INTERNATIONAL JOURNAL OF TECHO-ENGINEERING

ISSN : 2057-5688

PRICE BASED RESOURCE ALLOCATION FOR EDGE COMPUTING A MARKET EQUILIBRIUM APPROACH

CHITIKENA ARUNA
PG Scholar, Department of Computer Science,
SVKP & Dr K S Raju Arts & Science College,
Penugonda, W.G.Dt., A.P, India.
chitikena.aruna@gmail.com

A.N.RAMAMANI*
Associate Professor in Computer Science,
SVKP & Dr K S Raju Arts & Science College,
Penugonda, W.G.Dt., A.P, India
manivasu6@gmail.com

Abstract:

The emerging edge computing paradigm promises to deliver superior user experience and enable a wide range of Internet of Things (IoT) applications. In this paper, we propose a new market-based framework for efficiently allocating resources of heterogeneous capacity-limited edge nodes (EN) to multiple competing services at the network edge. By properly pricing the geographically distributed ENs, the proposed framework generates a market equilibrium (ME) solution that not only maximizes the edge computing resource utilization but also allocates optimal resource bundles to the services given their budget constraints. When the utility of a service is defined as the maximum revenue that the service can achieve from its resource allotment, the equilibrium can be computed centrally by solving the Eisenberg-Gale (EG) convex program. We further show that the equilibrium allocation is Pareto-optimal and satisfies desired fairness properties including sharing incentive, proportionality, and envy-freeness. Also, two distributed algorithms, which efficiently converge to an ME, are introduced. When each service aims to maximize its net profit (i.e., revenue minus cost) instead of the revenue, we derive a novel convex optimization problem and rigorously prove that its solution is exactly an ME. Extensive numerical results are presented to validate the effectiveness of the proposed techniques.

Index Terms—Market equilibrium, Fisher market, fairness, algorithmic game theory, edge computing, fog computing.

1 INTRODUCTION

The last decade has witnessed an explosion of data traffic over the communication network attributed to the rapidly growing cloud computing and pervasive mobile devices. This trend is expected to continue for the foreseeable future with a whole new generation of applications including 4K/8K UHD video, tactile Internet, virtual/augmented reality (VR/AR), and a variety of IoT applications [1]. As the cloud infrastructure and number of devices continue to expand at an accelerated rate, a tremendous burden will be put on the network. Hence, it is imperative for operators to develop innovative solutions to meet the soaring traffic demand and accommodate diverse requirements of various services and use cases in future networks. Thanks to the economy of scale and supercomputing capability advantages, cloud computing will likely continue to play a prominent role in the future computing landscape. However, cloud data centers (DC) are often geographically distant from the end-user, which induces enormous network traffic, along with significant communication delay and jitter. Therefore, despite the immense power, cloud computing alone is facing growing limitations in satisfying the stringent requirements in terms of latency, reliability, security, mobility, and localization of new systems and applications (e.g., embedded artificial intelligence, missioncritical communication, 5G wireless systems) [1]. To this end, edge computing (EC) [2], also known as fog computing (FC) [1], has emerged as a novel computing paradigm that complements the cloud and addresses many shortcomings in the traditional cloud model. In EC, storage, computing, control, and networking resources are placed closer to end-users, things, and sensors. The size of an EN is flexible ranging from smartphones, smart access points (AP), base stations (BS) to edge clouds [3]. For example, a smartphone is the edge between wearable devices and the cloud, a home gateway is the edge between smart appliances and the cloud, a telecom central office is the edge between mobile devices and the core network. By providing elastic resources and intelligence at the edge, EC offers many remarkable capabilities, such as local data processing and analytics, distributed caching, location awareness, resource pooling and scaling, enhanced privacy and security, and reliable connectivity. EC is also a key enabler for ultra-reliable low-latency applications (e.g., AR, autonomous driving). A myriad of benefits and other use cases (e.g., offloading, caching, advertising, healthcare, smart homes/grids/cities) of EC can be found in [1]–[3]. Today, EC is still in the developing stages and presents many new challenges, such as network architecture design, programming models and abstracts, IoT support, service placement, resource provisioning and management, security and privacy, incentive design, and reliability and scalability of edge devices [1]–[3]. In this paper, we focus on the EC resource allocation problem. Unlike cloud computing, where computational capacity of large DCs is virtually unlimited and network delay is high, EC is characterized by relatively low network latency but considerable processing delay due to the limited computing power of ENs. Also, there are a massive number of distributed computing nodes compared to a small number of large DCs. Additionally, ENs may come with different sizes (e.g., number of computing units) and configurations (e.g., computing speed) ranging from a smartphone to an edge cloud with tens/hundreds of servers. These nodes are dispersed in numerous locations with varying network and service delay towards end-users. On the other hand, different services may have different requirements and properties. Some services can only be handled by ENs satisfying certain criteria. Furthermore, different services may be given different priorities. While every service not only wants to obtain as much resource as possible but also prefers to be served by its closest ENs with low response time, the capacities of ENs are limited. Also, due to the diverse preferences of the services towards the ENs, some nodes can be under-demanded while other are over-demanded. Thus, a fundamental problem is: given a set of geographically distributed heterogeneous ENs, how can we efficiently allocate their limited computing resources to competing services with different desires and characteristics, considering service priority and fairness? This work introduces a novel market-based solution framework which aims not only to maximize the resource utilization of the ENs but also to make every service happy with the allocation decision. The basic idea behind our approach is to assign different prices to resources of different ENs. In particular, highly sought-after resources are priced high while prices of under-demanded resources are low. We assume that each service has a certain budget for resource procurement. The budget can be virtual or real money. Indeed, budget is used to capture service priority/differentiation. It can also be interpreted as the market power of each service. Given the resource prices, each service buys the favorite resource bundle that it can afford. When all the resources are fully allocated, the resulting prices and allocation form a market equilibrium (ME). If there is only one EN, an ME can be found easily by adjusting the price gradually until demand equals supply or locating the intersection of the demand and supply curves. However, when there are multiple heterogeneous ENs and multiple services with diverse objectives and different buying power, the problem becomes challenging since the services have more options to buy resources. We consider two distinct market models in this work. In the first model, the money does not have intrinsic value to the services. Given resource prices, each service aims to maximize its revenue from the allocated resources, without caring about how much it has to pay as long as the total payment does not exceed its budget. This model arises in many real-world scenarios. For example, in 5G networks, the Mobile Edge Computing (MEC) servers of a Telco are shared among different network slices, each of which runs a separate service (e.g., voice, video streaming, AR/VR, connected vehicles, sensing) and serves a group of customers who pay for the service. The Telco can allot different budgets to the slices depending on their importance and/or potential revenue generation (e.g., the total fee paid by the users/subscribers of each slice). Similarly, an application provider (e.g., Uber, Pokemon Go) or a sensor network may own a number of ENs in a city and need to allocate the edge resources to handle requests of different groups of users/sensors. The budget can be decided based on criteria such as the populations of users/sensors in different areas and/or payment levels (subscription fees) of different groups of users. Another example is that a university (or other organizations) can grant different virtual budgets to different departments or research labs so that they can fairly share the edge servers on the campus. The first model may also emerge in the setting of cloud federation at the edge where several companies (i.e., services) pool their resources together and each of them contributes a fixed portion of resource of every EN. Here, the budgets are proportional to the initial contributions of the companies. Instead of resource pooling, these companies may agree upfront on their individual budgets, and then buy/rent a given set of ENs together. In these scenarios, it is important to consider both fairness and efficiency. Thus, conventional schemes such as social welfare maximization, maxmin fairness, and auction models may not be suitable. In particular, a welfare maximization allocation often gives most of the resources to users who have high marginal utilities while users with low marginal utilities receive a very small amount of resources, even nothing. Similarly, in auction models, the set of losers are not allocated any resource. Hence, these solutions can be unfair to some users. On the other hands, a maxmin fairness solution often allocates too many resources to users with low marginal utilities, hence, it may not be efficient. To strive the balance between fairness and efficiency, we advocate the General Equilibrium Theory [4], with a specific focus on the Fisher market model [5], as an effective solution concept for this problem. Specifically, the first model can be cast as a Fisher market in which services act as buyers and ENs act as different goods in the market. For the linear additive utility function as considered in this work, given resource prices, a service may have an infinite set of optimal resource bundles, which renders difficulty in designing distributed algorithms. We suggest several methods to overcome this challenge. Moreover, we show that the obtained allocation is Pareto-optimal, which means there is no other allocation that would make some service better off without making someone else worse off [6]. In other words, there is no strictly "better" allocation. Thus, a Paretooptimal allocation is efficient. We furthermore link the ME to the fair division literature [7] and prove that the allocation satisfies remarkable fairness properties including envy-freeness, sharing-incentive, and proportionality, which provides strong incentives for the services to participate in the proposed scheme. Indeed, these properties were rarely investigated explicitly in the ME literature. Envy-freeness means that every service prefers its allocation to the allocation of any other service. In an envy-free allocation, every service feels that its share is at least as good as the share of any other service, and thus no service feels envy. Sharing-incentive is another well-known fairness concept. It ensures that services get better utilities than what they would get in the proportional sharing scheme that gives each service an amount of resource from every EN proportional to its budget. Note that proportional sharing is an intuitive way to share resources fairly in terms of quantity. For the federation setting, sharing-incentive implies that every service gets better off by pooling their resources (or money) together. Finally, it is natural for a service to expect to obtain a utility of at least b/B of the maximum utility that it can achieve by getting all the resources, where b is the payment of the service and B is the total payment of all the services. The proportionality property guarantees that the utility of every service at the ME is at least proportional to its payment/budget. Thus, it makes every service feel fair in terms of the achieved utility. In the second model, the money does have intrinsic value to the services. The services not only want to maximize their revenues but also want to minimize their payments. In particular, each service aims to maximize the sum of its remaining budget (i.e., surplus) and the revenue from the procured resources, which is equivalent to maximizing the net profit (i.e., revenue minus cost). This model is prevalent in practice. For example, several service providers (SP), each of which has a certain budget, may compete for the available resources of an edge infrastructure provider (e.g., a Telco, a broker). The SPs only pay for their allocated resources and can take back their remaining budgets. Obviously, a SP will only buy a computing unit if the potential gain from that unit outweighs the cost. It is natural for the SPs to maximize their net profits in this case. The classical Fisher market model does not capture this setting since the utility functions of the services depend on the resource prices. It is worth mentioning that, conventionally, the optimal dual variables associated with the supply demand constraints (i.e., the capacity constraints of the ENs) are often interpreted as the resource prices [32] and common approaches such as network utility maximization (NUM) [33] can be used to compute an ME. However, these approaches do not work for our models that take budget into consideration. Indeed, the main difficulty in computing an ME in both models stems from the budget constraints which contain both the dual variables (i.e., prices) and primal variables (i.e., allocation). In the second model, the prices also appear in the objective functions of the services. Therefore, the ME computation problem becomes challenging. Note that the pair of equilibrium prices and equilibrium allocation has to not only clear the market but also simultaneously maximize the utility of every service (as elaborated in Section 4). Fortunately, for a wide class of utility functions, the ME in the first model can be found by solving a simple Eisenberg-Gale (EG) convex program [8]–[10]. However, the EG program does not capture the ME in the second model. Interesting, by reverse-engineering the structure of the primal and dual programs in the first model, we can rigorously construct a novel convex optimization problem whose solution is an ME of the second model. This technique can also be used to find the ME that considers other practical constraints (e.g., operation cost of the edge servers). Our main contributions include:

• Modeling. We formulate a new market-based EC resource allocation framework and advocate the General Equilibrium theory as an effective solution method for the proposed problem. • Centralized solution. The unique ME in the first model can be determined by the EG program. We also prove some salient fairness features of the ME. • Decentralized algorithms. We introduce several distributed algorithms that efficiently overcome the difficulty raised by the non-unique demand functions of the services and converge to the ME. • Extended Fisher market. We systematically derive a new convex optimization problem whose optimal solution is an exact ME in the extended Fisher market

model where buyers value the money. • Performance Evaluation. Simulations are conducted to illustrate the efficacy of the proposed techniques.

II. RELATED WORK

The potential benefits and many technical aspects of EC have been studied extensively in the recent literature. First, the hybrid edge/fog-cloud system can be leveraged to improve the performance of emerging applications such as cloud gaming and healthcare [11], [12]. A. Mukherjee et. al. [13] present a power and latency aware cloudlet selection strategy for computation offloading in a multi-cloudlet environment. The tradeoff between power consumption and service delay in a fog-cloud system is investigated in [14] where the authors formulate a workload allocation problem to minimize the system energy cost under latency constraints. A latency aware workload offloading scheme in a cloudlet network is formulated in [15] to minimize the average response time for mobile users. In [16], M. Jia et. al. explore the joint optimization of cloudlet placement and user-to-cloudlet assignment to minimize service latency while considering load balancing. A unified service placement and request dispatching framework is presented in [17] to evaluate the tradeoffs between the user access delay and service cost. Stackelberg game and matching theory are employed in [18] to study the joint optimization among data service operators (DSO), data service subscribers (DSS), and a set of ENs in a threetier edge network where the DSOs can obtain computing resources from different ENs to serve their DSSs. Another major line of research has recently focused on the joint allocation of communication and computational 4 resources for task offloading in the Mobile Edge Computing (MEC) environment [19]–[21]. MEC allows mobile devices to offload computational tasks to resourcerich servers located near or at cellular BSs, which could potentially reduce the devices' energy consumption and task execution delay. However, these benefits could be jeopardized if multiple users offload their tasks to MEC servers simultaneously. In this case, a user may not only suffer severe interference but also receive a very small amount of EC resource, which would consequently reduce data rate, increase transmission delay, and cause high task execution time on the servers. Hence, offloading decision, allocation and scheduling of radio resources, and computational resources should be jointly considered in an integrated framework

III. SYSTEM MODEL

An EC environment is depicted in Fig. ??. Besides local execution and remote processing at cloud DCs, data and requests from end-devices (e.g., smartphones, set-top-boxes, sensors) can be handled by the EC platform. Note that some data and computing need to be done in the local to keep data privacy. A request typically first goes to a Point of Aggregation (PoA) (e.g., switches/routers, BSs, APs), then it will be routed to an EN for processing. Indeed, enterprises, factories, organizations (e.g., hospitals, universities, museums), commercial buildings (shopping malls, hotels, airports), and other third parties (e.g., sensor networks) can also outsource their services and computation to the intelligent edge network. Furthermore, service/content/application providers like Google, Netflix, and Facebook can proactively install their content and services onto ENs to serve better their customers. In the EC environment, various sources (e.g., smartphones, PCs, servers in a lab, underutilized small/medium data centers in schools/hospitals/malls/enterprises, BSs, telecom central offices) can act as ENs. We consider a system encompassing various services and a set of geographically distributed ENs with different configurations and limited computing capacities. Each service has a budget for resource procurement and wants to 5 offload as many requests as possible to the edge network. The value of an EN to a service is measured in terms of the maximum revenue that it can generate by using the EN's resource. An EN may have different values to different services. Since some ENs (e.g., ones with powerful servers) can be over-demanded while some others are underdemanded, it is desirable to harmonize the interests of the services so that each service is happy with its allotment while ensuring high resource utilization. An intuitive solution is to assign prices to ENs and let each service choose its favorite resource bundle. We assume that there is a platform lying between the services and the ENs. Based on the information collected from the ENs (e.g., computing capacity) and the services (e.g., budgets, preferences), the platform computes an ME solution including resource prices and allocation, which not only maximizes the satisfaction of every service but also fully allocates the ENs' resources. In the first model, each service seeks solely to maximize its revenue under the budget constraint, without concerning about the money surplus after purchasing resources. This can be the case where the services and ENs belong to the same entity, and each service is assigned a virtual budget representing the service's priority. For instance, a Telco can give different budgets to different network slices, each of which runs a service (e.g., voice, video streaming, AR/VR, connected vehicles). In the second model, the remaining money does have intrinsic value to the services. In this case, each service aims to maximize the sum of its remaining budget and the revenue from the procured resources. For example, this can be the case where services and ENs are owned by different entities, and each SP (e.g., Google, Facebook, enterprises) has a certain budget for leasing resources from an infrastructure provider (e.g., a Telco). For simplicity, we assume that the values of ENs to the services are fixed. Our model can be extended to capture time-varying valuation in a multi-period model by considering each pair of an EN and a time slot as an independent EN

IV CONCLUSION

In this work, we consider the resource allocation for an EC system which consists geographically distributed heterogeneous ENs with different configurations and a collection of services with different desires and buying power. Our main contribution is to suggest the famous concept of General Equilibrium in Economics as an effective solution for the underlying EC resource allocation problem. The proposed solution produces an ME that not only Pareto-efficient but also possesses many attractive fairness properties. The potential of this approach are well beyond EC applications. For example, it can be used to share storage space in edge caches to different service providers. We can also utilize the proposed framework to share resources (e.g., communication, wireless channels) to different users or groups of users (instead of services and service providers). Furthermore, the proposed model can extend to the multi-resource scenario where each buyer needs a combination of different resource types (e.g., storage, bandwidth, and compute) to run its service. We will formally report these cases (e.g., network slicing, NFV chaining applications) in our future work. The proposed framework could serve as a first step to understand new business models and unlock the enormous potential of the future EC ecosystem. There are several future research directions. For example, we will investigate the ME concept in the case when several edge networks cooperate with each other to form an edge/fog federation.

Investigating the impacts of the strategic behavior on the efficiency of the ME is another interesting topic. Note that N. Chen et. al. have shown that the gains of buyers for strategic behavior in Fisher markets are small. Additionally, in this work, we implicitly assume the demand of every service is unlimited. It can be verified that we can add the maximum number of requests constraints to the EG program to capture the limited demand case, and the solution of this modified problem is indeed an ME. However, although the optimal utilities of the services in this case are unique, there can have infinite number of equilibrium prices. We are investigating this problem in our ongoing work. Also, integrating the operation cost of ENs into the proposed ME framework is a subject of our future work. Finally, how to compute market equilibria with more complex utility functions that capture practical aspects such as task moving expenses among ENs and data privacy is an interesting future research direction. It is also interesting to test the performance of the proposed approach on real datasets of an EC system when EC is widely deployed.

V. REFERENCES

[1] M. Chiang and T. Zhang, "Fog and IoT: an overview of research opportunities," *IEEE Internet Things J.*, vol. 3, no. 6, pp. 854–864, Dec. 2016.

[2] M. Satyanarayanan, "The emergence of edge computing," *Computer*, vol. 50, no. 1, pp. 30–39, Jan. 2017.

[3] W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu, "Edge computing: vision and challenges," *IEEE Internet Things J.*, vol. 3, no. 5, pp. 637–646, Oct. 2016.

[4] K.J. Arrow and G. Debreu, "Existence of equilibrium for a competitive economy," *Econometrica*, vol. 22, no. 3, pp. 265–290, 1954.

[5] W.C. Brainard and H.E. Scarf, "How to compute equilibrium prices in 1891," *Cowles Foundation*, Discussion Paper, no. 1272, 2000.

[6] A. Mas-Colell, M. D. Whinston, and J. R. Green, "Microeconomic Theory", 1st ed. New York: Oxford Univ. Press, 1995.

[7] H. Moulin, "Fair division and collective welfare," *MIT Press*, 2004.

[8] N. Nisan, T. Roughgarden, E. Tardos, and V. Vazirani, "Algorithmic Game Theory", Cambridge, U.K.: Cambridge Univ. Press, 2007.

[9] E. Eisenberg and D. Gale, "Consensus of subjective probabilities: The pari-mutual method," *Annals of Mathematical Statistics*, vol. 30, pp. 165–168, 1959.

[10] E. Eisenberg, "Aggregation of utility functions," Manage. Sci. 7, PP. 337–350, 1961.

[11] Y. Lin and H. Shen, "CloudFog: leveraging fog to extend cloud gaming for thin-client MMOG with high quality of service," *IEEE Trans. Parallel Distrib. Syst.*, vol. 28, no. 2, pp. 431–445, Feb. 2017.

[12] L. Gu, D. Zeng, S. Guo, A. Barnawi, and Y. Xiang, "Cost efficient resource management in fog computing supported medical cyberphysical system," *IEEE Trans. Emerg. Topics Comput.*, vol. 5, no. 1,
pp. 108-119, Jan.-Mar. 2017.

[13] A. Mukherjee, D. De, and D.G. Roy, "A power and latency aware cloudlet selection strategy for multi-cloudlet environment," *IEEE Trans. Cloud Comput.* , to appear.

[14] R. Deng, R. Lu, C. Lai, T. H. Luan, and H. Liang, "Optimal workload allocation in fog-cloud computing toward balanced delay and power consumption," *IEEE Internet Things J.*, vol. 3, no. 6, pp. 1171–1181, Dec. 2016.

[15] X. Sun and N. Ansari, "Latency aware workload offloading in the cloudlet network," *IEEE Commun. Lett.*, vol. 21, no. 7, pp. 1481–1484, Jul. 2017.

[16] M. Jia, J. Cao, andW. Liang, "Optimal cloudlet placement and user to cloudlet allocation in wireless metropolitan area networks," *IEEE Trans. Cloud Comput.*, to appear.

[17] L. Yang, J. Cao, G. Liang, and X. Han, "Cost aware service placement and load dispatching in mobile cloud systems," *IEEE Trans. Comput.*, vol. 65, no. 5, pp. 1440–1452, May 2016.

[18] H. Zhang, Y. Xiao, S. Bu, D. Niyato, F.R. Yu, and Z. Han, "Computing resource allocation in three-Tier IoT fog networks: a joint optimization approach combining stackelberg game and matching," *IEEE Internet Things J.*, to appear.

[19] S. Sardellitti, G. Scutari, and S. Barbarossa, "Joint optimization of radio and computational resources for multicell mobile-edge computing," *IEEE Trans. Signal Inf. Process. Netw.*, vol. 1, no. 2, pp. 89–103, Jun. 2015.

[20] X. Lyu, H. Tian, C. Sengul, and P. Zhang, "Multiuser joint task offloading and resource optimization in proximate clouds," *IEEE Trans. Veh. Technol.*, vol. 66, no. 4, pp. 3435–3447, Apr. 2017.

[21] X. Chen, "Decentralized computation offloading game for mobile cloud computing," *IEEE Trans. Parallel Distrib. Syst.*, vol. 26, no. 4, pp. 974–983, Apr. 2015.

[22] N.R. Devanur, C.H. Papadimitriou, A. Saberi, and V.V. Vazirani, "Market equilibrium via a primal–dual algorithm for a convex program," J. ACM vol. 55, no. 5, article 22, Nov. 2008.

[23] V.V. Vazirani and M. Yannakakis, "Market equilibrium under separable, piecewise-linear, concave utilities," J. ACM, vol. 58, no. 3, article. 10, May 2011.

[24] N. Chen, X. Deng, B. Tang, and H. Zhang, "Incentives for strategic behavior in Fisher market games", in *Proc. Conf. Artificial Intelligence (AAAI)*, pp. 453–459, Phoenix, Arizona, USA, Feb. 2016.

[25] X. Chen, D. Paparas, and M. Yannakakis, "The complexity of nonmonotone markets" *J. ACM*, vol. 64, no. 3, Article 20, Jun. 2017.

[26] J. Garg, R. Mehta, M. Sohoni, and V.V. Vazirani, "A complementary pivot algorithm for market equilibrium under separable, piecewise-linear concave utilities", *SIAM J. Comput.*, vol. 44, no. 6, pp. 1820–1847, 2015.

[27] Z. Liu, M. Lin, A. Wierman, S. Low, and L. L. H. Andrew, "Greening geographical load balancing,", *IEEE/ACM Trans. Netw.*, vol. 23, no. 2, pp. 657–671, Apr. 2015.

[28] L. Tang and H. Chen, "Joint pricing and capacity planning in the IaaS cloud market," *IEEE Trans. Cloud Comput.*, vol. 5, no. 1, pp. 57–70, Jan.-Mar. 2017.

[29] F. Wu and L. Zhang, "Proportional response dynamics leads to market equilibrium," in *Proc. the thirty-ninth annual ACM symposium on Theory of Computing (STOC)*, pp. 354–363, New York, NY,
USA, 2007.

ABOUT AUTHORS:

CHITIKENA ARUNA is currently pursuing

MCA in SVKP & Dr K S Raju Arts & Science College, affiliated to Adikavi Nannaya University, Rajamahendravaram. Her research interests include Software Engineering , Web Technologies and Iternet of Things.

A.N.Ramamani is working as Associate Professor in SVKP & Dr K S Raju

Arts&Science College, Penugonda, A.P.She received Master's degree in Computer Applications from Andhra University and Computer Science & Engineering from Jawaharlal Nehru Technological University, Kakinada, India. Her research interests

ISSN : 2057-5688

include Software Engineering, Web Technologies and Iternet of Things.