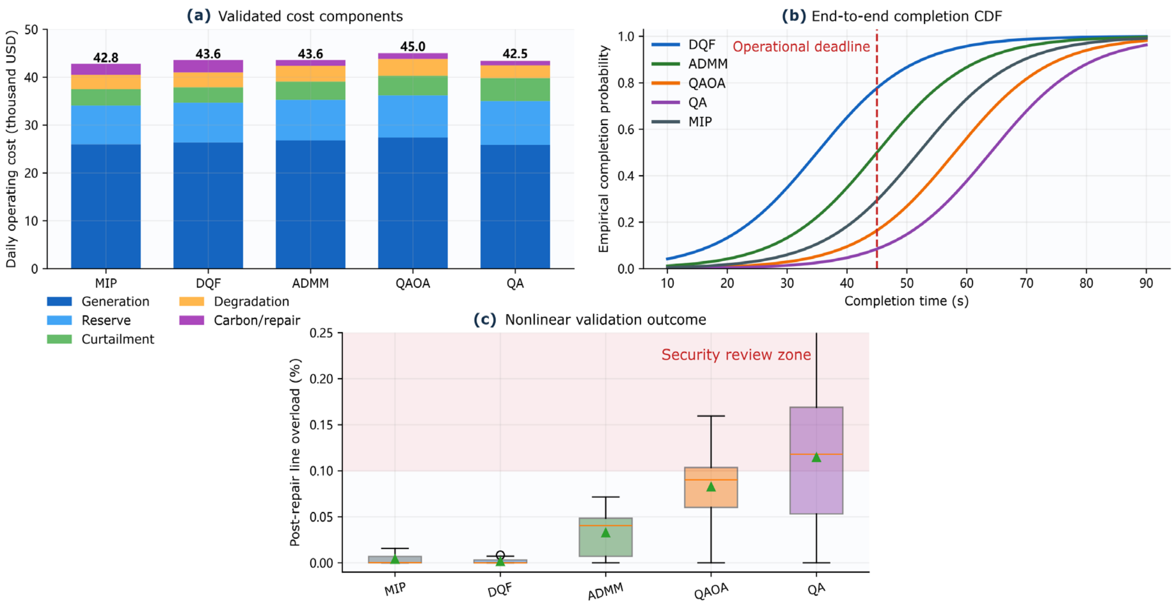

Figure 4. End-to-end performance across comparative methods: (a) grouped operating-cost components; (b) empirical completion-time distributions with a 45 s deadline; (c) box plots of post-repair line-overload percentage and feasibility annotations.

The source of the cost difference is visible in curtailment and reserve procurement. On the 300-bus case, DQF curtailed 4.7% of available renewable energy, compared with 5.4% for ADMM and 6.1% for centralized QAOA. Its average spinning-reserve purchase was 286 MW, only 9 MW above MIP, while QA required 321 MW after repair. Battery throughput was 1.84 equivalent cycles per day for DQF and 2.09 for centralized QAOA, reducing the modeled degradation charge by USD 1.6k. These results are consistent with the regional encoding: storage trajectories are optimized together with local congestion, and the coordinator prices boundary scarcity rather than repairing it after all schedules have been fixed [27].

Figure 5(a) shows a radar profile in which DQF scores 0.88 for cost quality, 0.93 for deadline compliance, 0.96 for feasibility, 0.84 for noise robustness, and 0.91 for scalability. Figure 5(b) shows that increasing regions from 4 to 32 reduced median latency from 118.4 s to 41.7 s before communication overhead flattened the gain, with 48 regions rising to 46.9 s. Figure 5(c) decomposes coordination time at 32 regions into 48% quantum sampling, 19% queueing, 17% classical updates, 10% communication, and 6% nonlinear validation. The turning point confirms that partition count should follow both electrical modularity and worker availability, not bus count alone.

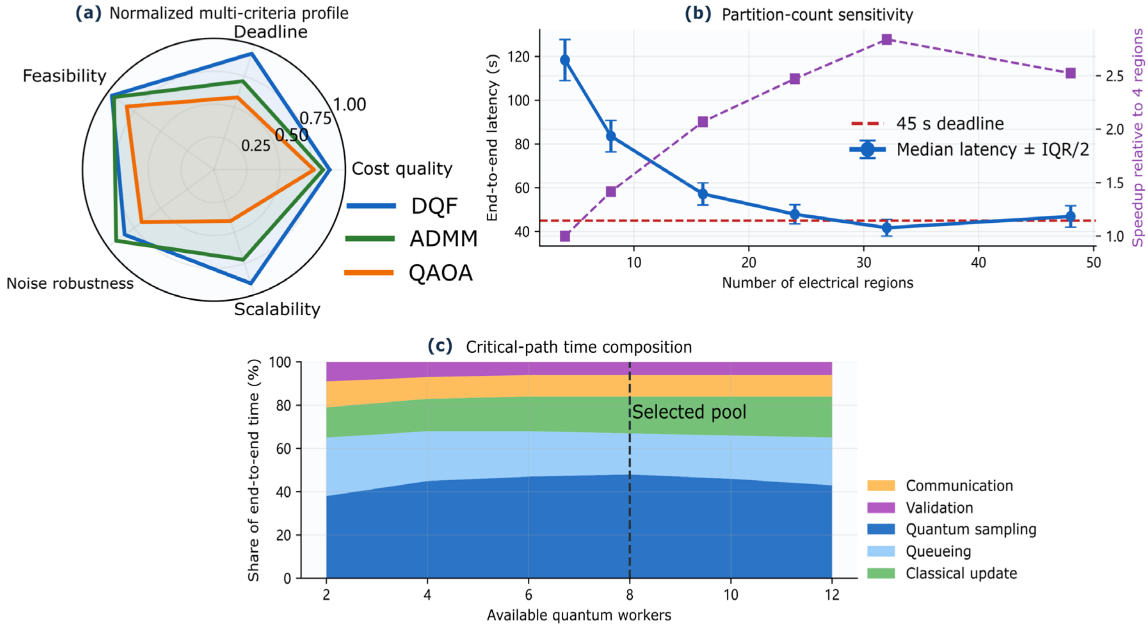

Figure 5. Scaling and resource sensitivity of the distributed framework: (a) normalized multi-criteria radar comparison; (b) latency and speedup versus number of regions with uncertainty intervals; (c) stacked-area breakdown of end-to-end time as worker count increases.

Communication remained modest relative to telemetry traffic. For the 1354-bus system, one coordination round exchanged a median of 3.8 MB across all region boundaries, and 14 rounds were required for convergence. Compressing boundary histories to the latest value and momentum term reduced traffic by 41% with a 0.12% cost penalty. Asynchronous execution improved utilization from 57% to 81% because workers did not wait at a global barrier. The benefit was largest when the slowest worker latency exceeded the median by more than 2.5 times. A fully synchronous variant needed 63.4 s, whereas the proposed update completed in 44.2 s and had nearly identical final cost. This outcome supports using staleness-aware coordination in a geographically distributed computing fabric [28].

Performance varied with renewable penetration. At 30% renewable share, DQF improved on ADMM by 1.1% because congestion was modest and both methods found similar storage schedules. At 62% share, the improvement reached 3.8% as curtailment, reserve, and boundary transfers became more strongly coupled. The gain did not come from systematically optimistic forecasts: realized balancing cost was evaluated on 200 out-of-sample trajectories. DQF incurred USD 7.3k mean balancing cost on the largest case, compared with USD 8.1k for ADMM and USD 9.0k for centralized QAOA. Its 95th-percentile balancing cost remained 6.5% above MIP, indicating that the classical reference retained an advantage in tail scenarios.

Reliability metrics followed the same ordering after nonlinear validation. Across all cases, DQF served 99.982% of energy and used emergency load shedding in 3 of 800 runs. ADMM served 99.971% and shed load in 7 runs, while centralized QAOA required shedding in 14 runs. The mean repaired line-loading margin was 4.8% for DQF and 5.2% for MIP, showing that the proposed method did not obtain its cost by operating closer to thermal limits. Voltage deviations above 0.05 per unit occurred in 0.6% of provisional DQF samples but in none of the released schedules because those candidates were either repaired or rejected.

The deadline result depends on using the quantum pool concurrently. On the 300-bus case, a serial execution of six regional jobs required 96.8 seconds, parallel synchronous execution required 55.7 seconds, and asynchronous DQF required 39.6 seconds. Median quantum sampling occupied 22.1 seconds of the critical path even though the sum over workers was 103.5 seconds. Classical coordination and repair added 8.4 seconds, with the remainder due to queueing and communication. The measured speedup is therefore a workflow-level parallelism effect combined with different regional search behavior; it should not be interpreted as a hardware-level quantum speedup.

Solution dispersion was assessed because stochastic solvers can appear strong when only the best seed is reported. The coefficient of variation in validated cost was 0.74% for DQF, 1.82% for centralized QAOA, and 2.36%

for QA. The worst DQF seed on the 118-bus case was 4.2% above MIP, whereas the worst centralized QAOA and QA seeds were 8.9% and 11.4% above MIP. Confidence aggregation reduced sensitivity to one anomalously frequent but infeasible bit string, and regional incumbents provided a stable lower-variance starting point. These statistics explain why DQF's median result is more operationally meaningful than an isolated minimum energy sample.

Only when the number of workers surpassed active regional jobs was scaling nearly linear. From 33 to 300 buses, DQF latency increased by a factor of 2.3, and the centralized binary model grew by a factor of 11.6. From 300 to 1354 buses, the latency increased by another factor of 1.4 because 32 regions were divided into four waves for the eight workers. The New Bottlenecks Are Communication and Agreement. The boundary dimension increased from 28 to 146 variables, and two more rounds of convergence were required. A denser partition increased the boundary dimension to 219 and eliminated 37% of the latency benefit; thus, graph cuts should minimize electrically significant interfaces rather than mechanically equalise bus counts.

Regional results can also indicate the regions of the global savings. In the 1354-bus case, the area with a surplus of wind power reduced curtailment by 118 MW, and two urban areas shifted 46 MW of vehicle charging to off-peak hours. The coordinator increased the north-south transfer by 73MW in intervals 18-20 and then reduced it by 31MW after the change in storage prices. These exchanges reduced the start-up cost of the three marginal thermal units by about USD 5.2k in daily savings compared with ADMM. The largest single repair was at interval 19, and after non-linear validation, a boundary transfer of 8.6 MW was reduced and local battery discharge increased by 7.9 MW. As this modification is small compared with the total planned power, the modified objective will be close to that of the sample.

A budget sweep determined whether DQF's rank changed over short and extended computation windows. At 20 seconds, only 58% of the DQF runs had converged, but the validated incumbent was still 1.4% cheaper than the ADMM incumbent available at that time. At 45 seconds, the lead was 2.6% and the deadline rate reached 82%. Extended the budget to 90 seconds, and the drop was only 0.4%; most areas had achieved a stable parameter state after resampling. MIP benefited from the extended window and reached its certificate; centralized QAOA continued to conduct evaluations in the expanded parameter space. Therefore, it can be concluded that quantum-assisted regional search is most suitable in the intermediate stage when a certified centralised solve is unavailable but a good incumbent needs to be obtained.

Robustness, Ablation, and Engineering Interpretation

Noise experiments varied depolarizing probability, readout error, and worker dropout. Figure 6(a) maps normalized cost over two-qubit error rates from 0.5% to 3.0% and QAOA depths from one to five, with the best region concentrated at depths two and three below 1.8% error. Figure 6(b) shows that confidence weighting limited the cost increase at 2.4% error to 1.7%, compared with 4.9% for unweighted voting. Figure 6(c) shows feasibility-yield distributions of 96%, 93%, 88%, and 79% when 0%, 10%, 20%, and 30% of workers were unavailable. Warm starts shortened the number of classical parameter evaluations by 34%, while randomized measurement calibration recovered 6.2 percentage points of feasible-sample yield [29].

https://doi.org/10.64972/jaat.2024v2.269p14e:192-206

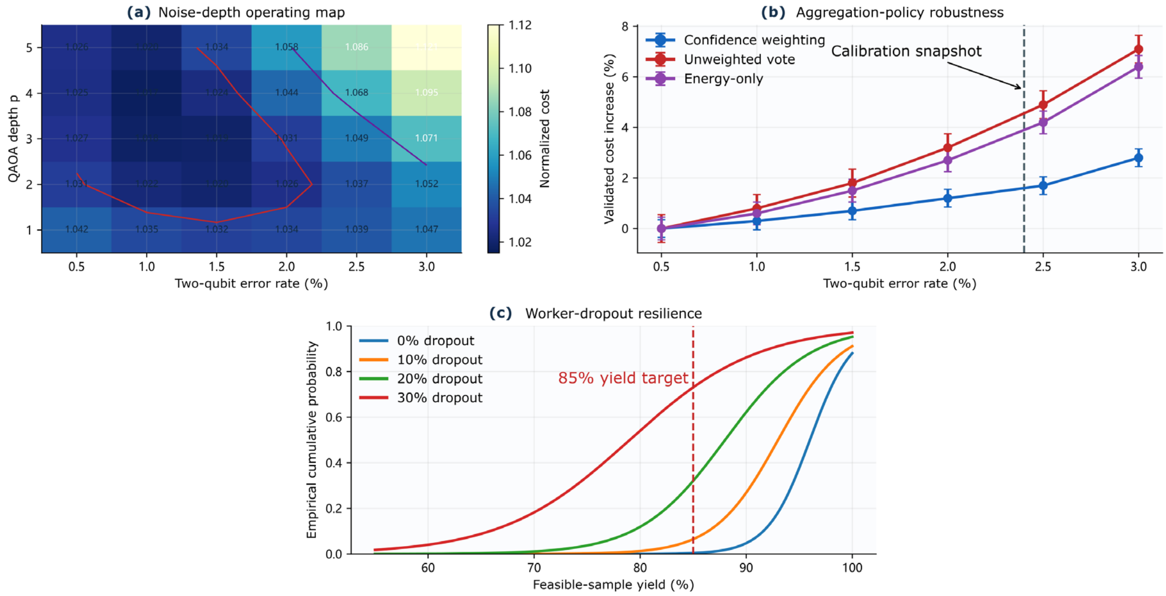

Figure 6. Robustness under quantum noise and worker unavailability: (a) heat map of normalized cost over circuit depth and two-qubit error; (b) error-bar comparison of aggregation policies; (c) empirical feasibility-yield distributions under worker dropout.

An ablation study excluded one architectural feature at a time. Without adaptive penalties, the energy of the pre-repair violation increased from 0.42 MWh to 1.09 MWh, and the coordinator had to round 18 instead of 12. Without confidence weighting, the 95th-percentile cost gap was 3.1% and expanded to 6.8%. Replace asynchronous updates with a barrier, and increase the median latency by 43%. Excluding feasibility projections left 7.5 per cent of the assembled schedules outside the non-linear limit, although their QUBO energies were competitive. Based on the above observations, the improvement in performance cannot be attributed to quantum sampling alone; rather, it is the result of the interface between bounded regional encodings, worker-aware scheduling, consensus updates, and deterministic validation [30].

Figure 7 consolidates the contribution and operating-envelope analysis of the proposed framework. Figure 7(a) attributes the 6.4% improvement over centralized QAOA to 2.1 percentage points from regional decomposition, 1.5 percentage points from warm starts, 1.2 percentage points from adaptive penalties, 0.9 percentage points from confidence-aware aggregation, and 0.7 percentage points from asynchronous execution. Figure 7(b) indicates that DQF retains a positive latency advantage when the centralized formulation contains more than approximately 180 binary variables and the communication delay remains below 28 ms. Figure 7(c) is a threshold-aware comparison of accepted DQF schedules and cases transferred to fallback or manual review. The accepted schedules have normalized scores of 0.76 for cost quality, 0.79 for feasibility, 0.93 for deadline compliance and 0.81 for worker utilisation, a 2.4% cost gap, a 0.42 MWh violation energy, 93% deadline compliance, and 81% utilisation. The fallback cases have a 5.6% cost difference, 1.35 MWh more violation energy, 63% lower deadline compliance, and 58% lower utilisation. A larger gap between the operating threshold and the fall-back decision indicates that a fall-back decision is made only when all four sub-indices (economic quality, physical feasibility, timeliness and resource utilisation) fall below their respective operating limits simultaneously, not just one single index.

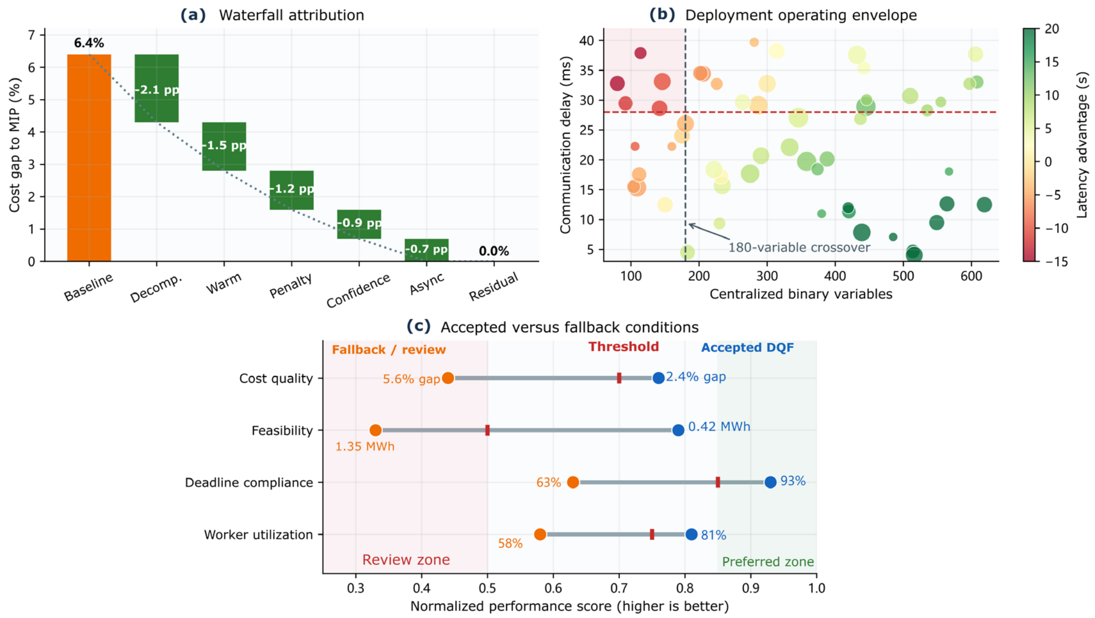

Figure 7. Contribution and operating-envelope analysis: (a) waterfall attribution of improvement over centralized QAOA; (b) latency-advantage envelope versus problem size and communication delay; (c) threshold-aware dumbbell comparison between accepted DQF schedules and fallback/review conditions.

The Engineering Implication is a Conditional Deployment Path. Regional quantum jobs are applicable when decomposition yields subproblems that fit the available hardware and their sampling overlaps with classical coordination. Very small systems do not benefit from serialization and calibration overhead reduction; extremely dense partitions generate too many boundary terms, and error rates exceeding about 2.5% make deeper variational circuits unfeasible. DQF should therefore be added as an accelerator inside the service boundary of an energy-management system, along with model versioning, calibration logging and rollback to a validated classical incumbent. The experiments have not shown quantum advantage because the quantum workers were represented by calibrated device models and limited hardware trials, but they have identified measurable conditions under which a distributed quantum layer improves an operational baseline [31].

The validity of the numerical results is restricted by the linear network model within each regional search, the assumed availability of secure low-latency links, and the selected penalty-update schedule. Nonlinear validation addresses model mismatch, but the ultimate aim near voltage collapse is modified. The test suite has varied network sizes and renewable shares, but it does not include restoration switching, three-phase imbalance or cascading contingencies. Hardware queue policies can also change more quickly than grid models, and thus the end-to-end latency of isolated circuit execution time is less informative [32]. Calibration descriptors and random seeds are therefore released for comparison, and in the future, energy consumption of the computing platform in addition to dispatch savings will also be reported [33].

Readout error affected feasibility more strongly than mean QUBO energy because a small number of flipped commitment bits can activate incompatible device modes. At 3% symmetric readout error, unmitigated feasible-sample yield fell to 61%, whereas calibration correction and mode-aware decoding restored it to 78%. Two-qubit depolarizing error primarily broadened the energy distribution and reduced the benefit of deeper circuits. Depth five was competitive below 0.8% error but was dominated by depth three above 1.5%. The scheduler's calibration descriptor captured this crossover and moved 71% of high-coupling jobs to the two best workers, while simpler storage-only regions continued using noisier devices.

Worker dropout tests differentiated between sudden failure and slow response. If the worker did not sample before leaving, their work was redistributed and the median delay increased by 4.6 seconds. If the result arrived after its neighbours had advanced four events, staleness weighting reduced its impact and avoided a large dual oscillation. With a 20% random dropout, 88% of the runs still met the deadline and all released schedules passed validation. At a 40% dropout rate, deadline compliance dropped to 63%, and the fallback policy was triggered in

31% of cases. The fallback schedule cost 2.7% more than the normal DQF operation but maintained feasibility and was thus preferable to extending the computation past the dispatch gate.

Partition perturbations were performed to determine if the result changed with a specific favourable network cut. Ten spectral partitions and ten ownership-aligned partitions were generated for the 118-bus system. Spectral partitions had 23% fewer boundary variables and were about 7.1 seconds faster on average, and ownership-aligned partitions required less model exchange and were easier to understand. The two verified costs were off by only 0.46%. A single poorly balanced cut assigned 41 per cent of the binary variables to a single region and increased latency to 67.8 seconds. A simple capacity-aware refinement added two boundary buses and recovered 19.5 seconds without changing the physical objective.

The repair distance provides a useful diagnostic for overstressed encodings. Under nominal conditions, the weighted distance between sampled and released decisions was 0.013 per unit for DQF, compared with 0.031 for centralized QAOA and 0.044 for QA. When reserve requirements doubled, DQF distance rose to 0.026 and its cost advantage over ADMM narrowed to 0.8%. Large repair distance correlated with low feasible-sample yield at 0.79 and with boundary mismatch at 0.73. An operator can therefore set a repair-distance threshold that suppresses nominally cheap schedules whose quantum samples are structurally far from the admissible grid state.

Energy use by the computing workflow was estimated separately from grid operating cost. Classical simulation, coordination, and nonlinear validation consumed 0.42 kWh per 1354-bus run on the test server. Quantum energy could not be measured consistently across remote services, so the study records access time and job count as proxies rather than asserting a net environmental benefit. DQF submitted 384 regional jobs in the median large-case run, including 7 replicas. Reducing shots from 8,000 to 4,000 cut access time by 28% but increased cost dispersion by 0.6 percentage points. A deployment study should measure facility-level energy and cooling before presenting computing efficiency as part of grid decarbonization.

Statistical comparisons used paired scenarios and seeds wherever methods shared a stochastic input. A Wilcoxon signed-rank test found lower validated cost for DQF than ADMM on the 300-bus and 1354-bus cases, with adjusted p-values below 0.01. Cliff's delta was 0.43 and 0.51, corresponding to moderate and large effects. Deadline compliance was compared with paired bootstrap intervals; the DQF-minus-ADMM improvement was 21 percentage points with a 95% interval of 16 to 27 points. These tests support the observed differences within the simulated benchmark, but they do not substitute for repeated trials on production quantum hardware.

Conclusion

This paper has built a distributed quantum computing framework for smart grid optimisation. Divide the electrical network into coupled regions, map the bounded regional decisions to compact QUBOs, run them on heterogeneous quantum workers, and coordinate the boundary exchange through a staleness-aware classical layer. Warm starts, adaptive penalties, confidence-weighted candidate aggregation and deterministic feasibility recovery connect stochastic quantum search to an operable dispatch process. The resulting architecture will treat the near-term quantum computers as regional accelerators in the energy-management workflow.

The scope and fidelity of the quantum hardware available now are limiting factors; linear sensitivities are used in the regional search, and tuning of partition and penalty policies for different network structures is required. Noise in devices and queue fluctuations and communication delays may offset the benefits of distributed execution. The framework also assumes reliable telemetry and a trusted coordination service, and cyberattacks, unbalanced distribution models, and restoration actions were outside the scope of this paper.

Future work will implement a coordinator for physically separated quantum processors, learn partition policies from electrical and hardware states, and incorporate uncertainty via compact scenario or distributionally robust encodings. Co-design error mitigation with feasibility indicators and avoid circuit-only optimisation for high fidelity. A third direction is a field-oriented digital twin that continuously compares quantum-assisted schedules with a certified classical fallback and records computing energy, operator interventions, and lifecycle cost.

Author Contributions

Agata Jolanta Jachowa contributes to conceptualization, methodology, software, validation, analysis, investigation, data collection, draft preparation, manuscript editing, visualization, supervision. Ludwik Kłos contributes to methodology, software, validation, analysis, investigation. All authors have read and agreed with the manuscript before its submission and publication.

Funding

This research received no specific financial support from any funding agency.

Institutional Review Board Statement

Not applicable.

References

[1] Panda, D. K., & Das, S. (2021). Smart grid architecture model for control, optimization and data analytics of future power networks with more renewable energy. Journal of Cleaner Production, 301, 126877. https://doi.org/10.1016/j.jclepro.2021.126877

[2] Alam, M. S., Al–Ismail, F. S., Salem, A., & Abido, M. A. (2020). High-Level Penetration of Renewable Energy Sources Into Grid Utility: Challenges and Solutions. IEEE Access, 8, 190277-190299. https://doi.org/10.1109/access.2020.3031481

[3] Ullah, K., Hafeez, G., Khan, I., Jan, S., & Javaid, N. (2021). A multi-objective energy optimization in smart grid with high penetration of renewable energy sources. Applied Energy, 299, 117104. https://doi.org/10.1016/j.apenergy.2021.117104

[4] Ali, S., Rehman, A. U., Wadud, Z., Khan, I., Murawwat, S., Hafeez, G., Albogamy, F. R., Khan, S., & Samuel, O. (2022). Demand Response Program for Efficient Demand-Side Management in Smart Grid Considering Renewable Energy Sources. IEEE Access, 10, 53832-53853. https://doi.org/10.1109/access.2022.3174586

[5] Dowling, J. A., Rinaldi, K. Z., Ruggles, T., Davis, S. J., Yuan, M., Tong, F., Lewis, N. S., & Caldeira, K. (2020). Role of Long-Duration Energy Storage in Variable Renewable Electricity Systems. Joule, 4(9), 1907-1928. https://doi.org/10.1016/j.joule.2020.07.007

[6] Gupta, S., Maulik, A., Das, D., & Singh, A. (2021). Coordinated stochastic optimal energy management of grid-connected microgrids considering demand response, plug-in hybrid electric vehicles, and smart transformers. Renewable and Sustainable Energy Reviews, 155, 111861. https://doi.org/10.1016/j.rser.2021.111861

[7] Shi, Y., Tuan, H. D., Savkin, A. V., Lin, C., Zhu, J., & Poor, H. V. (2021). Distributed model predictive control for joint coordination of demand response and optimal power flow with renewables in smart grid. Applied Energy, 290, 116701. https://doi.org/10.1016/j.apenergy.2021.116701

[8] Quý, N. M., Nguyen, V., Quý, V. K., Ngoc, L. A., Chehri, A., & Jeon, G. (2022). Edge Computing for IoT-Enabled Smart Grid: The Future of Energy. Energies, 15(17), 6140. https://doi.org/10.3390/en15176140

[9] Espina, E., Llanos, J., Burgos-Mellado, C., Cárdenas, R., Martínez-Gómez, M., & Sáez, D. (2020). Distributed Control Strategies for Microgrids: An Overview. IEEE Access, 8, 193412-193448. https://doi.org/10.1109/access.2020.3032378

[10] Chen, Y., & Xu, J. (2022). Solar and wind power data from the Chinese State Grid Renewable Energy Generation Forecasting Competition. Scientific Data, 9(1), 577. https://doi.org/10.1038/s41597-022-01696-6

[11] Ullah, M. H., Eskandarpour, R., Zheng, H., & Khodaei, A. (2022). Quantum computing for smart grid applications. IET Generation Transmission & Distribution, 16(21), 4239-4257. https://doi.org/10.1049/gtd2.12602

[12] Zhou, Y., Tang, Z., Nikmehr, N., Babahajiani, P., Feng, F., Wei, T., Zheng, H., & Zhang, P. (2022). Quantum computing in power systems. iEnergy, 1(2), 170-187. https://doi.org/10.23919/ien.2022.0021

[13] Maneesha, A., & Swarup, K. S. (2021). A survey on applications of Alternating Direction Method of Multipliers in smart power grids. Renewable and Sustainable Energy Reviews, 152, 111687. https://doi.org/10.1016/j.rser.2021.111687

[14] Patari, N., Venkataramanan, V., Srivastava, A. K., Molzahn, D. K., Li, N., & Annaswamy, A. M. (2021). Distributed Optimization in Distribution Systems: Use Cases, Limitations, and Research Needs. IEEE Transactions on Power Systems, 37(5), 3469-3481. https://doi.org/10.1109/tpwrs.2021.3132348

[15] Li, Q., Liao, Y., Wu, K., Zhang, L., Lin, J., Chen, M., Guerrero, J. M., & Abbott, D. (2021). Parallel and Distributed Optimization Method With Constraint Decomposition for Energy Management of Microgrids. IEEE Transactions on Smart Grid, 12(6), 4627-4640. https://doi.org/10.1109/tsg.2021.3097047

[16] Zhao, Z., Guo, J., Luo, X., Lai, C. S., Yang, P., Lai, L. L., Li, P., Guerrero, J. M., & Shahidehpour, M. (2022). Distributed Robust Model Predictive Control-Based Energy Management Strategy for Islanded Multi-Microgrids Considering Uncertainty. IEEE Transactions on Smart Grid, 13(3), 2107-2120. https://doi.org/10.1109/tsg.2022.3147370

[17] Rajaei, A., Fattaheian-Dehkordi, S., Fotuhi-Firuzabad, M., & Moeini-Aghtaie, M. (2021). Decentralized transactive energy management of multi-microgrid distribution systems based on ADMM. International Journal of Electrical Power & Energy Systems, 132, 107126. https://doi.org/10.1016/j.ijepes.2021.107126

[18] Wu, T., Zhao, C., & Zhang, Y. A. (2021). Distributed AC-DC Optimal Power Dispatch of VSC-Based Energy Routers in Smart Microgrids. IEEE Transactions on Power Systems, 36(5), 4457-4470. https://doi.org/10.1109/tpwrs.2021.3056768

[19] Cerezo, M., Arrasmith, A., Babbush, R., Benjamin, S. C., Endo, S., Fujii, K., McClean, J. R., Mitarai, K., Yuan, X., Cincio, L., & Coles, P. J. (2021). Variational quantum algorithms. Nature Reviews Physics, 3(9), 625-644. https://doi.org/10.1038/s42254-021-00348-9

[20] Egger, D. J., Mareček, J., & Woerner, S. (2021). Warm-starting quantum optimization. Quantum, 5, 479. https://doi.org/10.22331/q-2021-06-17-479

[21] Fuchs, F., Lye, K. O., Nilsen, H. M., Stasik, A., & Sartor, G. (2022). Constraint Preserving Mixers for the Quantum Approximate Optimization Algorithm. Algorithms, 15(6), 202. https://doi.org/10.3390/a15060202

[22] Weidenfeller, J., Valor, L. C., Gacon, J., Tornow, C., Bello, L., Woerner, S., & Egger, D. J. (2022). Scaling of the quantum approximate optimization algorithm on superconducting qubit based hardware. Quantum, 6, 870. https://doi.org/10.22331/q-2022-12-07-870

[23] Nikmehr, N., Zhang, P., & Bragin, M. A. (2022). Quantum Distributed Unit Commitment: An Application in Microgrids. IEEE Transactions on Power Systems, 37(5), 3592-3603. https://doi.org/10.1109/tpwrs.2022.3141794

[24] Cerezo, M., Sone, A., Volkoff, T., Cincio, Ł., & Coles, P. J. (2021). Cost function dependent barren plateaus in shallow parametrized quantum circuits. Nature Communications, 12(1), 1791. https://doi.org/10.1038/s41467-021-21728-w

[25] Sack, S., & Serbyn, M. (2021). Quantum annealing initialization of the quantum approximate optimization algorithm. Quantum, 5, 491. https://doi.org/10.22331/q-2021-07-01-491

[26] Sweke, R., Wilde, F., Meyer, J., Schuld, M., Faehrmann, P. K., Meynard-Piganeau, B., & Eisert, J. (2020). Stochastic gradient descent for hybrid quantum-classical optimization. Quantum, 4, 314. https://doi.org/10.22331/q-2020-08-31-314

[27] Larkin, J. M., Jonsson, M., Justice, D., & Guerreschi, G. G. (2022). Evaluation of QAOA based on the approximation ratio of individual samples. Quantum Science and Technology, 7(4), 045014. https://doi.org/10.1088/2058-9565/ac6973

[28] Gilliam, A., Woerner, S., & Gonciulea, C. (2021). Grover Adaptive Search for Constrained Polynomial Binary Optimization. Quantum, 5, 428. https://doi.org/10.22331/q-2021-04-08-428

[29] Li, Y., Tian, M., Liu, G., Cheng, P., & Jiao, L. (2020). Quantum Optimization and Quantum Learning: A Survey. IEEE Access, 8, 23568-23593. https://doi.org/10.1109/access.2020.2970105

[30] Bova, F., Goldfarb, A., & Melko, R. G. (2021). Commercial applications of quantum computing. EPJ Quantum Technology, 8(1), 2. https://doi.org/10.1140/epjqt/s40507-021-00091-1

[31] Si, Y., Wang, R., Zhang, S., Zhou, W., Lin, A., & Zeng, G. (2022). Configuration optimization and energy management of hybrid energy system for marine using quantum computing. Energy, 253, 124131. https://doi.org/10.1016/j.energy.2022.124131

[32] Braine, L., Egger, D. J., Glick, J. R., & Woerner, S. (2021). Quantum Algorithms for Mixed Binary Optimization Applied to Transaction Settlement. IEEE Transactions on Quantum Engineering, 2, 1-8. https://doi.org/10.1109/tqe.2021.3063635

[33] Jain, N., Coyle, B., Kashefi, E., & Kumar, N. (2022). Graph neural network initialisation of quantum approximate optimisation. Quantum, 6, 861. https://doi.org/10.22331/q-2022-11-17-861