Circuit Partitioning and Transmission Cost Optimization in Distributed Quantum Circuits (IEEE TCAD)

Xinyu Chen, Zilu Chen, Pengcheng Zhu, Xueyun Cheng, and Zhijin Guan

Nantong University

March 11, 2025

Motivation: Challenges in Distributed Quantum Computing

- NISQ devices have limited qubits, preventing large-scale quantum computation.
- Distributed quantum computing partitions circuits across multiple QPUs.
- Excessive inter-QPU communication increases error rates and latency.

Background

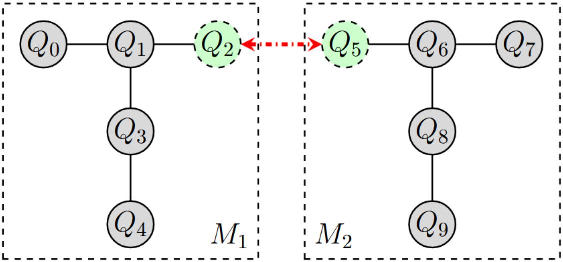

Figure: An example of a distributed quantum architecture which consists of two IBMQ quito architectures.

Methodology

- Circuit Partitioning as a Graph Cut Problem
- Dynamic Lookahead Transmission Optimization

Circuit Partitioning as a Graph Cut Problem

- Quantum circuits are represented as a qubit-weighted graph.
- Partitioning aims to minimize inter-QPU quantum gates.
- The problem is formulated as a minimum cut optimization task.

Introduction to QUBO

- **Quadratic Unconstrained Binary Optimization (QUBO)** is a mathematical model used for solving combinatorial optimization problems.
- It involves minimizing a quadratic function of binary variables, where each variable can be either 0 or 1.

QUBO Objective Function (from wiki)

The general form of the QUBO objective function is:

$$f(x) = x^T Qx = \sum_{i=1}^{n} \sum_{j=i}^{n} Q_{ij} x_i x_j$$

where:

- $x = (x_1, x_2, \ldots, x_n)$ is a vector of binary variables.
- Q is an upper-triangular matrix of real coefficients defining interactions between variables.

Example of a QUBO Problem

Consider a simple QUBO problem with three binary variables:

$$f(x) = Q_{11}x_1^2 + Q_{22}x_2^2 + Q_{33}x_3^2 + Q_{12}x_1x_2 + Q_{13}x_1x_3 + Q_{23}x_2x_3$$

Represented in matrix form:

$$Q = \begin{pmatrix} Q_{11} & Q_{12} & Q_{13} \\ 0 & Q_{22} & Q_{23} \\ 0 & 0 & Q_{33} \end{pmatrix}$$

Including Constraints in QUBO

- Constraints in QUBO are introduced as penalties.
- Penalty functions penalize infeasible solutions.
- General form with constraints:

$$f(x) = \sum_{i=1}^{n} \sum_{j=i}^{n} Q_{ij} x_i x_j + \sum_{k} P_k \cdot \text{penalty}_k(x)$$

where:

  - P_k are penalty coefficients.
  - $\text{penalty}_k(x)$ measures constraint violations.

Natural QUBO Formulations (from A Tutorial)

- The Number Partitioning Problem: For $S = \{3, 1, 1, 2, 2, 1\}$, one possible partition is $S_1 = \{1, 1, 1, 2\}$ and $S_2 = \{2, 3\}$, both summing to 5.
- The Max Cut Problem: In a triangle graph with vertices $\{A, B, C\}$ and edges $\{(A, B), (B, C), (C, A)\}$, one maximum cut is $S = \{A\}$ and $\bar{S} = \{B, C\}$, resulting in two edges crossing the cut.
- Various algorithms can solve QUBO problems, including **Quantum annealing methods using quantum computers**. (Ref. [1811.11538][s41598-019-53585-5])

Dynamic Lookahead Transmission Optimization

- Transmission qubit selection is crucial for reducing communication.
- Looks ahead dynamically to evaluate impact of transmission choices.
- Prioritizes merging quantum gate executions to minimize state transfers.

Performance Gains

- Reduces transmission cost by up to 73.85% compared to prior methods.
- Achieves significant speedup in circuit partitioning using QUBO.
- Demonstrates scalability for large circuits (up to 10,000 qubits).

Comparison with Existing Methods

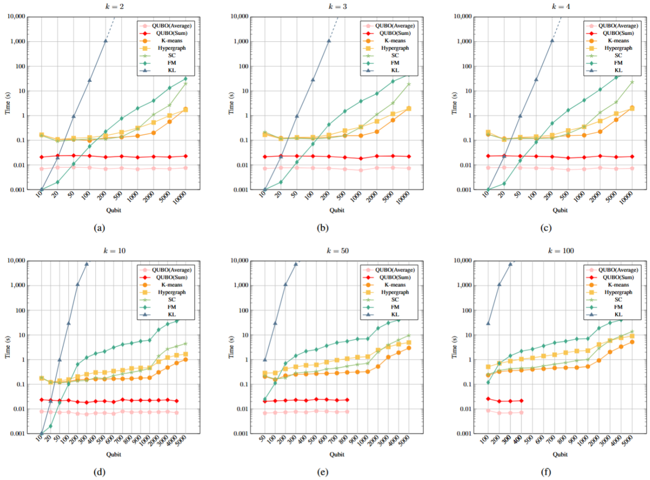

Figure: Transmission cost reduction across different methods.

Conclusion and Future Work

- Proposed QUBO-based partitioning reduces inter-QPU gates.
- Dynamic lookahead scheduling significantly lowers transmission cost.
- Approach enables more efficient execution of distributed quantum algorithms.

Future Directions

- Extend QUBO approach for more than 100 partitions.
- Incorporate error rates and latency in transmission models.
- Test method in real quantum hardware environments.