The International Maritime Transport and Logistics (MARLOG 15) - ISSN 2974-3141
http://dx.doi.org/10.21622/MARLOG.2026.15.1.123
Smart & Resilient Logistics Corridors Future Gateway to the Green Trade

AI-, DIGITAL-TWIN-, AND QUANTUM-ENABLED ROLLING-HORIZON OPTIMIZATION FOR YARD ALLOCATION AND GATE APPOINTMENT COMPLIANCE IN MODULAR CONTAINER TERMINALS

Roberto Revetria, Anastasiia Rozhok

Department of Political and International Sciences, University of Genoa, Piazza Emanuele Brignole 3A, Central Tower, Genoa, 16126, Italy

Keywords: Maritime Renewable Energy Solutions, Sustainable shipping, Maritime decarbonization, Zero-emission maritime transport, Bibliometrics Analysis

ABSTRACT

Container terminals operate under increasingly tight capacity, energy, and service-level constraints, while facing growing uncertainty in arrival patterns, handling operations, and hinterland interactions. This paper proposes a hybrid decision-support framework that integrates advanced Artificial Intelligence techniques, Digital Twin modeling, and quantum-inspired optimization to address yard planning and operational optimization in container terminals. The core problem is formulated as a block-to-area assignment under soft and hard capacity constraints, priority rules, and rolling-horizon forecasts. The problem is mapped to a Quadratic Unconstrained Binary Optimization (QUBO) formulation and solved using both classical and quantum-inspired approaches, including the Quantum Approximate Optimization Algorithm (QAOA). Two alternative modeling strategies are investigated: a one-hot encoding with deterministic feasibility repair, and a feasibility-by-design binary encoding that reduces the number of decision variables and enforces assignment constraints structurally. A Digital Twin of the container yard acts as a truth model and constraint oracle, translating abstract optimization decisions into operational Key Performance Indicators such as reshuffling effort, crane interference, and truck waiting times. Computational experiments on a modular yard scenario highlight the trade-offs between encoding choice, feasibility, solution quality, and computational cost under time-bounded quantum-inspired optimization. The results demonstrate that quantum-inspired methods can effectively support combinatorial decision-making in port operations when combined with appropriate encodings and simulation-based validation. Rather than replacing existing planning tools, the proposed framework positions quantum optimization as a complementary accelerator within a scalable, explainable, and industry-ready Digital Twin-based decision-support architecture.

1. PROBLEM CONTEXT AND RELEVANCE

1.1 Operational delays in container terminals and systemic impacts

Container terminals represent critical nodes of global supply chains, acting as interfaces between maritime transport and inland logistics systems. Even small inefficiencies in terminal operations can propagate downstream, generating significant economic, industrial, and societal impacts. In particular, delays in container handling and prolonged dwell times directly affect vessel turnaround, truck waiting times, and inventory availability for manufacturing and commercial activities.

At a global level, major container ports handle extremely large volumes. Leading European ports such as Rotterdam, Antwerp-Bruges, Hamburg, Valencia, and Gioia Tauro each process between 3 and 15 million TEU per year, while major North American ports such as Los Angeles, Long Beach, New York-New Jersey, Savannah, and Vancouver handle comparable volumes. Across these systems, even a modest increase of one day in average container dwell time can immobilize hundreds of thousands of TEU at any given time, translating into billions of euros or dollars of working capital tied up in inventories.

From an operational perspective, prolonged container dwell times increase yard congestion, reduce effective storage capacity, and exacerbate re-handling (reshuffling) operations, which in turn degrade crane productivity and increase energy consumption. At the gate, congestion caused by misaligned appointments or delayed container availability leads to truck queues, higher emissions, and increased logistics costs. From the perspective of shipping lines and terminal operators, delays trigger demurrage and detention penalties, erode service reliability, and reduce the attractiveness of ports within competitive port systems.

From the standpoint of industrial and commercial users, unreliable container availability disrupts production schedules, forces higher safety stocks, and weakens the responsiveness of supply chains. Recent disruptions have highlighted how port congestion and terminal inefficiencies can rapidly amplify into broader supply-chain crises, affecting sectors ranging from automotive and electronics to food and pharmaceuticals. Consequently, reducing container processing times and delays is not only an operational objective but also a strategic requirement for port competitiveness and supply-chain resilience.

1.2 The planning challenge: uncertainty, congestion, and rolling decisions

Container terminal operations are inherently stochastic. Arrival times of vessels and trucks, container dwell times, and block sizes are subject to uncertainty that increases with the planning horizon. At the same time, terminals must make frequent planning decisions under tight time constraints, often revising plans on a daily or even intra-day basis as new information becomes available.

Traditional static planning approaches struggle to cope with this environment. Optimizing plans over long horizons with deterministic assumptions often leads to solutions that rapidly become infeasible or inefficient when forecasts change. Conversely, purely reactive approaches may ensure feasibility but sacrifice global efficiency and service quality.

The problem is further complicated by the strong coupling between yard allocation, crane productivity, and gate operations. Decisions taken to optimize one subsystem (e.g., aggressive yard saturation to maximize storage utilization) can have adverse effects on others (e.g., increased re-handling, slower truck service, missed vessel cut-offs). As a result, there is a need for integrated, adaptive planning approaches that explicitly account for uncertainty, congestion, and service-level constraints.

2. APPLICATION SCENARIO: MODULAR YARD ALLOCATION WITH GATE APPOINTMENTS

2.1 Terminal structure and modular yard design

We consider a container terminal with two primary flows: import (sea → yard → landside) and export (landside → yard → sea). All containers are assumed to be 20-foot units (TEU) for simplicity.

The yard is organized into autonomous modular storage areas indexed by $a \in A$. Each area is characterized by:

- ground dimensions (L_a , W_a) expressed in container slots,
- a maximum stacking height H_a (tiers),
- a set of rail-mounted gantry cranes (RMGs), with at most N_a cranes operating simultaneously in the same area.

The theoretical maximum capacity of area a is:

$$C_a^{\text{theo}} = H_a \cdot L_a \cdot W_a. \tag{1}$$

To ensure operational accessibility and limit excessive reshuffling, a maximum saturation coefficient $S_a \in (0, 1]$ is defined. This yields a soft operational capacity:

$$C_a^{\text{sat}} = S_a \cdot C_a^{\text{theo}}. \tag{2}$$

Exceeding C_a^{sat} is allowed locally at the cost of reduced productivity and increased re-handling, whereas exceeding C_a^{theo} is strictly forbidden.

2.2 Blocks, arrivals, and departures

Containers are grouped into homogeneous blocks (orders) indexed by $b \in B$. All containers belonging to the same block arrive together and depart together. The number of containers in block b, denoted by Q_b, is a random variable with known mean Q and standard deviation SQ.

Each block has a planned arrival date A_b and a planned departure date P_b, both subject to uncertainty. The dwell time $\Delta_b = P_b - A_b$ has mean T days and standard deviation ST. Export blocks are subject to vessel cut-off times, while import blocks are subject to delivery commitments and free-time limits.

Planning is performed on a daily basis. At each day t, the terminal receives an updated rolling forecast for the next D days, including expected arrivals, departures, and block sizes. Forecast uncertainty is lower for near-term events and increases with the forecast horizon.

2.3 Assignment constraints and autonomy of yard areas

Each block must be assigned to exactly one yard area and must remain in that area for its entire dwell time. This constraint reflects the modular design of the terminal and the operational preference to avoid inter-area relocations.

Let $x_{b,a}$ be a binary decision variable equal to 1 if block b is assigned to area a. The assignment constraint is:

$$\sum_{a \in \mathcal{A}} x_{b,a} = 1 \qquad \forall b \in \mathcal{B}. \tag{3}$$

Given the assignment, the daily occupancy of area a is determined by the set of blocks present during each day and their sizes. Capacity constraints must be satisfied over time, accounting for both hard theoretical limits and soft saturation thresholds.

2.4 Gate appointment system and service constraints

Landside trucks (trailers) access the terminal through a gate appointment system. Each truck is assigned a specific appointment time within the day for container delivery (export) or pickup (import). Appointments are designed to smooth arrivals and limit congestion.

Let $Q^{gate}(t,h)$ denote the number of trucks waiting at the gate on day t at intra-day time h. To ensure acceptable service levels and limit emissions and congestion, the number of waiting trucks should not exceed a predefined threshold R. Exceedances may be penalized or controlled through probabilistic (chance) constraints.

Truck waiting time at the gate represents a critical performance indicator and directly affects logistics costs and terminal attractiveness.

2.5 Performance objective

The primary objective of the terminal is to minimize the average delay experienced by containers relative to their planned arrival and departure dates. Delays include late availability of export containers at vessel cut-offs and late delivery of import containers to landside customers. Export flows are prioritized due to space scarcity and contractual obligations with shipping lines.

Secondary objectives include minimizing truck waiting times at the gate, limiting yard congestion and re-handling, and avoiding demurrage and detention penalties. These objectives are inherently stochastic due to uncertainty in arrivals, departures, and service processes.

3. STATE OF THE ART: COMPUTATIONAL LIMITS AND EMERGING PARADIGMS

3.1 Classical optimization: expressiveness versus scalability

The majority of planning and control problems arising in container terminal operations can be formulated as mixed-integer programming (MIP) or mixed-integer linear programming (MILP) models. These formulations are attractive due to their expressive power, allowing the explicit modeling of assignment decisions, capacity constraints, sequencing, and precedence relations. In particular, integrated models have been proposed to jointly optimize quay crane schedules, internal transport dispatching, and yard crane operations.

However, the computational complexity of such formulations grows rapidly with the number of containers, equipment units, and time periods. Yard allocation problems with time-dependent occupancy constraints already induce large binary decision spaces; when combined with rolling horizons and uncertainty, exact solution approaches quickly become impractical for operational use. As a result, most MIP-based approaches rely on problem decomposition, relaxations, or truncated horizons, which may compromise global optimality and robustness.

In the context of daily re-planning, where solutions must be computed within minutes, the trade-off between model fidelity and solvability becomes a fundamental limitation of classical exact optimization methods.

3.2 Metaheuristics: flexibility at the cost of predictability

To address scalability issues, a wide range of metaheuristic and hybrid approaches have been applied to container terminal problems. Genetic algorithms, tabu search, simulated annealing, and variable neighborhood search are commonly used to generate high-quality solutions within reasonable time limits. Hybrid strategies combining heuristics with local exact optimization or rolling-horizon frameworks are also widespread.

While metaheuristics offer flexibility and scalability, they present several limitations in the present context. First, their performance is highly sensitive to parameter tuning and problem-specific design choices. Second, they provide limited guarantees on solution quality and feasibility under uncertainty. Third, when applied repeatedly in rolling-horizon settings, their cumulative computational cost may become prohibitive, particularly when extensive simulation-based evaluation is required.

3.3 Reinforcement learning: adaptivity under constraints

Reinforcement learning (RL) has gained significant attention as a means of learning adaptive control policies for complex, stochastic systems such as container terminals. RL approaches are particularly attractive for real-time dispatching and dynamic decision-making, as they can react to system states without solving an optimization problem from scratch at each decision point.

Nevertheless, several challenges limit the direct applicability of RL to yard allocation and gate appointment compliance problems. RL methods typically require large numbers of training episodes, leading to sample inefficiency when high-fidelity simulations are used. Moreover, enforcing hard constraints - such as yard capacity limits or appointment compliance - is nontrivial and often handled through penalty terms or action masking, which can destabilize training or lead to conservative policies. Finally, the interpretability of learned policies and their robustness to distributional shifts remain open concerns in safety-critical logistics environments.

3.4 Quantum and quantum-inspired optimization: a pragmatic perspective

Quantum computing has recently emerged as a promising paradigm for combinatorial optimization, particularly through Quadratic Unconstrained Binary Optimization (QUBO) and Ising model formulations. From a practical standpoint, current quantum hardware does not offer a general-purpose replacement for classical optimization solvers due to limitations in qubit count, noise, and connectivity.

However, quantum and quantum-inspired approaches can be viewed as *combinatorial accelerators* rather than universal solvers. In this role, they are well suited to generating high-quality candidate solutions for structured subproblems, such as block-to-area assignment or equipment allocation, where the decision space is discrete and highly combinatorial. Hybrid workflows, in which quantum or quantum-inspired solvers produce candidate assignments that are subsequently refined or validated by classical methods, represent a realistic and effective near-term strategy.

3.5 Digital twins as constraint oracles and truth models

Digital twins based on discrete-event simulation provide a high-fidelity representation of terminal operations, capturing congestion effects, resource interactions, and stochastic variability that are difficult to model analytically. Rather than serving merely as visualization tools, digital twins can act as *constraint oracles* and *truth models*, evaluating the feasibility and performance of candidate plans under realistic operating conditions.

In the proposed framework, the digital twin plays a central role by: (i) validating candidate solutions generated by optimization or learning components, (ii) providing unbiased performance estimates under uncertainty, and (iii) generating synthetic data for training surrogate models and reinforcement learning agents. By combining fast, low-fidelity simulation for screening with high-fidelity digital twins for validation, it is possible to balance computational efficiency and modeling accuracy.

3.6 Integrated perspective

The limitations of individual approaches motivate an integrated paradigm. Classical optimization provides structure and constraint awareness but lacks scalability under uncertainty. Metaheuristics offer scalability but limited robustness guarantees. Reinforcement learning enables adaptivity but struggles with sample efficiency and hard constraints. Quantum optimization can accelerate discrete search but requires hybridization. Digital twins provide realism and validation but are computationally expensive if used alone.

This work adopts an integrated architecture in which digital twins serve as the reference model of reality, machine learning provides prediction and surrogate modeling capabilities, quantum or quantum-inspired solvers accelerate combinatorial decision-making, and classical optimization enforces structure and feasibility. Such a combination is particularly well suited to rolling-horizon container terminal planning, where decisions must be both fast and reliable under uncertainty.

4. FORMAL MATHEMATICAL FORMULATION

4.1 Sets and indices

- A: set of yard areas (modules), index a.
- B: set of container blocks (orders), index b.
- T: discrete set of planning days, index t.
- H: discrete set of intra-day time slots (e.g., hours) for gate operations, index h.
- K: set of trucks (trailers), index k.

4.2 Parameters

- L_a, W_a: ground dimensions (slots) of area a.
- H_a: maximum stacking height (tiers) of area a.
- $C_a^{theo} = H_a L_a W_a$: hard theoretical capacity of area a.
- $S_a \in (0, 1]$: operational saturation coefficient of area a.
- $C_a^{sat} = S_a C_a^{theo}$: soft capacity of area a.
- N_a: maximum number of RMG cranes operating simultaneously in area a.
- Q_b: random variable representing the number of containers in block b.
- A_b: random variable representing the arrival day of block b.
- P_b: random variable representing the departure day of block b.
- τ_b: target completion time for block b (cut-off or delivery deadline).
- R: maximum acceptable number of trucks waiting at the gate.
- w_b: priority weight of block b (higher for export blocks).
- α, μ, η : nonnegative penalty coefficients.

4.3 Decision variables

- $x_{b,a} \in \{0, 1\}$: equals 1 if block b is assigned to area a.
- $u_a(t) \geq 0$: slack variable representing soft-capacity violation in area a on day t.

4.4 Auxiliary variables

- $I_b(t) \in \{0, 1\}$: equals 1 if block b occupies yard space on day t.
- $O_a(t)$: total occupancy of area a on day t.
- $Q^{gate}(t, h)$: number of trucks waiting at the gate at day t, time slot h.
- W_k: waiting time of truck k at the gate.

- c_b: actual completion time of block b.
- T_b = max(0, c_b − τ_b): tardiness of block b.

4.5 Block assignment constraints

Each block must be assigned to exactly one yard area:

$$\sum_{a \in \mathcal{A}} x_{b,a} = 1 \qquad \forall b \in \mathcal{B}. \tag{4}$$

4.6 Yard occupancy dynamics

Block presence is defined as:

$$I_b(t) = \begin{cases} 1, & \text{if } A_b \le t < P_b, \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

The occupancy of area a on day t is:

$$O_a(t) = \sum_{b \in \mathcal{B}} x_{b,a}\, Q_b\, I_b(t). \tag{6}$$

4.7 Capacity constraints

Hard capacity constraints:

$$O_a(t) \le C_a^{\text{theo}} \qquad \forall a \in \mathcal{A}, \forall t \in \mathcal{T}. \tag{7}$$

Soft saturation constraints:

$$O_a(t) \le C_a^{\text{sat}} + u_a(t) \qquad \forall a \in \mathcal{A}, \forall t \in \mathcal{T}. \tag{8}$$

4.8 Gate congestion constraints

Gate congestion may be controlled either through penalties or probabilistic constraints. A soft constraint formulation penalizes excess queue length:

$$\max\left(0, Q^{\text{gate}}(t,h) - R\right) \qquad \forall t \in \mathcal{T}, \forall h \in \mathcal{H}. \tag{9}$$

Alternatively, a chance constraint can be imposed:

$$\mathbb{P}\left(Q^{\text{gate}}(t,h) \le R \;\; \forall h \in \mathcal{H}\right) \ge 1 - \varepsilon. \tag{10}$$

4.9 Objective function

The stochastic objective minimizes expected tardiness, truck waiting, yard congestion, and soft-capacity violations:

$$\min_{x,u} \; \mathbb{E}\left[\sum_{b \in \mathcal{B}} w_b T_b + \mu \sum_{k \in \mathcal{K}} W_k + \eta \sum_{t \in \mathcal{T}} \sum_{h \in \mathcal{H}} \max\left(0, Q^{\text{gate}}(t,h) - R\right) + \alpha \sum_{a \in \mathcal{A}} \sum_{t \in \mathcal{T}} u_a(t)\right], \tag{11}$$

subject to constraints (4)-(8).

4.10 Rolling-horizon structure

At each planning day t, decisions are computed for a finite horizon $t, \ldots, t + D$ based on the current forecast. Only decisions related to newly arriving blocks are fixed, while future decisions may be revised as updated information becomes available. The optimization is repeatedly solved in a receding-horizon fashion, ensuring adaptability to uncertainty and forecast updates.

4.11 Discussion

The above formulation defines a stochastic, time-coupled assignment and congestion-control problem with hard and soft constraints. Exact solution via classical MIP becomes rapidly intractable for realistic terminal sizes and rolling-horizon operation. This structure motivates the hybrid solution approach proposed in this work, in which digital twins provide unbiased evaluation of stochastic performance, machine learning supports forecasting and surrogate modeling, and quantum or quantum-inspired solvers accelerate the solution of the discrete assignment subproblem.

5. CASE STUDY AND IMPLEMENTATION: YARD LAYOUT, QUBO FORMULATION, AND SOLUTION STRATEGIES

5.1 Yard layout and physical capacity assumptions

The container terminal yard is modeled as a modular system composed of autonomous storage areas. Each area corresponds to a rectangular yard block served by rail-mounted gantry cranes (RMGs) and is assumed to be internally homogeneous in terms of storage geometry and handling capabilities. Containers are grouped into homogeneous blocks and each block must remain within the same area for its entire dwell time.

Each yard area $a \in$ A is characterized by three physical parameters:

- L_a: number of ground slots along the longitudinal direction,
- W_a: number of ground slots along the transversal direction,
- H_a: maximum stacking height.

All containers are assumed to be standard 20-foot units (TEU). The theoretical maximum storage capacity of area a is therefore:

$$C_a^{\text{theo}} = L_a \cdot W_a \cdot H_a \quad [\text{TEU}].$$

Reference layout used in the case study

For the case study considered in this paper, all yard areas are assumed to have identical geometry:

$$L_a = 1, \quad W_a = 1, \quad H_a = 12 \qquad \forall a \in \mathcal{A}.$$

This choice corresponds to a vertical-stack abstraction, where each area represents an aggregate stack with a maximum theoretical capacity of:

$$C_a^{\text{theo}} = 12 \text{ TEU}.$$

This abstraction is intentionally adopted to isolate the strategic block-to-area assignment problem and to keep the instance size compatible with quantum and quantum-inspired optimization methods. It does not limit the generality of the formulation, which directly extends to realistic yard blocks with larger values of L_a and • W_a.

Operational saturation and accessibility.

In real terminal operations, operating at full theoretical capacity significantly degrades accessibility and crane productivity due to re-handling and reshuffling requirements. To capture this effect, an operational saturation coefficient $S_a \in$ (0, 1] is introduced. The corresponding soft (operational) capacity is:

$$C_a^{\text{sat}} = S_a \cdot C_a^{\text{theo}}.$$

In the case study, the following value is adopted:

$$S_a = 0.70 \quad \Rightarrow \quad C_a^{\text{sat}} = 8.4 \text{ TEU}.$$

Exceeding C_a^{sat} is allowed but penalized, while exceeding C_a^{theo} is strictly forbidden.

5.2 Application scenario and decision variables

The terminal handles both import (sea-to-land) and export (land-to-sea) container flows. Containers are grouped into blocks (orders), each characterized by:

- a forecasted size $\hat{Q}_b$ (TEU),
- a forecasted arrival and departure window,
- a priority level, with export blocks receiving higher priority.

Planning is performed on a daily rolling horizon. At each day t_0, decisions are computed for a finite horizon T_{t_0} = {t_0, ..., t_0 + D} using the most recent forecast.

The core decision variable is:

$$z_{b,a} = \begin{cases} 1 & \text{if block } b \text{ is assigned to area } a, \\ 0 & \text{otherwise.} \end{cases}$$

Each block must be assigned to exactly one area for its entire dwell time.

5.3 QUBO formulation of the assignment problem

The constrained assignment problem is transformed into a Quadratic Unconstrained Binary Optimization (QUBO) problem by embedding constraints into quadratic penalty terms.

One-hot assignment constraint

For each block b, the constraint:

$$\sum_{a \in \mathcal{A}} z_{b,a} = 1$$

is enforced via the penalty:

$$P_{\text{A}} = \lambda_{\text{A}} \sum_b \left(1 - \sum_a z_{b,a}\right)^2,$$

where λ_A is a sufficiently large penalty coefficient.

Forecasted occupancy

Using forecasted block sizes $\hat{Q}_b$ and presence indicators $\hat{I}_b(t)$, the forecasted occupancy of area a on day t is:

$$\hat{O}_a(t) = \sum_b \hat{Q}_b \, \hat{I}_b(t) \, z_{b,a}.$$

Capacity constraints via slack variables

Soft and hard capacity constraints are enforced by introducing nonnegative slack variables $s_{a,t}^S$ and $s_{a,t}^H$:

$$\hat{O}_a(t) - C_a^{\text{sat}} - s_{a,t}^{\text{S}} = 0, \qquad \hat{O}_a(t) - C_a^{\text{theo}} - s_{a,t}^{\text{H}} = 0.$$

Slack variables are encoded in binary using M bits:

$$s = \sum_{m=0}^{M-1} 2^m y_m, \qquad y_m \in \{0,1\}.$$

The corresponding penalty terms are weighted by λ_S (soft capacity) and λ_H (hard capacity), with $\lambda_H \gg \lambda_S$.

Proxy objective

A linear proxy cost $g_{b,a}$ is introduced to guide the assignment based on export priority, expected delay risk, or historical suitability:

$$P_{\text{obj}} = \sum_{b,a} g_{b,a} z_{b,a}.$$

Final QUBO

The resulting unconstrained objective is:

$$\min_{x \in \{0,1\}^N} P_{\text{obj}} + P_{\text{A}} + P_{\text{S}} + P_{\text{H}} = x^\top Q x + c^\top x + d,$$

where x collects assignment and slack variables.

5.4 Mapping from QUBO variables to qubits

Each binary variable corresponds to one qubit. For example, with A = 2 areas and B = 4 blocks, the assignment variables are mapped as:

$[z_{1,1}, z_{1,2}, z_{2,1}, z_{2,2}, z_{3,1}, z_{3,2}, z_{4,1}, z_{4,2}]$,

which correspond to qubits 1 through 8. Slack variables are mapped to subsequent qubits.

The QUBO energy defines a diagonal cost Hamiltonian $\widehat{H_C}$ such that:

$$\hat{H}_C|x\rangle = E(x)|x\rangle,$$

where $|x\rangle$ is a computational basis state. Quantum optimization algorithms search for low-energy bitstrings corresponding to high-quality assignments.

5.5 Solution strategies

Two complementary solution strategies are considered.

Classical QUBO optimization

For full-scale instances, the QUBO is solved using classical metaheuristics (e.g., Tabu Search). This approach scales well and provides robust baseline solutions suitable for operational deployment.

Quantum-inspired optimization

For reduced instances with a limited number of binary variables, the same QUBO is solved using the Quantum Approximate Optimization Algorithm (QAOA) as implemented in MATLAB. Due to the exponential growth of statevector simulation, this strategy is applied to small subproblems and used as a candidate generator. Candidate solutions are subsequently validated and ranked using high-fidelity digital twin simulation.

5.6 Role of the digital twin

The digital twin serves as a constraint oracle and truth model, evaluating the operational feasibility and true performance of candidate assignments under stochastic conditions. This separation enables efficient combinatorial optimization while preserving realism and robustness in decision-making.

6. EXPERIMENTAL RESULTS: QAOA-BASED CANDIDATE GENERATION WITH FEASIBILITY RESTORATION VS. FEASIBILITY-BY-DESIGN ENCODING

6.1 Experimental setup and evaluation metrics

We report a controlled experiment on a reduced, yet physically interpretable, modular yard assignment instance that is compatible with statevector-based QAOA simulation. The goal is to compare two quantum-inspired pipelines:

- Solution 2 (S2): QAOA on a one-hot assignment QUBO ($z_{b,a}$ variables) followed by a deterministic one-hot repair that projects the sampled bitstring to a feasible assignment.
- Solution 3 (S3): QAOA on a feasibility-by-design encoding where, for A = 2 areas, each block uses a single binary variable x_b ("area-2 if x_b=1, else area-1"), thereby eliminating the one-hot constraint.

Both approaches include soft and hard capacity terms via binary-encoded slack variables, and are solved with the same QAOA configuration (NumLayers = 2, NumShots = 300). In both runs, the internal classical optimizer used to tune QAOA parameters terminated due to reaching the maximum number of function evaluations, hence solutions should be interpreted as time-bounded results.

Physical yard parameters

Each area is modeled using the vertical-stack abstraction with (L, W, H) = (1, 1, 12) TEU slots/tiers, yielding C^{theo} = 12 TEU per area. The saturation coefficient is S = 0.70, hence C^{sat} = 8.4 TEU.

Block forecasts and proxy costs

We consider B = 4 blocks with forecasted sizes $\hat{Q}$ = [6, 5, 4, 3] TEU and a one-day horizon (D = 1). The proxy assignment costs favor blocks 1-2 in area 1 and blocks 3-4 in area 2:

$$g = \begin{bmatrix} 0 & 0.2 \\ 0 & 0.2 \\ 0.2 & 0 \\ 0.2 & 0 \end{bmatrix}.$$

Capacity penalties. Soft and hard capacities are enforced through quadratic penalties with λ_S= 2 and λ_H = 60; slack variables are encoded with M = 2 bits per constraint (values 0-3). In S2, the one-hot penalty is set to λ_A = 20.

Evaluation metrics. To compare solutions we report: (i) feasibility (one-hot satisfaction), (ii) per-area occupancy and soft/hard exceedance, (iii) proxy objective value (lower is better), (iv) number of binary variables (qubits under QAOA), (v) wall-clock time. Note that the raw QUBO BestFunctionValue values are not directly comparable across encodings because the two QUBOs differ in variable sets and constant offsets.

6.2 Results

Table 1 summarizes the outcomes.

Table 1: Comparison of S2 (QAOA + repair) vs. S3 (A=2 feasibility-by-design encoding).

Metric	S2: QAOA + repair	S3: 1-bit encoding
Binary variables (qubits)	16 (8 assign + 4 soft + 4 hard)	12 (4 assign + 8 slack)
QAOA configuration	Layers=2, Shots=300	Layers=2, Shots=300
One-hot feasible (raw)	No	Yes (by construction)

Metric	S2: QAOA + repair	S3: 1-bit encoding
One-hot feasible (final)	Yes (after repair)	Yes
Final assignment Z	$\begin{bmatrix}1 & 0\\1 & 0\\0 & 1\\0 & 1\end{bmatrix}$	$\begin{bmatrix}0 & 1\\1 & 0\\1 & 0\\1 & 0\end{bmatrix}$
Area 1 occupancy (TEU)	11	12
Area 2 occupancy (TEU)	7	6
Soft exceedance (Area 1)	2.6	3.6
Soft exceedance (Area 2)	0	0
Hard exceedance (both areas)	0	0
Proxy objective	0.0	0.6
Wall-clock time (s)	193.4	65.4

6.3 Discussion

Feasibility mechanisms

S2 demonstrates that QAOA may return bitstrings that violate the one-hot structure when λ_A is insufficient relative to the scale of competing capacity penalties and/or when the QAOA parameter search is prematurely terminated. The deterministic repair step resolves this by projecting the sampled assignment onto the feasible set, yielding a one-hot compliant solution without violating hard capacity. In contrast, S3 enforces one-hot feasibility by design through a reduced encoding (one binary variable per block for A = 2), eliminating the need for an explicit one-hot penalty and subsequent repair.

Quality trade-offs under time-bounded QAOA

On this instance, S2 (after repair) achieves a lower proxy objective (0.0 vs. 0.6) and a smaller soft capacity exceedance in the most loaded area (2.6 vs. 3.6 TEU). This indicates that, under the current penalties and time budget, the repair-based approach preserved the intended priority structure embedded in $g_{b,a}$ while mitigating saturation-related accessibility degradation.

Computational efficiency and qubit reduction

S3 reduces the number of binary variables from 16 to 12 (25% fewer qubits for QAOA) and yields a substantially lower wall-clock time (65.4 s vs. 193.4 s) under the same QAOA configuration. This supports the key methodological point: encoding choices that reduce qubit count and enforce feasibility structurally can improve tractability in quantum-inspired workflows, especially under statevector simulation constraints.

Operational interpretation and the role of the digital twin

Both solutions respect hard capacity, while exceeding the saturation threshold in Area 1. In practice, this implies increased re-handling and reshuffling, which must be quantified via a high-fidelity digital twin (e.g., AnyLogic/SimPy) to translate soft exceedance into expected delays at crane and gate level. The QAOA/QUBO layer thus acts as a fast candidate generator, whereas the digital twin provides the operational "truth model" for ranking and robustness checks.

Next steps

Two natural extensions are: (i) penalty calibration (in particular λ_A) and QAOA tuning (layers, shots, optimizer budget) to reduce infeasible sampling and improve solution quality; (ii)

hybrid post-optimization where repaired/encoded candidates are refined via classical local search, with objective terms computed from digital-twin measurements (delay, truck queueing, and service levels).

7. PENALTY SENSITIVITY AND ENCODING ABLATION

7.1 Motivation

The experimental comparison between Solution 2 (QAOA with one-hot repair) and Solution 3 (feasibility-by-design encoding for A = 2) highlights that solution quality and feasibility are strongly affected not only by the quantum algorithm itself, but also by penalty calibration and variable encoding. This section provides a focused ablation analysis to clarify these effects and to guide reproducible parameter choices.

7.2 Sensitivity to the one-hot penalty λ_A (Solution 2)

In Solution 2, feasibility with respect to the one-hot constraint

$$\sum_{a=1}^{A} z_{b,a} = 1, \quad \forall b$$

is enforced via a quadratic penalty weighted by λ_A. The experimental results show that, with λ_A = 20, QAOA samples bitstrings that violate the one-hot structure, such as [1, 1] or [0, 0] assignments for a given block. This behavior can be explained by the relative scale of competing penalties.

Let $\hat{Q}_B$ denote the block size and λ_H the hard-capacity penalty. Violating the one-hot constraint allows a block to effectively reduce the marginal contribution to capacity penalties by distributing its load across multiple areas. When

$$\lambda_A < \mathcal{O}\left(\lambda_H \cdot \hat{Q}_b^2\right),$$

the QUBO energy landscape favors capacity relief over structural feasibility, especially under time-bounded QAOA optimization.

Implication

Increasing λ_A by one or two orders of magnitude would suppress infeasible assignments, but at the cost of a stiffer energy landscape that is harder to optimize. This motivates the separation of roles adopted in Solution 2: QAOA as a candidate generator, followed by a deterministic feasibility projection.

7.3 Encoding ablation: one-hot vs. binary (A = 2)

For A = 2 areas, the one-hot encoding with variables $z_{b,1}$, $z_{b,2}$ can be replaced by a single binary variable

$$x_b = \begin{cases} 0 & \text{block } b \text{ assigned to area 1,} \\ 1 & \text{block } b \text{ assigned to area 2.} \end{cases}$$

This removes the one-hot constraint entirely and reduces the number of assignment variables from 2B to B.

Observed effects

Compared to Solution 2, the feasibility-by-design encoding in Solution 3 yields:

- a 25% reduction in total binary variables (qubits);
- guaranteed feasibility with respect to assignment constraints;

• a significant reduction in wall-clock time under statevector QAOA simulation.

However, the reduced encoding also changes the geometry of the QUBO landscape. In the reported experiment, Solution 3 converges faster but produces a higher proxy objective and larger soft-capacity exceedance under the same QAOA time budget. This indicates that fewer variables do not automatically imply better solution quality when the optimization is prematurely terminated.

7.4 Interpretation for hybrid quantum-digital twin workflows

The ablation study supports a key methodological insight:

Encoding choices primarily control feasibility and computational tractability, while penalty calibration and quantum optimization depth control solution quality.

From a system-design perspective:

- Solution 2 is preferable when priority structures are critical and can be refined by classical post-processing and digital twin validation.
- Solution 3 is preferable when rapid generation of feasible candidates is required, for example in rolling-horizon or real-time decision support.

In both cases, the digital twin acts as the final arbiter, translating soft-capacity exceedance into operational KPIs such as re-handling moves, crane interference, truck waiting times, and vessel delays.

7.5 Scalability considerations

As the number of areas increases (A > 2), the feasibility-by-design approach generalizes to $\lceil log_2 A \rceil$ bits per block, preserving structural feasibility while controlling qubit growth. This suggests a clear research direction for scaling quantum-inspired optimization in container terminal applications.

8. DIGITAL TWIN INTEGRATION AND SCALABILITY

8.1 Role of the Digital Twin in the Optimization Loop

In the proposed framework, the Digital Twin (DT) of the container terminal plays a central and non-substitutable role. While the QUBO-QAOA layer generates candidate solutions for block-to-area assignments under simplified abstractions, the Digital Twin acts as a constraint oracle and a truth model, translating abstract capacity and priority decisions into operational performance indicators.

The DT explicitly represents: (i) yard topology and crane interference, (ii) stochastic arrival and departure processes, (iii) handling policies and reshuffling rules, (iv) truck and gate appointment dynamics, (v) vessel-side constraints and berth interactions.

As a result, each candidate assignment produced by the optimization layer can be evaluated against realistic Key Performance Indicators (KPIs), such as:

- average and tail container dwell time;
- number of re-handling and reshuffling moves;
- RMG utilization and interference delays;
- truck waiting time at gates;
- vessel demurrage and detention risk.

This separation of concerns allows the optimization layer to remain computationally tractable while preserving decision relevance.

8.2 Hybrid Quantum-Digital Twin Workflow

Figure 1 conceptually summarizes the hybrid workflow. The process unfolds as follows:

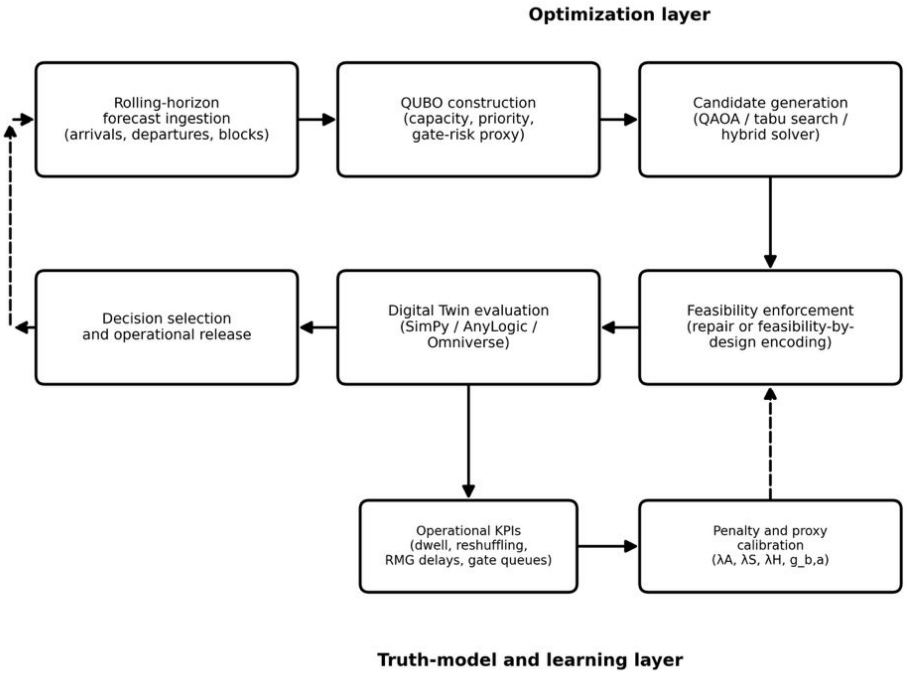

Figure 1. Hybrid Quantum-Digital Twin Workflow

1. ***Forecast ingestion.*** Rolling-horizon forecasts of block arrivals and departures are collected for the next D days.
2. ***QUBO construction.*** Forecasts are aggregated into a reduced assignment problem, encoding capacity, priority, and policy constraints.
3. ***Quantum-inspired candidate generation***. QAOA (or classical QUBO solvers) generate a limited set of candidate assignments.
4. ***Feasibility enforcement.*** Structural feasibility is guaranteed either by encoding (Solution 3) or by deterministic repair (Solution 2).
5. ***Digital Twin evaluation.*** Each candidate is simulated in the DT to compute operational KPIs.
6. ***Selection and feedback.*** The best-performing assignment is selected and KPI feedback is used to recalibrate penalties and priorities.

This loop can be executed daily, or intra-day in a rolling-horizon setting, without requiring full re-optimization from scratch.

8.3 From Soft Capacity to Operational Impact

Soft-capacity exceedance in the QUBO formulation does not directly correspond to infeasibility, but to an increased probability of operational inefficiencies. The Digital Twin bridges this semantic gap by mapping soft exceedance to measurable impacts.

For example, exceeding the saturation threshold C^{sat} in an area leads to:

- higher expected reshuffling counts;
- increased crane travel distances;
- non-linear growth in service time variance.

Through simulation, these effects can be quantified and used to construct surrogate penalty functions that replace or augment the static QUBO penalties. This enables a learning loop where Digital Twin outputs progressively inform the optimization model.

8.4 Scalability in Yard Size and Planning Horizon

The experiments reported in this work focus on reduced instances compatible with statevector-based QAOA simulation. However, the proposed methodology is explicitly designed to scale beyond these limits.

Scaling in yard size

For A > 2 areas, feasibility-by-design encodings generalize by assigning $\lceil log_2 A \rceil$ binary variables per block. While this increases the number of qubits, it avoids quadratic growth associated with one-hot encodings. Moreover, the problem naturally decomposes across independent yard modules, enabling parallel optimization.

Scaling in planning horizon

As the horizon D increases, the number of blocks grows linearly, but temporal coupling remains limited due to forecast uncertainty. This motivates:

- rolling-horizon optimization with partial re-planning;
- fixing near-term assignments while optimizing future blocks;
- scenario sampling within the Digital Twin to stress-test robustness.

Solver heterogeneity

In large-scale deployments, quantum hardware is expected to handle only small, critical subproblems (e.g., highly congested yard modules), while classical heuristics address the remaining instances. The Digital Twin provides a common evaluation layer across solvers, ensuring consistent decision criteria.

8.5 Implications for Real-Time Decision Support

The integration of QUBO-based optimization with a Digital Twin enables a new class of decision-support systems for container terminals. Rather than replacing existing planning tools, the proposed approach augments them with:

- rapid exploration of combinatorial alternatives;
- explicit handling of uncertainty and variability;
- continuous calibration through operational feedback.

This hybrid architecture aligns with emerging industrial practices, where optimization, simulation, and AI components are orchestrated to balance optimality, robustness, and explainability.

8.6 Research Outlook

Future research directions include: (i) learning QUBO penalty parameters directly from Digital Twin data, (ii) integrating reinforcement learning to guide candidate selection, (iii) exploiting quantum hardware for specific bottleneck subproblems, (iv) extending the framework to multi-terminal and hinterland-integrated scenarios.

These directions position Digital Twin-enabled quantum-inspired optimization as a promising paradigm for next-generation port operations management.

9. CONCLUSIONS

This paper investigated the integration of advanced Artificial Intelligence techniques, Digital Twin technology, and quantum-inspired optimization for the planning and operation of container terminal yards. Focusing on the block-to-area assignment problem under capacity, priority, and uncertainty constraints, we proposed a unified framework that

combines Quadratic Unconstrained Binary Optimization (QUBO) formulations, Quantum Approximate Optimization Algorithm (QAOA), and high-fidelity Digital Twin simulation.

From a methodological perspective, the study demonstrates that the effectiveness of quantum-inspired optimization in real-world logistics problems depends less on the quantum algorithm alone and more on the quality of the problem encoding and constraint handling. The comparative analysis between a one-hot encoding with deterministic feasibility repair and a feasibility-by-design binary encoding shows that reducing the number of binary variables and embedding feasibility directly into the model can substantially improve computational tractability. However, encoding choices also influence solution quality under time-bounded optimization, highlighting an inherent trade-off between speed, feasibility, and optimality.

The results further confirm the central role of the Digital Twin as a truth model and constraint oracle. While the QUBO-QAOA layer efficiently explores combinatorial alternatives, the Digital Twin translates abstract decisions into operational Key Performance Indicators, such as reshuffling effort, crane interference, and truck waiting times. This separation enables scalable decision support without oversimplifying the operational reality of container terminals.

From an application standpoint, the proposed hybrid architecture is well suited for rolling-horizon and real-time decision support in modular container terminals. Quantum-inspired solvers can be selectively applied to congested or critical subproblems, while classical heuristics and simulation-based evaluation ensure robustness and explainability. Importantly, the framework aligns with current industrial practices, positioning quantum computing as a complementary accelerator rather than a disruptive replacement of existing planning systems.

Future research directions include learning QUBO penalty parameters directly from Digital Twin data, tighter integration with reinforcement learning for adaptive policy refinement, and experimental validation on larger terminal layouts using quantum-inspired and emerging quantum hardware. Overall, this work contributes a realistic and scalable pathway for bringing quantum-inspired optimization into port and terminal operations management.

10. REFERENCES

[1] Y. Gao, D. Chang, and C. H. Chen, "A digital twin-based approach for optimizing operation energy consumption at automated container terminals," Journal of Cleaner Production, vol. 384, 2023.

[2] A. Sonavane and A. Aylani, "Exploring the use of quantum algorithms for logistics optimization," in Quantum Computing and Artificial Intelligence in Logistics, Taylor & Francis, 2023.

[3] A. ZarinchangMokalla, "Intelligent digital twin for optimizing warehouse operations: Embedded optimization components for enhanced order-picking efficiency," 2024.

[4] J. Bowles, A. Dauphin, P. Huembeli, and J. Martinez, "Quadratic unconstrained binary optimisation via quantum-inspired annealing," arXiv preprint arXiv:2108.xxxxx, 2021.

[5] J. R. Jiang and C. W. Chu, "Classifying and benchmarking quantum annealing algorithms based on quadratic unconstrained binary optimization for solving NP-hard problems," IEEE Access, vol. 11, 2023.

[6] F. Phillipson, "Quantum computing in logistics and supply chain management: An overview," arXiv preprint arXiv:2402.17520, 2024.

[7] M. A. Salam and S. A. Khan, "Simulation based decision support system for optimization: A case of Thai logistics service provider," Industrial Management & Data Systems, vol. 116, no. 5, 2016.

[8] Z. Wang, X. Zhu, and L. Wang, "Optimization of multi-equipment intelligent scheduling in automated terminals integrating quantum co-evolution and deep learning," in Proc. Int. Conf. on Artificial Intelligence Applications, Springer, 2024.

[9] A. Paul, K. Singh, C. P. Li, and O. A. Dobre, "Digital twin-aided vehicular edge network: A large-scale model optimization by quantum-DRL," IEEE Transactions on Network Science and Engineering, 2024.

[10] A. Sonavane and A. Aylani, "Exploring the use of quantum algorithms," in Artificial Intelligence in Logistics and Supply Chain, 2025.