Optimal Service Placement for 6G Edge Computing with Quantum-Centric Optimisation in Real Quantum Hardware

Dang Van Huynh, *Member, IEEE,* Octavia A. Dobre, *Fellow, IEEE,* and Trung Q. Duong, *Fellow, IEEE*

***Abstract*—Mobile edge computing (MEC) is emerging as a transformative enabler for the sixth-generation (6G) wireless networks. This letter investigates the potential of a quantum-centric optimisation approach for service placement in 6G MEC. Specifically, we formulate a mixed-integer binary programming problem that aims to minimise both service costs and delay by optimising service placement decisions, subject to constraints on service availability and the computing budget of edge servers. The formulated problem is NP-hard, making it computationally challenging for classical methods to solve. To address this, we develop a quantum-centric optimisation solution that efficiently finds optimal binary solutions for the problem, demonstrating potential for tackling medium-to-large-scale instances. Simulation results validate the effectiveness of the quantum-centric approach by showcasing the convergence pattern of the optimisation process on real quantum hardware and the associated increase in running time compared to a classical method.**

***Index Terms*—6G networks, mobile edge computing, QAOA, quantum optimisation, service placement.**

I. Introduction

Mobile edge computing has been considered as an important component for 6G wireless networks, providing the advanced computing architecture needed to meet stringent 6G performance requirements, such as ultra-low latency, enhanced energy efficiency, and high reliability. By decentralising resources and bringing critical infrastructure—such as computing, storage, networking, and communication capabilities—closer to end users, edge computing allows for rapid processing and data exchange. This proximity to mobile devices enables a new wave of responsive, data-intensive services, including autonomous driving, virtual and augmented reality, the industrial Internet-of-Things (IIoT), and smart healthcare, all of which are foundational to the vision of 6G networks. As a result, edge computing holds the potential to fundamentally reshape network capabilities, extending 6G beyond traditional connectivity to provide seamless, intelligent, and adaptable services that respond in real time to users' needs [1].

To fully realise the potential of edge computing within 6G networks, however, several open challenges must be addressed, including the co-design of computation and communication processes, efficient resource scheduling, and dynamic service provisioning. Among these, optimal service placement is crucial, as it ensures dynamic adaptability, strengthens security, and maximises resource utilisation across distributed edge servers [2]–[5]. Service placement involves determining the ideal distribution of services to balance demand with available resources effectively and efficiently. Recent research identifies optimal service placement as a central challenge in edge computing, primarily due to its inherent complexity; typically, it is formulated as a mixed-integer programming (MIP) problem, which is computationally intensive and often intractable for classical optimisation techniques [6]. Addressing this challenge is essential for achieving the resource management and operational flexibility required to support the demanding, real-time applications that define the future of 6G networks.

With the rapid advancement of quantum computing technology, quantum-centric optimisation has emerged as a promising approach for addressing challenging combinatorial optimisation problems, particularly in the realm of MIP [7]. The advent of quantum computing presents substantial opportunities to address challenging combinatorial optimisation problems that are intractable for classical methods. The quantum approximate optimisation algorithm (QAOA), first introduced in [8], has attracted significant attention due to its potential to tackle large-scale, classically intractable problems. Studies, such as those in [9], [10], have shown that QAOA can effectively navigate complex solution spaces, providing a feasible route for optimising problems that exceed the computational limits of classical algorithms. These findings offer strong evidence for the practical utility of quantum optimisation across various domains.

Recent work exploring quantum optimisation within edge computing, such as in [11], demonstrates that quantum-based solutions are poised to become a viable alternative to classical algorithms. As edge computing in 6G networks demands highly efficient resource allocation and service placement—challenges often framed as MIP problems—quantum optimisation presents a compelling new approach. By leveraging quantum computing's ability to process and solve complex optimisation tasks at scale, quantum-centric solutions offer a path forward for realising the full potential of edge computing in next-generation networks.

Given the growing demands of 6G wireless networks, edge computing has become essential to enable high-efficiency, low-latency, and resource-intensive applications. However, achieving the adaptability and efficiency required for these next-generation services necessitates solving complex optimisation problems, often formulated as MIP tasks that are challenging for classical computing. With the promising advancements in quantum computing, quantum-centric optimisation solutions offer a compelling approach to tackle these intractable problems. This study aims to explore the potential of quantum optimisation in advancing the dynamic adaptability of future edge computing systems, creating opportunities to enable a

D. V. Huynh and O. A. Dobre are with the Faculty of Engineering and Applied Science, Memorial University, St. John's, NL A1B 3X5, Canada (e-mails: vdhuynh@mun.ca, odobre@mun.ca).

T. Q. Duong is with the Faculty of Engineering and Applied Science, Memorial University, St. John's, NL A1C 5S7, Canada, and with the School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, BT7 1NN Belfast, U.K., (e-mail: tduong@mun.ca).

This work was supported in part by the Canada Excellence Research Chair (CERC) Program CERC-2022-00109 and in part by the Canada Research Chair Program CRC-2022-00187. Corresponding author is Trung Q. Duong

new generation of applications and transforming how 6G networks manage resources. The main contributions of this study are summarised as follows:

- We formulate the dynamic service placement problem in 6G edge computing as an optimisation task, addressing the critical requirements for adaptability and resource efficiency in next-generation networks.
- We transform this classical optimisation problem into a quadratic unconstrained binary optimisation (QUBO) format and a Hamiltonian representation, enabling compatibility with quantum processors.
- We develop and implement a quantum-centric optimisation approach to effectively solve the problem under the constraints of edge computing environments.
- We conduct extensive simulations to evaluate the potential of quantum optimisation in solving computationally intensive problems in wireless networks, demonstrating its viability as a future solution.

II. System Model and Problem Formulation

In this letter, we consider a typical edge computing system comprising M edge servers (ES) and N user equipment (UE). To execute computational tasks requested by the UEs, the ESs must install the appropriate services. Only when the proper service is installed on the ESs can the requested tasks from the UEs be processed. An illustration of the considered system model is provided in Fig. 1.

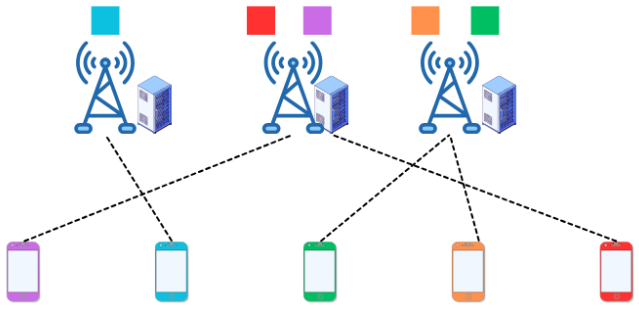

Fig. 1. An illustration of the service placement problem in edge computing.

A. Service Placement Model

In the considered system model, the computational task requested by the n-th UE is characterised by three parameters: C_n, the required CPU cycles to process the task; D_n, the delay tolerance (i.e., the maximum allowable delay) of the task; and t_n, a binary parameter indicating whether the task is requested (i.e., $t_n = 1$) or not (i.e., $t_n = 0$). In addition, tasks are associated with a cost value, ρ_n, which models the service cost. The objective of service placement is to determine which service should be installed at each particular server. An optimal service placement strategy can improve the system's dynamic adaptability, reduce service costs, and enhance the security aspect of the system.

We assume that there are a total of N services in the system. Let x_{mn} be a binary decision variable, where $x_{mn} = 1$ if the m-th server hosts the service for the n UE, and $x_{mn} = 0$ otherwise. This can be mathematically expressed as follows:

$$x_{mn} = \begin{cases} 1, & \text{if the } n\text{-th service is hosted at the } m\text{-th ES,} \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

Due to limitations in computing resources, the total computational load assigned to any server must not exceed its capacity $F_m^{\max}$:

$$\sum_{n=1}^{N} x_{mn} f_n \leq F_m^{\max}, \quad \forall m \in \{1, 2, \ldots, M\}, \tag{2}$$

where f_n denotes the computing resources required by the ES to handle a single requested task from the UEs.

To ensure that each requested task from the UEs can be processed, each requested service should be placed on at least one server. This requirement can be expressed as follows:

$$\sum_{m=1}^{M} x_{mn} \geq t_n, \quad \forall n \in \{1, 2, \ldots, N\}. \tag{3}$$

B. Delay Model

In this letter, we aim at minimising both the service cost of service placement and the total latency for processing the tasks. The delay includes wireless transmission delay and edge processing delay. The wireless channel gain h_{mn} between the n-th UE and the m-th ES is modelled as $h_{mn} = \sqrt{g_{mn}}\tilde{h}_{mn}$, where g_{mn} denotes the large-scale fading, and $\tilde{h}_{mn}$ accounts for small-scale Rayleigh fading. For each UE-ES pair, the transmission rate R_{mn} is calculated based on a dedicated bandwidth allocation, assuming no interference from other UEs. This non-interference assumption is achieved by allocating an isolated bandwidth B to each UE-ES link, ensuring that all transmissions operate independently. Consequently, the signal-to-noise ratio (SNR), denoted by γ_{mn}, for each pair is given by: $\gamma_{mn} = \frac{|h_{mn}|^2 P_{\text{tx}}}{\sigma^2}$, where P_{tx} is the transmission power and σ^2 is the noise variance. Using γ_{mn}, the transmission rate R_{mn} is computed as:

$$R_{mn} = B \log_2(1 + \gamma_{mn}), \tag{4}$$

As a result, the end-to-end delay, including transmission delay and the edge processing delay can be calculated as follows

$$D_{mn} = \frac{S_n}{R_{mn}} + \frac{C_n}{f_n}, \tag{5}$$

where S_n is the task size (bits); R_{mn} is the transmission rate (bits/s); f_n is the designed processing rate of the MEC for each task.

C. Optimisation Problem Formulation

Based on the above development, the optimisation problem addressed in this study is formulated as (6). The objective is to minimise the total cost associated with service placement and the total end-to-end delay by finding the optimal service

placement decisions.

$$\min_{\mathbf{x}} \sum_{m=1}^{M}\sum_{n=1}^{N}\rho_n x_{mn} + \sum_{m=1}^{M}\sum_{n=1}^{N} x_{mn}D_{mn}, \tag{6a}$$

$$\text{s.t.} \sum_{n=1}^{N} x_{mn} f_n \le F_m^{\max}, \forall m, \tag{6b}$$

$$\sum_{m=1}^{M} x_{mn} \ge t_n, \forall n, \tag{6c}$$

$$x_{mn} \in \{0,1\}, \quad \forall m, n. \tag{6d}$$

In (6), constraint (6b) represents the computing capacity constraint for the ESs, while constraint (6c) ensures service availability. The formulated problem is clearly a mixed-integer (binary) programming problem, which is NP-hard and intractable for classical optimisation methods.

III. Proposed Quantum-Centric Optimisation Solution

In this paper, we develop a quantum-centric optimisation approach based on the quantum approximate optimisation algorithm (QAOA) to tackle problem (6). To achieve this, we first transform the classical integer program into a quadratic unconstrained binary optimisation (QUBO) formulation. Then, from the QUBO form, we convert it into a Hamiltonian expression for execution on a quantum processor.

A. Transformation of Classical Problem to QUBO Problem

To transform the classical problem into a QUBO form, we introduce penalty parameters λ_0, λ_1, λ_2, and λ_3 to reformulate problem (6) as follows.

$$\min_{\mathbf{x}} \Big[\lambda_0 \sum_{m=1}^{M}\sum_{n=1}^{N}\rho_n x_{mn} + \lambda_1 \sum_{m=1}^{M}\left(\sum_{n=1}^{N} x_{mn} f_n - F_m^{\max}\right)^2 + \lambda_2 \sum_{n=1}^{N}\left(t_n - \sum_{m=1}^{M} x_{mn}\right)^2 + \lambda_3 \sum_{n=1}^{N}\sum_{m=1}^{M} x_{mn} D_{mn}\Big]. \tag{7}$$

In this formulation, λ_1 and λ_2 are penalty parameters that enforce the constraints on the computing capacity of the ESs, (6b) and service availability, (6c), respectively. By setting λ_1 and λ_2 to sufficiently large values, the optimisation process is guided toward feasible service placement decisions that satisfy these constraints. Additionally, the parameters λ_0 and λ_3 balance the trade-off between service cost and delay in the objective function, ensuring that both aspects are considered effectively in the optimisation process.

It is important to note that, in order to implement the quantum-centric optimisation problem, the optimisation problem must be expressed in a quadratic form, given by

$$\min_{\mathbf{x}} \quad \mathbf{x}^T \mathbf{Q} \mathbf{x}. \tag{8}$$

Therefore, we must construct the $\mathbf{Q}$ matrix based on the formulation of the problem in (7). Details of the construction process are provided in the Appendix.

B. Transformation of QUBO Problem to the Hamiltonian

In order to transform the QUBO problem into the the Hamiltonian, we first convert the binary variables to spin variables $z_{mn} \in \{-1, 1\}$. We convert each binary variable $x_{mn} \in \{0, 1\}$ to a spin variable z_{mn} using:

$$x_{mn} = \frac{1 - z_{mn}}{2}. \tag{9}$$

Substituting this expression into the QUBO objective function allows us to rewrite it in terms of spin variables.

Firstly, we substitute (9) into each part of the objective function:

$$\sum_{m=1}^{M}\sum_{n=1}^{N}\rho_n x_{mn} = \sum_{m=1}^{M}\sum_{n=1}^{N}\rho_n \frac{1 - z_{mn}}{2} = \frac{1}{2}\sum_{m=1}^{M}\sum_{n=1}^{N}\rho_n - \frac{1}{2}\sum_{m=1}^{M}\sum_{n=1}^{N}\rho_n z_{mn}. \tag{10}$$

Then, we process the constraint (6b), the second term of (7)

$$\lambda_1 \sum_{m=1}^{M}\left(\sum_{n=1}^{N} x_{mn} f_n - F_m^{\max}\right)^2 = \lambda_1 \sum_{m=1}^{M}\left(\sum_{n=1}^{N}\frac{f_n(1 - z_{mn})}{2} - F_m^{\max}\right)^2. \tag{11}$$

This will yield a combination of constants, linear terms in z_{mn}, and quadratic terms in $z_{mn}z_{mk}$. Similarly, the availability penalty term is transformed as follows

$$\lambda_2 \sum_{n=1}^{N}\left(t_n - \sum_{m=1}^{M} x_{mn}\right)^2 = \lambda_2 \sum_{n=1}^{N}\left(t_n - \sum_{m=1}^{M}\frac{1 - z_{mn}}{2}\right)^2. \tag{12}$$

Finally, the delay term in spin variables is given by

$$\lambda_3 \sum_{m=1}^{M}\sum_{n=1}^{N} x_{mn} D_n = \frac{\lambda_3}{2}\sum_{m=1}^{M}\sum_{n=1}^{N} D_n(1 - z_{mn}) = \frac{\lambda_3}{2}\sum_{m=1}^{M}\sum_{n=1}^{N} D_n - \frac{\lambda_3}{2}\sum_{m=1}^{M}\sum_{n=1}^{N} D_n z_{mn}. \tag{13}$$

After expanding all terms, we combine them into a Hamiltonian in terms of spin variables z_{mn}. Each spin variable z_{mn} corresponds to a Pauli Z operator, where $z_{mn} = Z_{mn}$ Thus, each linear term in z_{mn} translates to a single Z_{mn} operator, and each quadratic term in $z_{mn}z_{mk}$ becomes a product of two Pauli Z operators.

The Hamiltonian takes the form:

$$H_C = \sum_i I_i Z_i + \sum_{i<j} J_{ij} Z_i Z_j, \tag{14}$$

where I_i are coefficients for each linear Z_i term (from the linear terms in z_{mn}) and J_{ij} are coefficients for each quadratic $Z_i Z_j$ term.

C. Proposed Algorithm

Based on the above development, we propose a quantum-centric optimisation algorithm to solve the formulated service placement problem, as presented in Algorithm 1. The algorithm begins by taking as input the system parameters, penalty

parameters, and quantum backend settings. Next, we construct the parameterised quantum circuit (ansatz) for quantum measurement. The optimisation process then aims to optimise the parameters $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ on the ansatz to minimise the objective value. This process is repeated until convergence is reached or the maximum number of iterations is attained.

Algorithm 1 : Proposed quantum-centric optimisation algorithm for solving the service placement problem (6).

1: **Input:** Problem parameters M, N, ρ, f_n, $F_m^{\max}$, t_n, D_n; penalty coefficients λ_0, λ_1, λ_2, λ_3; quantum backend; classical optimiser.
2: Design a parameterised ansatz circuit with number of layers p, initialized with parameter vector $(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0)$.
3: Initialise a classical optimiser (e.g., COBYLA) for iterative parameter adjustment.
4: Implement a function `Obj`$(\boldsymbol{\beta}, \boldsymbol{\gamma})$ to:
- Bind parameters $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ to the ansatz circuit.
- Execute the circuit on the quantum backend to compute $\langle \boldsymbol{\beta}, \boldsymbol{\gamma} | H_C | \boldsymbol{\beta}, \boldsymbol{\gamma} \rangle$.
- Return the measured expectation value as the objective for optimisation.

5: *Optimise Ansatz Parameters*
6: **repeat**
7: Update $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ using the classical optimiser on `Obj`.
8: Track the objective function values to observe convergence.
9: **until** convergence
10: **Output:** Optimised binary service placement matrix $\mathbf{x}$, minimised objective value.

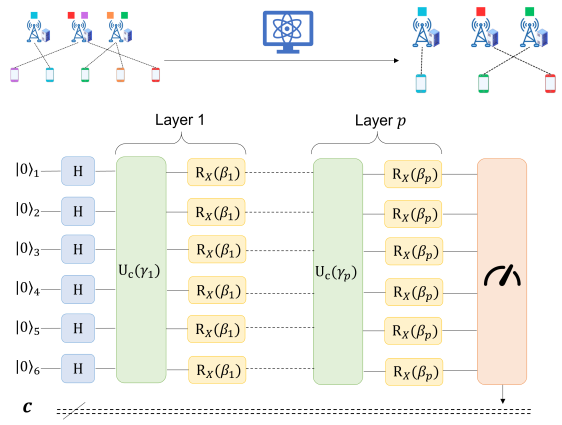

Fig. 2. A example of the parameterised quantum circuit in the quantum-centric optimisation solution for the 6-qubit service placement problem.

Regarding the design of the ansatz, Fig. 2 illustrates an example of the parameterised quantum circuit for a problem with six variables (6 qubits). The circuit consists of p layers, starting with Hadamard gates $H^{\otimes n}|0\rangle$ to create superpositions. Each layer applies two key components: $U_C(\gamma_k) = \exp(-i\gamma_k H_C)$, which encodes the cost function into the circuit, and $R_X(\beta_k) = \exp(-i\beta_k H_M)$, which explores the solution space by introducing mixing. The circuit is parameterised by angles $\gamma_1, \ldots, \gamma_p$ and $\beta_1, \ldots, \beta_p$, and the optimisation process adjusts these parameters to minimise the objective function.

IV. Simulation Results and Discussions

A. Parameter Setting

For simulations, we consider a system model with $M = 3$ ESs and $N = \{2, 3, \ldots, 7\}$ UEs. The transmission power of the UEs is set to 20 dBm, and the allocated bandwidth for each UE-ES link is 1 MHz. The large-scale fading for the wireless transmission between the n-th UE and the m-th ES is modeled as $g_{mn} = 10^{\mathbf{PL}(d_{mn})/10}$, where $\mathbf{PL}(d_{mn}) = -35.3 - 37.6 \log_{10} d_{mn}$ and d_{mn} is the distance between the n-th UE and the m-th ES [6]. The processing rate of each ES for handling tasks is set to 2 GHz, while the maximum computing capacity of each server is 3 GHz. The simulations are conducted in a Python environment, utilising packages such as `qiskit`, `qiskit_algorithms`, and `matplotlib` to implement the proposed algorithm and visualise the numerical results. For real quantum hardware processing, we run the code on the `ibm_quebec` backend.

B. Numerical Results

1) Convergence behaviour of the proposed algorithm: Fig. 3 illustrates the convergence behaviour of the proposed algorithm in a scenario with 9 optimisation variables, evaluated on both a simulator and real IBM quantum hardware (`ibm_quebec`). The plot presents the objective function value over 60 iterations, comparing the performance between the Qiskit Aer simulator and `ibm_quebec` backend. Initially, both platforms exhibit significant fluctuations in the objective function values, reflecting the exploration phase of the optimisation process. After approximately 30 iterations, both the simulator and real quantum hardware begin to converge towards minimised objective function values. Notably, the final objective value obtained on the simulator is slightly lower than that on the real IBM hardware, indicating better optimisation performance in the simulated environment. This difference can be attributed to the noise and decoherence effects inherent in quantum hardware, which can slightly degrade the optimisation accuracy. Despite these hardware limitations, the overall convergence trend on `ibm_quebec` closely aligns with the simulator, demonstrating the algorithm's robustness and effectiveness in both simulated and real quantum environments.

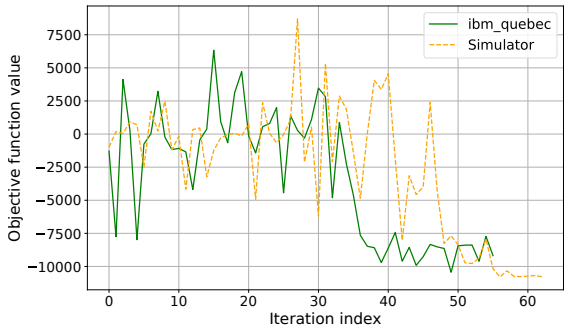

Fig. 3. Convergence behaviour of the proposed algorithm with the scenario of 9 optimisation variables on Qiskit Aer simulator and IBM quantum computer (PINQ^2).

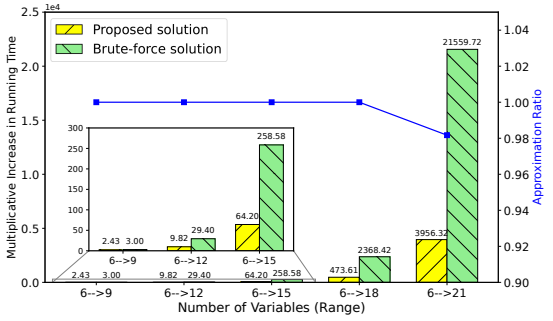

Fig. 4. The multiplicative increase of running time and the approximation ratio of the proposed solution compared with the classical solution.

2) Running time comparisons and approximation ratio analysis: Fig. 4 illustrates the multiplicative increase in running time for both the proposed quantum-centric method and the brute-force method as the number of variables increases from 6 to 21. The x-axis denotes the range of variables in steps, while the y-axis (left) represents how many times the running time increases when moving from 6 variables to the indicated number of variables. The figure clearly demonstrates the rapid increase in the brute-force method's running time, which rises significantly as the number of variables increases, peaking at approximately 21,560 times for the increase of 15 variables from 6 to 21 scenario. In contrast, the running time for quantum-based method increases more moderately, reaching 3956 times for the same case. A zoomed-in inset highlights the results for smaller variable ranges (6→9, 6→12, and 6→15), where the differences between the two methods are less pronounced. In addition, the right y-axis plots the approximation ratio (brute-force solution over quantum-centric solution), which remains at 1 for most cases, except for a slight dip (0.98) in the last case. This indicates that QAOA provides near-optimal solutions for most instances. Importantly, the results reveal that the quantum-centric solution has the potential to provide near-optimal solutions for problems that are classically intractable.

V. Conclusion

In this letter, we have explored the potential of quantum-centric optimisation for addressing the service placement problem in 6G edge computing. The formulated optimisation problem minimises both service placement costs and total delay while meeting the computing capacity constraints of edge servers and ensuring service availability. Numerical results demonstrate the effectiveness of our approach, showing clear convergence behaviour and a multiplicative increase in runtime efficiency compared to classical solutions as the system model scales. In future work, we will extend our quantum-centric optimisation approach to dynamic 6G networks, compare it with classical algorithms, and evaluate its scalability with advancing quantum hardware.

Appendix: Construction of the **Q** Matrix for the QUBO Formulation

The Q matrix for the QUBO problem represents the objective function for optimising service placement in an edge computing network. Given binary variables x_{mn}, where M is the number of edge servers (ESs) and N is the number of user equipment (UEs), the matrix entries are defined as follows:

The **diagonal entries** $Q_{(m,n),(m,n)}$ represent linear terms associated with individual placement decisions and are calculated as:

$$Q_{(m,n),(m,n)} = \lambda_0 \rho_n - 2\lambda_1 F_m^{\max} f_n - 2\lambda_2 t_n + \lambda_3 D_{mn} \quad (15)$$

where $\lambda_0 \rho_n$ applies a service cost penalty with ρ_n as the cost for the n-th UE, $-2\lambda_1 F_m^{\max} f_n$ enforces a penalty related to the ES's maximum capacity $F_m^{\max}$ when serving the n-th UE with computing requirement f_n, and $-2\lambda_2 t_n$ applies an availability penalty when a service request t_n is active for the n-th UE. The delay penalty term $\lambda_3 D_{mn}$ penalizes the delay between the n-th UE and ES m, where D_{mn} represents the delay cost.

The **off-diagonal entries** $Q_{(m,n),(p,j)}$ represent quadratic interactions between placement decisions and are given by:

$$Q_{(m,n),(p,j)} = \begin{cases} \lambda_1 f_n f_j, & \text{if } m = p \text{ and } n \neq j, \\ \lambda_2, & \text{if } m \neq p \text{ and } n = j, \\ 0, & \text{otherwise.} \end{cases} \quad (16)$$

where $\lambda_1 f_n f_j$ is the capacity constraint penalty applied when different UEs n and j are served by the same ES m (thus controlling for total computational load). The term λ_2 is an availability constraint penalty applied when the same the n-th UE is placed on different ESs m and p, ensuring that each requested service is only placed once. All other off-diagonal terms are set to zero, as they represent independent decisions without interaction.

References

[1] K. B. Letaief, Y. Shi, J. Lu, and J. Lu, "Edge artificial intelligence for 6G: Vision, enabling technologies, and applications," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 1, pp. 5–36, Jan. 2022.

[2] J. Li, F. Lin, L. Yang, and D. Huang, "AI service placement for multi-access edge intelligence systems in 6G," *IEEE Trans. Netw. Sci. Eng.*, vol. 10, no. 3, pp. 1405–1416, May/Jun. 2023.

[3] L. Gu, W. Zhang, Z. Wang, D. Zeng, and H. Jin, "Service management and energy scheduling toward low-carbon edge computing," *IEEE Trans. on Sustainable Comput.*, vol. 8, no. 1, pp. 109–119, Jan.-Mar. 2023.

[4] X. Li, X. Zhang, and T. Huang, "Joint task offloading and service placement for mobile edge computing: An online two-timescale approach," *IEEE Trans. Cloud Comput.*, vol. 11, no. 4, pp. 3656–3671, Oct.-Dec. 2023.

[5] L. Chen, C. Shen, P. Zhou, and J. Xu, "Collaborative service placement for edge computing in dense small cell networks," *IEEE Trans. Mobile Comput.*, vol. 20, no. 2, pp. 377–390, Feb. 2021.

[6] D. V. Huynh, V.-D. Nguyen, O. A. Dobre, S. R. Khosravirad, and T. Q. Duong, "Adaptive service placement, task offloading and bandwidth allocation in task-oriented URLLC edge networks," in *Proc. 2023 IEEE Int. Conf. Commun. (ICC'23)*, Rome, Italy, 2023, pp. 5755–5760.

[7] R. Mandelbaum, A. D. Córcoles, and J. Gambetta, "IBM's big bet on the quantum-centric supercomputer: Recent advances point the way to useful classical-quantum hybrids," *IEEE Spectrum*, vol. 61, no. 9, pp. 24–33, Sep. 2024.

[8] E. Farhi, J. Goldstone, and S. Gutmann, "A quantum approximate optimization algorithm," *arXiv preprint arXiv:1411.4028*, 2014.

[9] R. Shaydulin *et al.*, "Evidence of scaling advantage for the quantum approximate optimization algorithm on a classically intractable problem," *Sci. Adv.*, vol. 10, no. 22, pp. 1–10, May 2024.

[10] D. J. Egger, J. Mareček, and S. Woerner, "Warm-starting quantum optimization," *Quantum*, vol. 5, p. 479, Jun. 2021.

[11] C. Mastroianni, F. Plastina, J. Settino, and A. Vinci, "Variational quantum algorithms for the allocation of resources in a cloud/edge architecture," *IEEE Trans. Quantum Eng.*, vol. 5, pp. 1–18, May 2024.