Consiglio Nazionale delle Ricerche
Istituto di Calcolo e Reti ad Alte Prestazioni

Quantum Computing Management of a Cloud/Edge Architecture

C. Mastroianni, L. Scarcello, A. Vinci, S. Celia

RT-ICAR-CS-22-9 **Dicembre 2022**

Consiglio Nazionale delle Ricerche, Istituto di Calcolo e Reti ad Alte Prestazioni (ICAR)
– Sede di Cosenza, Via P. Bucci 8-9C, 87036 Rende, Italy, URL: www.icar.cnr.it
– Sezione di Napoli, Via P. Castellino 111, 80131 Napoli, URL: www.icar.cnr.it
– Sezione di Palermo, Via Ugo La Malfa, 153, 90146 Palermo, URL: www.icar.cnr.it

INDEX

Quantum Computing Management of a Cloud/Edge Architecture

CARLO MASTROIANNI, ICAR-CNR, Italy
LUIGI SCARCELLO, ICAR-CNR, Italy
ANDREA VINCI, ICAR-CNR, Italy
SABRINA CELIA, ICAR-CNR, Italy

Modern Cloud/Edge architectures are composed of computing nodes belonging to multiple layers, including Cloud facilities, Edge/Fog nodes and sensors/actuators. In this paper, we present an architecture that includes also quantum computing devices, in two ways: (i) quantum devices can become, in the next future, a viable alternative for executing computation that is intractable classically and (ii) they can be exploited to assist resource management and scheduling within the architecture itself. Furthermore, we describe the procedure through which a typical resource assignment problem, which has NP-hard complexity, can be transformed into a formulation that can be tackled by QAOA, a renowned hybrid quantum algorithm, and we present some preliminary results obtained for a simple instance, a knapsack problem where an Edge node needs to select and retrieve a set of processes from the Cloud.

1 INTRODUCTION

Recently, new paradigms are emerging to distribute the computation among centralized and decentralized nodes, exploiting heterogeneous types of platforms, with very different capabilities and characteristics: sensors/actuators, mobile devices, personal computers, Cloud data centers, etc. The so-called continuous computing architecture [3, 5, 15] aims to integrate all these platforms and foster their cooperation.

Research on Quantum Computing is exploring a further alternative, which, in the next years, can help to tackle combinatorial optimization problems that are intractable on classical computers. Quantum computers leverage the intrinsic parallel nature of quantum operations, based on the superposition principle of quantum mechanics: in particular, an operation can be performed on a number of configurations that is exponential with respect to the number of adopted resources, i.e., the qubits. Today, the most promising approach is to devise variational hybrid algorithms [7], in which quantum computation is driven by a set of parameters that are optimized classically, in a cycle that aims at finding the best solution with a significant speed-up with respect to classical approaches.

A computing architecture that includes the quantum computing alternative can be composed of three layers: (i) a *device layer*, which includes sensors/actuators and smart objects; (ii) an *Edge/Fog layer*, which includes Edge nodes located close to the devices and (iii) a *Cloud/Quantum layer* that includes Cloud infrastructures and quantum computing devices. The efficient and flexible management of this architecture relies on resource management algorithms that are able to determine where the computation should be performed (assignment problem) and when (scheduling problem). Resource management can be handled at any of the three layers: for example, sensor devices can self-organize in wireless sensor networks [6], Edge nodes can determine which processes can be offloaded to the device layer or to the Cloud [2], and management algorithms can be used to schedule hundreds or thousands of processes on the nodes of a Cloud data center [1, 7].

In this paper, we discuss how a typical resource management problem can be transformed and solved by a renowned hybrid quantum algorithm, i.e., the Quantum Approximate Optimization Algorithm (QAOA) [10]. In particular, we refer to the problem of assigning a number of processes to a set of Edge nodes or to the Cloud. This is a very important problem, as Edge nodes can offer advantages under multiple aspects – lower latency, more efficient usage of local data, higher level of security, etc. – but they have limited computing and memory capacity, so they can host only a subset of processes.

Afterward, given the current limitation of available quantum computers, we discuss a simplified problem in which there is only one Edge node that can retrieve a subset of processes from the Cloud [16]: the aim is to maximize the added value related to the selection while matching the constraint related to the node capacity. This becomes a knapsack problem, which, despite its simple formulation, is a reference problem for computer scientists [11]. Indeed, it is NP-hard and no known algorithm is both correct and polynomial-time in all cases. We show how this problem can be transformed into an Ising problem [8] and solved by the QAOA algorithm. In particular, a quantum diagonal operator (a "Hamiltonian" operator) is built from the Ising expression, and the problem becomes finding the minimum eigenvalue of this operator and the corresponding eigenvector, i.e., the "ground state" in the Physics language. The components of this eigenvector correspond to the values of the binary variables that specify whether the single processes should be assigned to the Edge node or remain in the Cloud.

The intent of this paper is to introduce the researchers in the Edge/Cloud computing field to the opportunities of Quantum Computing, and report some preliminary results for a simple problem. The paper is organized as follows: Section 2 illustrates the reference architecture and the resource management problem; Section 3 describes how a knapsack problem can be reformulated and solved with QAOA; Section 4 reports preliminary results for an instance of the knapsack problem and Section 5 concludes the paper.

2 EDGE MODELING

2.1 Reference Architecture

Figure 1 shows the quantum-assisted Edge/Cloud computing continuum architecture exploited as a reference in the present paper. As mentioned in the introductory section, the architecture consists of three layers: the Device layer, the Edge/Fog layer, and the Cloud/Quantum layer. At the bottom, the Device layer comprehends a set of heterogenous end-devices, which span from connected sensors/actuators to smart objects and users' smartphones. A device can be connected to a specific Edge node or can access the internet directly. Devices can gather sensed information, act upon an environment, do local processing, and request processing to the Edge/Cloud continuum.

The Edge/Fog layer consists of a network of heterogeneous computing nodes available at different levels of the networking infrastructure that connects the device layer and the internet. These nodes can be deployed near the devices, e.g., in-home Edge computing nodes, or made available by network providers, e.g., in the nearing of cellular radio towers. The computational capabilities and resources vary widely, but, in general, the nearer to the end devices are the nodes, the lower their capabilities. However, processing on Edge nodes results in lowering the network latency in communication with the end devices.

The top layer includes and integrates Cloud and quantum resources. They are positioned in the same level as they need to be tightly integrated with each other, since hybrid algorithms combine classical and quantum computation, as mentioned in the introductory section. Adding quantum resources to the typical Cloud/Edge architecture can help to solve hard problems for which quantum algorithms promise to offer a computational speedup in the coming years, specifically: (i) optimization problems, which can be approached with variational

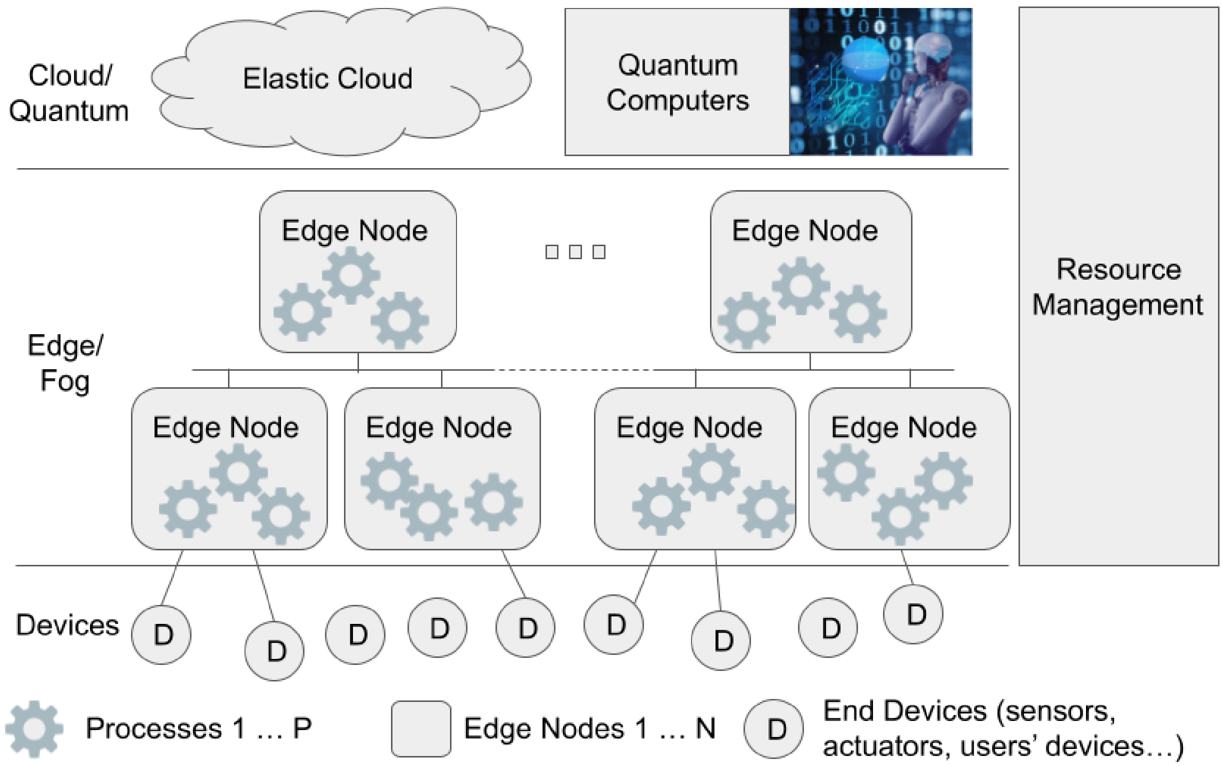

Fig. 1. Quantum-assisted Edge/Cloud computing continuum.

quantum algorithms [7], and machine learning problems, which can be tackled with quantum machine learning algorithms [4]. In this context, quantum computers can be exploited not only to perform computation but also to execute algorithms for the resource management of the architecture itself. Assignment and scheduling algorithms are typically NP-hard, therefore, if their instances are dimensionally large, classical computation could be infeasible. Quantum algorithms can be useful as the exploitation of quantum parallelism promises to speed up the computation.

2.2 Assignment Problem Formulation

Given a set of processes $\mathcal{P} = \{1, ..., P\}$ and a set of computational Edge nodes in the Edge/Cloud continuum $\mathcal{N} = \{1, ..., N\}$, there is the problem of assigning each process to an Edge node or to the Cloud. The execution of processes on Edge nodes can bring important benefits, for example, in terms of service latency and access to local data that have peculiar security and privacy characteristics. Therefore, the assignment needs to maximize the effectiveness of the processes, for example in terms of service latency, and at the same time it must take into account the limited availability of the computation resources on each node.

The problem can be formulated as a binary linear programming problem, which is known to be NP-hard [9], as follows:

$$\max \sum_{i \in \mathcal{P}, j \in \mathcal{N}} v_{ij} x_{ij} \tag{1}$$

$$\sum_{j \in \mathcal{N}} x_{ij} <= 1, \qquad \forall i \in \mathcal{P} \tag{2}$$

$$\sum_{i \in \mathcal{P}} w_i x_{ij} <= B_j, \qquad \forall j \in \mathcal{N} \tag{3}$$

The binary variable x_{ij} is equal to 1 if the process i is assigned to the Edge node j, and 0 otherwise. If a process i is not assigned to any Edge node, it is considered to be assigned to the Cloud. The objective is to assign the processes to nodes and maximize the overall gain function while matching the capacity constraints of the Edge nodes. Each process $i \in \mathcal{P}$ is assigned a value v_{ij}, which is the value gain of executing process i on a node $j \in \mathcal{N}$, w.r.t. executing the process on the Cloud (e.g., a measure of its effectiveness in terms of latency or occupied network bandwidth), and an integer weight w_i, which represents the amount of computing resources required by process i for its execution. Each Edge node j has a capacity B_j, defined as an integer. It is worth noting that the Cloud has no constraints, as its capacity is considered much larger than the available Edge nodes.

The number of binary variables needed for this problem is equal to:

$$P \cdot N + P + \sum_{j \in \mathcal{N}} \lceil \log_2(B_j + 1) \rceil \tag{4}$$

where $P \cdot N$ is the number of binary variables x_{ij}. Moreover, to convert the inequalities constraints (2) into equations, we need P slack binary variables, one for each inequality. Also, to convert the inequalities constraints (3), for each node j, we need a number of slack binary variables $\lceil \log_2(B_j + 1) \rceil$ that allow giving every possible value to the residual capacity of the node j, i.e., the capacity that remains not assigned to processes, defined as:

$$B_j - \sum_{i \in \mathcal{P}} w_i x_{ij} \tag{5}$$

A simplified model is obtained when there is only one Edge node and a set of processes, and the problem is to decide whether or not to assign each process to the node or run it into the Cloud. Even in this simplified version, the problem is still NP-hard, since it becomes a knapsack problem:

$$\max \sum_{i \in \mathcal{P}} v_i x_i \tag{6}$$

$$\sum_{i \in \mathcal{P}} w_i x_i <= B \tag{7}$$

In this case, we have a set of binary variables x_i, which are equal to 1 if the process i is offloaded to the Edge node, and 0 otherwise, and the capacity of the Edge node is equal to B. The number of binary variables needed by the knapsack problem is:

$$P + \lceil \log_2(B + 1) \rceil \tag{8}$$

3 QUANTUM FORMULATION OF THE EDGE KNAPSACK PROBLEM

The optimization problem shown in the previous section can be tackled with quantum computing, following two possible avenues: (i) using a gate-based quantum computer and a hybrid variational algorithm, such as QAOA; (ii) using a quantum annealing computer that, through analog physical processes, aims to achieve the ground state of a Hamiltonian operator [12]. Both avenues present pros and cons. Specifically, quantum annealing can exploit a larger number of qubits, currently up to 1,000, but the connectivity graph among qubits is limited, which hinders the possibility of solving general problems. In this paper, we focus on the gate-based approach, and on its adoption for solving the knapsack problem, in which there is only one Edge node, and the problem is to decide which processes should be offloaded from the Cloud to this node.

The first step of this transformation is to convert the inequality constraint (7) into an equation that includes $\lceil \log_2(B+1) \rceil$ binary slack variables, denoted with the symbol y:

$$\sum_{i=1}^{P} w_i x_i + \sum_{i=1}^{\lceil \log_2(B+1) \rceil} 2^{(i-1)} y_i = B \tag{9}$$

The next step is to include the constraint (7) into the objective function, and obtain an extended objective function, defined as:

$$\min\Bigg(- \sum_{i=1}^{P} v_i x_i + A \cdot \Big(B - \sum_{i=1}^{P} w_i x_i - \sum_{i=1}^{\lceil \log_2(B+1) \rceil} 2^{(i-1)} y_i \Big)^2 \Bigg) \tag{10}$$

$$A = 1 + \sum_{i=1}^{P} v_i \tag{11}$$

The maximization problem has been converted to a minimization problem, and the constraint has been transformed into a penalty, which is equal to 0 only when the constraint is satisfied by the values of the binary variables. When the constraint is not satisfied, the value of the penalty is equal or larger than A. Since the value of A is defined, in Eq. (11), to be larger than the maximum possible value of the first term of Eq. (10), a violation of the constraint cannot be compensated by the minimization of the first term. This ensures that, in the desired solution, the values of the binary variables match the constraint.

Now, we need to perform the following variable substitutions:

$$x_i = \frac{1 - z_i}{2}, \ i = 1, ..., P \tag{12}$$

$$y_i = \frac{1 - z_{(P+i)}}{2}, \ i = 1, ..., \lceil \log_2(B+1) \rceil \tag{13}$$

At this point, the binary variables have been substituted by discrete variables z, numbered from 1 to $P + \lceil \log_2 (B+1) \rceil$, which can assume values $\{+1, -1\}$, leading to an Ising problem [8], i.e., a problem formulated as:

$$min \left(\sum_{i=1}^{Q} h_i \cdot z_i - \sum_{i=1}^{Q} \sum_{j=1}^{i-1} J_{ij} \cdot z_i \cdot z_j \right) \tag{14}$$

where the number of discrete variables is denoted as $Q = P + \lceil \log_2(B+1) \rceil$.

Now, each discrete variable is associated with a qubit, and the quantum computing problem is defined over a registry of Q qubits. Starting from the Ising expression, a Hamiltonian operator is built with sums and tensor products of two basic one-qubit operators, the identity $\mathbf{I}$ and the Pauli operator $\mathbf{Z}$. For each term in Eq. 14, the operator $\mathbf{Z}_i$ substitutes the variable z_i, and the identity operator $\mathbf{I}$ is added for each variable z that does not appear in the term. Moreover, the multiplications between the z variables are substituted with the tensor products between the corresponding $\mathbf{Z}$ operators. For example, with $Q = 5$ the term $z_1 \cdot z_4$ is substituted with $\mathbf{Z}_1 \otimes \mathbf{I}_2 \otimes \mathbf{I}_3 \otimes \mathbf{Z}_4 \otimes \mathbf{I}_5$ or, more succinctly, $\mathbf{Z}_1 \otimes \mathbf{Z}_4$ or, even more briefly, $\mathbf{Z}_1\mathbf{Z}_4$, where the identity operators are implicit. The Hamiltonian operator that corresponds to Eq. 14 is:

$$\mathbf{H} = \sum_{i=1}^{Q} h_i \cdot \mathbf{Z}_i - \sum_{i=1}^{Q} \sum_{j=1}^{i-1} J_{ij} \cdot \mathbf{Z}_i \otimes \mathbf{Z}_j \tag{15}$$

Now, the problem becomes finding the ground state of the Hamiltonian operator, as mentioned in the introductory section. A register of Q qubits is prepared by QAOA to achieve with maximum probability the ground state, in which each qubit, after measurement, collapses to one of the basis states $|0\rangle$ or $|1\rangle$. The solution to the problem is obtained by setting each binary variable to 0 (if the corresponding qubit collapses to $|0\rangle$) or 1 (if the corresponding qubit collapses to $|1\rangle$).

4 RESULTS

Here, we show some preliminary results obtained for a knapsack problem with $P = 3$, and $B = 3$. The number of qubits is $Q = 5$, see expression (8). The values and the weights are assigned as follows:

- $(v_1, v_2, v_3) = (4, 5, 1)$
- $(w_1, w_2, w_3) = (1, 3, 1)$

and thus the value of the penalty is $A = 11$.

The problem is expressed as:

$$\max\ (4x_1 + 5x_2 + x_3) \tag{16}$$

$$x_1 + 3x_2 + x_3 <= 3 \tag{17}$$

After the definition of slack variables, and the inclusion of the constraint into the objective function, we obtain the following formulation, corresponding to the general expression (10):

$$\begin{aligned}\min\ \big(&-70 \cdot x_1 - 203 \cdot x_2 - 67 \cdot x_3 - 66 \cdot y_1 - 132 \cdot y_2 + 66 \cdot x_1 \cdot x_2 \\ &+ 22 \cdot x_1 \cdot x_3 + 22 \cdot x_1 \cdot y_1 + 44 \cdot x_1 \cdot y_2 + 66 \cdot x_2 \cdot x_3 + 66 \cdot x_2 \cdot y_1 \\ &+ 132 \cdot x_2 \cdot y_2 + 22 \cdot x_3 \cdot y_1 + 44 \cdot x_3 \cdot y_2 + 44 \cdot y_1 \cdot y_2 + 11 \cdot x_1^2 + 99 \cdot x_2^2 \\ &+ 11 \cdot x_3^2 + 11 \cdot y_1^2 + 44 \cdot y_2^2 + 99\big)\end{aligned} \tag{18}$$

After the variable substitutions (12,13), we obtain the Ising problem and the Hamiltonian operator, expressed as:

$$\begin{aligned}\mathbf{H} = &-9 \cdot Z_5 - 30.5 \cdot Z_4 - 10.5 \cdot Z_3 - 11 \cdot Z_2 - 22 \cdot Z_1 + 16.5 \cdot Z_4 Z_5 \\ &+ 5.5 \cdot Z_3 Z_5 + 16.5 \cdot Z_3 Z_4 + 5.5 \cdot Z_2 Z_5 + 16.5 \cdot Z_2 Z_4 + 5.5 \cdot Z_2 Z_3 \\ &+ 11 \cdot Z_1 Z_5 + 33 \cdot Z_1 Z_4 + 11 \cdot Z_1 Z_3 + 11 \cdot Z_1 Z_2 + 50\end{aligned} \tag{19}$$

Figure 2 shows the probability that the final measurement on the QAOA circuit gives, respectively, the best solution and an admissible solution, i.e., a solution that satisfies the constraint (17). The results have been obtained with the simulator *ibmq_qasm_simulator* and the real hardware *ibmq_montreal* of the IBM Quantum Platform. When using the simulator, the probabilities increase with the depth of the QAOA circuit, i.e., with the number of repetitions of the gates, as foreseen by [10]. Conversely, when using the real hardware, more then 20 repetitions increase the effect of noise, which explains the lower probabilities to find the best solution.

5 CONCLUSIONS AND FUTURE WORK

The paper highlighted how Quantum Computing could be exploited in the Edge/Cloud continuum. In particular, it first presented a Quantum-assisted Edge/Cloud computing architecture, where quantum computing resources can support both infrastructural behaviors and developed applications. Then, the paper proposed an approach, leveraging Quantum Computing, for solving the process assignment problem in the Edge/Cloud continuum, i.e., choosing where (Edge or Cloud) to assign a set of processes. The problem has been transformed and solved with the Quantum Approximate Optimization Algorithm. A preliminary set of results show the feasibility of the approach in the domain of the Edge/Cloud continuum.

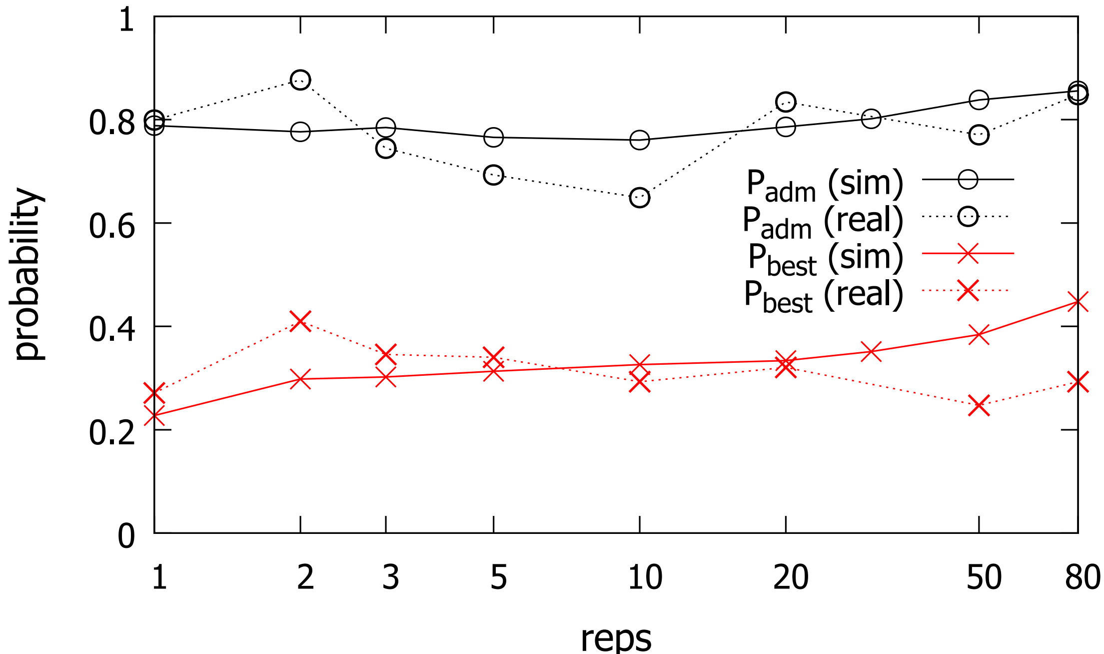

Fig. 2. Probability of finding an admissible solution, P_{adm}, and the best solution, P_{best}, with simulation and real quantum hardware.

Current research is devoted to experiments with real devices, which are prone to noise and decoherence errors. An interesting preliminary outcome is that, when increasing the size of the problem, the success probability decreases due to the noise, but the execution time hardly depends on the number of qubits. The latter result is consistent with the $O(n)$ circuit depth of the QAOA algorithm, with n the number of qubits, since the QAOA gates are executed in parallel on the qubits. Moreover, we are currently working on the formalization and the analysis of the results obtained in the more general assignment problem, presented in Section 2.2, with multiple Edge nodes.

6 ACKNOWLEDGMENTS

This work was partially funded by the Italian MUR Ministry under the project PNRR National Centre on HPC, Big Data and Quantum Computing, PUN: B93C22000620006.

REFERENCES

[1] Albino Altomare, Eugenio Cesario, and Andrea Vinci. 2019. Data analytics for energy-efficient clouds: design, implementation and evaluation. *International Journal of Parallel, Emergent and Distributed Systems* 34, 6 (2019), 690–705.

[2] Marica Amadeo, Antonella Molinaro, Stefano Yuri Paratore, Albino Altomare, Andrea Giordano, and Carlo Mastroianni. 2017. A Cloud of Things framework for smart home services based on Information Centric Networking. In *2017 IEEE 14th International Conference on Networking, Sensing and Control (ICNSC)*. 245–250.

[3] Luciano Baresi, Danilo Filgueira Mendonça, Martin Garriga, Sam Guinea, and Giovanni Quattrocchi. 2019. A unified model for mobile-edge-cloud continuum. ACM Transactions on Internet Technology (TOIT) 19, 2 (2019), 1–21.
[4] Jacob Biamonte, Peter Wittek, Nicola Pancotti, Patrick Rebentrost, Nathan Wiebe, and Seth Lloyd. 2017. Quantum machine learning. Nature 549, 7671 (2017).
[5] Ricardo Bianchini, Samee U. Khan, and Carlo Mastroianni. 2016. Guest Editors' Introduction: Special Issue on Green and Energy-Efficient Cloud Computing: Part I. IEEE Transactions on Cloud Computing 4, 2 (2016), 119–121.
[6] Miriam Carlos-Mancilla, Ernesto López-Mellado, and Mario Siller. 2016. Wireless sensor networks formation: approaches and techniques. Journal of Sensors (2016).
[7] Marco Cerezo, Andrew Arrasmith, Ryan Babbush, Simon C Benjamin, Suguru Endo, Keisuke Fujii, Jarrod R McClean, Kosuke Mitarai, Xiao Yuan, Lukasz Cincio, et al. 2021. Variational quantum algorithms. Nature Reviews Physics 3, 9 (2021).
[8] Barry A Cipra. 2000. The Ising model is NP-complete. SIAM News 33, 6 (2000).
[9] Michele Conforti, Gérard Cornuéjols, Giacomo Zambelli, et al. 2014. Integer programming. Vol. 271. Springer.
[10] Edward Farhi, Jeffrey Goldstone, and Sam Gutmann. 2014. A Quantum Approximate Optimization Algorithm. https://doi.org/10.48550/ARXIV.1411.4028 arXiv:1907.11513
[11] Harisankar Haridas, Sriram Kailasam, and Janakiram Dharanipragada. 2014. Cloudy knapsack problems: An optimization model for distributed cloud-assisted systems. In 14-th IEEE International Conference on Peer-to-Peer Computing. 1–5.
[12] Philipp Hauke, Helmut G Katzgraber, Wolfgang Lechner, Hidetoshi Nishimori, and William D Oliver. 2020. Perspectives of quantum annealing: methods and implementations. Reports on Progress in Physics 83, 5 (May 2020).
[13] Harald Mueller, Spyridon V. Gogouvitis, Houssam Haitof, Andreas Seitz, and Bernd Bruegge. 2016. Poster Abstract: Continuous Computing from Cloud to Edge. In 2016 IEEE/ACM Symposium on Edge Computing (SEC). 97–98.
[14] Junhua Wang, Tingting Liu, Kai Liu, BaekGyu Kim, Jiang Xie, and Zhu Han. 2018. Computation Offloading Over Fog and Cloud Using Multi-Dimensional Multiple Knapsack Problem. In 2018 IEEE Global Communications Conference (GLOBECOM). 1–7.