Decomposition of Large-Scale Quadratic Unconstrained Binary Optimization Problems for Quantum Annealers and Quantum-Inspired Annealers

Jehn-Ruey Jiang   Qiao-Yi Lin

Department of Computer Science and Information Engineering
National Central University, Taiwan

Engineering Proceedings 2026, 134, 29

Presented at ECICE 2025, Yunlin, Taiwan, 14–16 November 2025

Motivation: Large QUBO Models Meet Limited Annealing Resources

- Quantum annealers and quantum-inspired annealers solve combinatorial optimization through QUBO/Ising formulations.
- Practical QUBO models often contain constraints that introduce many auxiliary variables and dense couplings.
- Large monolithic formulations may exceed device capacity or become difficult for heuristic annealers to solve effectively.
- We therefore ask: Can a constrained QUBO be decomposed into compact subproblems while retaining solution quality?

Main contribution

Two decomposition mechanisms for linear inequalities: **one-way-one-hot (1W1H)** and **slack variable range search (SVRS)**.

QUBO Formulation

A QUBO problem seeks $\boldsymbol{x} \in \{0,1\}^n$ minimizing

$$f(\boldsymbol{x}) = \boldsymbol{x}^T Q \boldsymbol{x} = \sum_i Q_{ii} x_i + \sum_{i<j} Q_{ij} x_i x_j.$$

Constraints are incorporated as penalty terms:

$$F(\boldsymbol{x}) = \alpha f(\boldsymbol{x}) + \beta g(\boldsymbol{x}).$$

- α: objective weight; β: constraint penalty weight.
- β must discourage infeasible solutions without overwhelming the original objective.
- Our target constraint is a knapsack-style inequality

$$\boldsymbol{w}^T \boldsymbol{x} \leq C.$$

Baseline Constraint Encodings

Exactly-one (one-way-one-hot) constraint:

$$\sum_{i=1}^{k} x_i = 1, \qquad \beta \left(\sum_{i=1}^{k} x_i - 1 \right)^2 .$$

Linear inequality using binary slack:

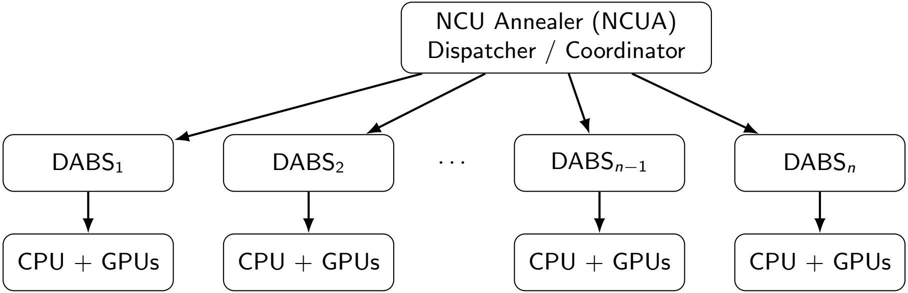
$$\boldsymbol{a}^T \boldsymbol{x} + \sum_{t=0}^{m-1} 2^t s_t = b,$$

with penalty

$$\beta \left(\boldsymbol{a}^T \boldsymbol{x} + \sum_{t=0}^{m-1} 2^t s_t - b \right)^2 ,$$

where $m \geq \lceil \log_2(b+1) \rceil$ is required to cover the full slack range.

NCUA: Distributed Quantum-Inspired Annealing Platform



- NCUA decomposes QUBO/Ising instances and dispatches subproblems to networked CPU–GPU solver nodes.
- Each node uses the Diverse Adaptive Bulk Search (DABS) engine with GPU-parallel candidate evaluation.
- Decomposition naturally matches NCUA's horizontally scalable meta-solver architecture.

1W1H Decomposition

For $\boldsymbol{w}^T\boldsymbol{x} \leq C$, introduce indicators y_k, where $y_k = 1$ means $\boldsymbol{w}^T\boldsymbol{x} = k$:

$$\beta\left(\boldsymbol{w}^T\boldsymbol{x} - \sum_{k=1}^{C} k y_k\right)^2 + \beta\left(\sum_{k=1}^{C} y_k - 1\right)^2.$$

Fix one indicator $y_k = 1$ and all others to zero. The kth independent subproblem becomes

$$\alpha\boldsymbol{x}^T Q\boldsymbol{x} + \beta(\boldsymbol{w}^T\boldsymbol{x} - k)^2, \qquad 1 \leq k \leq C.$$

Trade-off

Each subproblem is very small and highly parallelizable, but the method creates C **subproblems**. A large admissible range therefore causes a severe orchestration burden.

SVRS Decomposition

Encode slack using only m bits within a restricted window:

$$S = \sum_{t=0}^{m-1} 2^t s_t.$$

The feasible slack range is partitioned into windows of size 2^m. The number of subproblems is approximately

$$q = \left\lceil \frac{C}{2^m} \right\rceil.$$

The kth subproblem is

$$\alpha \boldsymbol{x}^T Q \boldsymbol{x} + \beta \left(\boldsymbol{w}^T \boldsymbol{x} + \sum_{t=0}^{m-1} 2^t s_t - C - 2^{km} \right)^2,$$

for $k = 0, 1, \ldots, q-1$.

Key design parameter

Larger m: fewer but larger subproblems. Smaller m: more but more compact subproblems.

SVRS Search Strategy and Early Stopping

Parallel mode

- Solve all slack windows independently.
- Suitable for distributed CPU–GPU or multi-annealer environments.
- Select the best solution across all windows.

Sequential mode

- Solve windows in the order $k = 0, 1, \ldots$.
- Track the best objective value found so far.
- Stop after p consecutive windows without improvement (e.g., $p = 3$).

Rationale used in the paper

For the knapsack formulation considered, solutions with $\boldsymbol{w}^T\boldsymbol{x}$ closer to capacity C are treated as promising candidates, motivating the window ordering.

Experimental Setup: P08 Knapsack Instance

- Dataset: 0/1 knapsack instance **P08**.
- Number of items: **24**.
- Capacity: $C = 6{,}404{,}180$.
- Item weights: **31,385–951,111**.
- Item values: **65,731–2,067,838**.
- Known optimum objective value: **13,549,094**.
- Solver used for evaluation: **Compal GPU Annealer (CGA)**.

Evaluation goal

Compare monolithic 1W1H, decomposed 1W1H, and SVRS in terms of subproblem count, variables per subproblem, practicality, and solution quality.

Experimental Results

	Original 1W1H	1W1H Decomp.	SVRS $m = 10$	SVRS $m = 23$
Subproblems	1	6,404,180	6,225	1
Variables/subproblem	6,404,204	24	34	47

- Original 1W1H is too large to run on CGA.
- 1W1H decomposition produces tiny 24-variable subproblems, but 6.4 million subproblems are impractical.
- SVRS with $m = 10$ balances both dimensions: 6,225 subproblems, 34 variables each, and reaches the known optimum 13,549,094.
- SVRS with $m = 23$ collapses to one 47-variable problem, but its solution quality is inadequate on P08.

Why SVRS Provides a Better Scalability–Quality Trade-off

1W1H

- Very simple independent subproblems
- Strong fit for massive parallelism
- Number of subproblems grows directly with C
- Best for small/medium admissible ranges

SVRS

- Window size controlled by m
- Balances subproblem count and size
- Supports parallel or sequential execution
- Enables early stopping in sequential search

Decomposition is not only about making each QUBO smaller; it must also control how many QUBOs are created.

Conclusions and Future Work

- We proposed two QUBO decomposition mechanisms for linear inequalities: **1W1H** and **SVRS**.
- 1W1H gives compact, independent subproblems but scales poorly with a large admissible range.
- SVRS windows the slack space and offers a tunable balance between subproblem size and subproblem count.
- On P08, SVRS with 10 slack variables reached the known optimum while using far fewer resources than monolithic 1W1H and far fewer subproblems than exhaustive 1W1H decomposition.

Future work

Integrate both mechanisms into NCUA and investigate adaptive/overlapping search ranges and early-stopping policies for SVRS.

Thank You

Questions?

Jehn-Ruey Jiang and Qiao-Yi Lin
National Central University