A Study of Edge Computing Enabled IoT Systems

Thesis submitted in partial fulfillment
for the Award of Degree

Doctor of Philosophy

by

Sumit Kumar

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY
(BANARAS HINDU UNIVERSITY)
VARANASI - 221005

Roll No. 17071007 *Year 2022*

Chapter 7

Conclusion and Future Directions

This chapter summarizes the important conclusions obtained from the contributions in this thesis. Additionally, it provides promising future directions to explore the edge computing and networking resource allocation problems.

7.1 Conclusion

In this thesis, we studied the essential problem of allocating edge computing and networking resources to the app users of IoT devices. We propose game-theoretical approaches to solve this problem from the perspective of different actors, such as app users, app vendors, and edge infrastructure providers. The following are the primary objectives achieved in this thesis. We 1) improved the app users' QoS while minimizing costs, 2) maximized the benefits of the app vendors, 3) maximized resource utilization and balanced the load on edge servers, and 4) maximized the benefits of the edge infrastructure providers.

In Chapter 2, we reviewed the existing literature on edge computing and networking resource allocation to find the research gap and limitations of the existing work. Broadly, resource allocation approaches are classified into three groups: 1) QoS-driven resource allocation, 2) Cost-effective resource allocation, and 3) Optimization-based re-

source allocation. In the literature, the allocation of edge computing resources for task offloading and edge networking resources for group communication are NP-Complete problems. Hence, most approaches take considerable time to get the desirable solution because of the large multidimensional search space. To the best of our knowledge, only few researchers have studied the resource allocation problem from the perspective of app vendors, and no study has investigated the trade-off between the app vendors' benefits and QoS for their app users. A few studies examined how to improve the resource utilization of multi-tenant edge servers.

In chapter 3, we investigated the resource allocation problem from the perspective of app users and proposed a protocol to establish the trade-off between the usage cost and the Quality of Service (QoS) while balancing the load on edge servers. In this study, we formulated the App User Allocation (AUA) problem as a UAGame, a potential game that admits at least one Pure Nash Equilibrium (PNE). An App User Allocation (AUA) algorithm that runs in parallel on the edge servers is designed to find the PNE and converge quickly. The time complexity of AUA algorithm has a bound $O(\max_{n_x}(M \times (\frac{n_x}{2})^2))$ for convergence to PNE. We also theoretically analyzed the solution's optimality in terms of Price of Stability (PoS) and found the bound for PoS that is at most $\alpha \cdot H(n)$. The simulation also depicted that the AUA algorithm minimizes the overall cost and improves the QoS compared to other state-of-the-art approaches.

In chapter 4, we studied the resource allocation problem from the app vendors' perspective, where they compete for the resources on the multi-tenant edge servers. This problem is referred to as the Edge Resource Allocation (ERA) problem. We proposed an ERAGame, a game-theoretic approach that formulates the ERA problem as a potential game. The goal of the proposed game is to maximize resource utilization and increase app users assigned to Edge servers while minimizing costs incurred by app vendors. To accomplish this objective, the ERAGame employs the ERA algorithm

for convergence to the Pure Nash Equilibrium (PNE) solution. This way, the ERA problem can be solved in a distributed manner. The ERA algorithm takes at most $\max_{q_{i,x}}(T \times q_{i,x})$ iterations to reach the PNE. We got bound $O(log\ n)$ for the price of stability of ERAGame under the ERA algorithm. We conducted extensive experiments and compared the ERA algorithm's performance with baseline and state-of-the-art approaches. The experimental findings validate that the ERA algorithm allocates the optimal bundle of computing resources by maximizing resource utilization and assigns the most app users to edge servers at a lower cost.

In Chapter 5, we investigated the problem from the perspective of edge infrastructure. We proposed an approach for allocating edge resources in a distributed manner that deals with the geographically dynamic user density and bottleneck resources on the edge server while increasing resource utilization. Our proposed approach formulated the Load Balancing (ERA) problem as the benefit function of edge servers' resources, which is defined by the revenue generated by the resource utilization and the resource overhead cost. We then proposed an ERA algorithm that accommodates the IoT and mobile users from overloaded to underutilized edge servers. This way, ERA algorithm improves the QoS by reducing the latency experienced by the users and the Quality of Experience (QoE) by maximizing the number of served users. Experimental results show that ERA algorithm archives minimum latency while serving maximum users compared to state-of-the-art approaches, which is a significant advancement.

In Chapter 6, we proposed a PSGame to construct a multicast tree that minimizes the overall cost of IoT devices equipped with sensors for group communication. We designed a path selection algorithm (PSA) to converge at Nash equilibrium by following PSGame rules. This algorithm used a proposed cost-sharing scheme to improve the quality of the Nash equilibrium. We proved that the proposed PSA reaches at least one Nash equilibrium. The time-bound for the convergence at the Nash equilibrium is $O(n.r_{max})$. The proposed algorithm's performance is analyzed theoretically, proving

that the price of stability for the obtained solution will always be less than $O(log(n))$. We also conducted experiments to numerically evaluate the algorithm's performance and compared the results with other game-theoretic, heuristic, and meta-heuristic algorithms. Results show that the proposed algorithm constructs the multicast tree of lesser cost with minimal overhead than the other state-of-the-art algorithms.

7.2 Future Directions

Based on the research work presented in this thesis, this study opened many new avenues for future investigation of the resource allocation problem. The following are possible future directions:

1) The effects of mobility and trajectories of app users on the resource allocation games proposed in the thesis can be investigated. Future research could look into app users' mobile participation in the proposed games, including existing app users' or app vendors' departures and new app users' or app vendors' arrivals.

2) The edge servers work as the caching server between the central cloud and the app users. It stores content closer to the app users, which enables quick retrieval by app users. Future research can study the impact of resource allocation to app users/ vendors on the performance of edge caching.

3) In some applications, app users may have a variable resource and QoS requirement over time. For example, Multimedia based application requires more bandwidth and storage type resources. On the other hand, computational-based applications require more CPU-type resources. The resource allocation games proposed in this thesis can be improved to accommodate app users' diverse capabilities and needs over time.