Joint Resource Allocation and Minimizing Carbon Footprint in Geo-distributed Fog Computing

Cuong T. Do, Pham Chuan, Choong Seon Hong
Email: {dtcuong, pchuan, cshong}@.khu.ac.kr
Department of Computer Engineering, Kyung Hee University

Abstract
In this paper, we consider the emerging problem of joint resource allocation and minimizing carbon footprint problem for video streaming service in Fog computing. To solve the large-scale optimization, we develop a distributed algorithm based on the proximal algorithm. The numerical results show that our algorithm converges to near optimum within fifteen iterations, and is insensitive to step sizes.

1. Introduction

Recently, Cisco has introduced Fog computing as a new paradigm which can transform the network edge into a distributed computing infrastructure for applications that take advantage of the billions of devices already connected to the Internet of Things (IoT) [1]. The Fog is located below the Cloud in a widely distributed manner and serves as an optimized transfer medium for services and data within the Cloud. Since Fog has wide geographical distribution, the Fog paradigm is well positioned for big data and real time analysis and it supports mobile computing and data streaming. In Fog Computing model, data, processing and applications are concentrated in devices at the network edge, rather than existing almost entirely in the Cloud, to isolate them from the Cloud systems and place them closer to the end-user. Putting computing resource near the edge allows Fog to perform low latency processing while latency tolerant and large scope aggregation can still be efficiently performed on powerful resources in the core of the Cloud. Data center resources may still be used with Fog computing, but they do not dominate over the entire picture.

In industry, many companies are ready for adopting Fog computing. Any company that delivers content can start using Fog computing. A good example is Netflix who is able to reach a large number of globally distributed customers. The delivery of video-on-demand service would not be efficient enough if it is based on the data management in one or two central data centers. Fog computing thus allows providing very large amounts of streamed data by delivering the data directly into the vicinity of the customer [2].

Today, geographical resource allocation and energy cost are managed independently, leading to poor performance and high costs in many cases [3]. The objectives of the two decisions can also be misaligned and lead to sub-optimal equilibria. In this paper, we study the joint resource allocation and minimizing carbon footprint problem for streaming service in Fog computing. We assume that physical devices called Fog computing nodes (FCNs) are placed in the network infrastructure to deliver video streaming service from content providers. For example, streaming services are hosted at the network edge such as ``smart'' routers and switches with more application-level functionality, or even end devices such as set-top-boxes or access points. By doing so, Fog reduces service latency, and improves QoS. FCN aggregates video demands from nearby end users. As the tenants of a Fog provider, the content providers want to delivery content from the data center as much as possible to FCN to increase their utility. Because the applications and end users are heterogeneous, the utility varies significantly depending on the geographical distribution of end-users. Thus, we need to control the fraction of traffic (from huge number of end-users to data center) to maximize the utility of content providers and minimize the carbon footprint at the data center.

In a large scale systems, fast distributed resource allocation and social welfare maximization are critical problems. Traditional solutions to such problems rely on primal/dual decomposition and gradient methods [4], however these methods have slow convergence speed and sensitivity to step sizes and require strict convex assumptions. Related work has considered geo-distributed and large scale system such as work in [5] for video streaming in cloud computing. In [5], the authors propose a distributed algorithm with fast convergence speed but it requires strict convex assumptions like primal/dual decomposition and gradient methods. Here, the joint resource allocation and minimizing carbon footprint problem is a very large-scale convex optimization due to large number of FCN (dozen of thousand devices [1]). Thus, for reasons of performance and scalability, we introduce a distributed solution for the joint resource allocation and minimizing carbon footprint problem. Our algorithm is based on proximal algorithms [6], a powerful algorithm that recently has been applied in many large scale distributed convex optimization problems. Comparing to conventional methods such as gradient methods, proximal algorithms have faster convergence speed with modest accuracy, insensitivity to step sizes, and robustness without strong assumptions such as strict convexity of the objective function.

2. System model

We consider a content provider run Fog computing services over a data center and N Fog Computing Nodes (FCNs) are located at the edge of network in distinct geographical regions to serve video streaming for end users as illustrated in Fig. 1. We assume that FCNs send a request to the data center. After the data center have finished serving request, it sends the response video streaming back to FCNs. We use x_n to denote the amount of video streaming to FCN_n from the data center. We assume that the reservation of egress network bandwidth from the data center to FCNs has already been enabled by the detailed engineering techniques proposed in [7]. Thus, the data center can guarantee the bandwidth to serve video streaming of clients at FCNs. To model the social welfare, we consider both the carbon footprint cost at the data center and utility of the content provider, which are detailed below.

We consider an affine utility function that is the de facto utility function widely used in the literature [8]. An affine utility function at FCN_n has the following form:

$$U_n(x_n) = \alpha_n x_n. \tag{1}$$

Where α_n is a conversion factor that translates user-perceived request video streaming into utility (e.g. revenue). Because the applications and end users are heterogeneous, α_n varies significantly depending on the geographical distribution of end-users.

For the environmental cost, the carbon footprint of energy at the data center can also be taken into account.

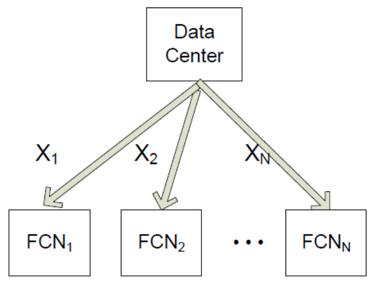

Fig.1.Multiple Fog Computing Nodes with a Data Center.

The cost function $C(\cdot)$ considered in existing work [3] is given as follows

$$C(y) = c \cdot r \cdot PUE \cdot P(y), \tag{2}$$

where c denotes the carbon footprint cost in term of \$/g at the data center, r is the average carbon emission rate g/KWh, PUE is the power usage effectiveness and $P(y)$ represents the server power at the data center. $P(y)$ represents the server power at the data center, which is a function of the total of requested video streaming y and can be obtained empirically. From a measurement study by Google [9], a commonly used server power function is given by

$$P(y) = \varsigma \cdot P_{\{idle\}} + \left(P_{\{peak\}} - P_{\{idle\}}\right) \cdot y \cdot \kappa, \tag{3}$$

where κ is a conversion factor that translates requested video streaming into workload, ς is workload capacity of the data center, $P_{\{idle\}}$ is server idle power and $P_{\{peak\}}$ is peak power.

3. Joint Resource Allocation and Minimizing Carbon Footprint Problem

We now formulate joint resource allocation and minimizing carbon footprint problem. Putting the utility and cost function, the social welfare maximization problem can be written as

$$\max_{\{x_n \geq 0, y \geq 0\}} \sum_{n=1}^{N} U_n(x_n) - C(y) \ , \tag{4}$$

$$\sum_{n=1}^{N} x_n = y \leq \varsigma / \kappa, \tag{5}$$

The constraint is the capacity constraint at data center. If Problem (4) is small then it would be easy to solve. However, in Fog, Problem (4) is an extremely large-scale optimization problem. Thus, we need a distributed algorithm to solve such a large scale problem.

We rewrite Problem (4) as follows

$$\max_{\{x_n \geq 0, z_n \geq 0\}} \sum_{n=1}^{N} U_n(x_n) - C\left(\sum_{n=1}^{N} z_n\right) , \tag{6}$$

$$x_n = z_n, \forall n, \tag{7}$$

$$\sum_{n=1}^{N} z_n \leq \varsigma / \kappa, \tag{8}$$

Where $z = [z_1, z_2, \dots, z_N]$ is an auxiliary variable. Since Problem (6) is a sharing problem, we have the iterations based on proximal algorithms as follows

$$x_n^{k+1} := prox_{\lambda U_n}(z_n^k - u_n^k), \forall n \ , \tag{9}$$

$$z^{k+1} := prox_{\lambda c}\left(x^{k+1} + u^k\right), \quad (10)$$
$$u_n^{k+1} := u_n^{k+1} + x_n^{k+1} - z_n^{k+1}, \forall n, \quad (11)$$

where x is the vector $[x_1, x, \dots, x_N]$, u_n is an interpretation of Lagrange multiplier. The proximal operator $prox$ is defined by

$$prox_{\lambda f}(x^k) = \min_{\lambda f} \quad f(x) + 1/(2\lambda) \parallel x - x^k \parallel_2^2, \quad (12)$$

where $f : \Re^n \to \Re \cup \{+\infty\}$ is a closed proper convex function, k is the iteration counter, and x^k denotes the kth iteration of the algorithm, $\parallel \cdot \parallel_2^2$ is the usual Euclidean norm [6]. Thus, we have a distributed algorithm (Algorithm 1) to solve the joint resource allocation and minimizing carbon footprint problem.

Algorithm 1: Optimal Distributed Solution for (6)

1: Choose randomly initial values u^0, x^0, z^0

2: *while not convergence*

3: Each FCN_n updates request video streaming x_n^{k+1} by using x-update in (9). Send its request x_n^{k+1} and u_n^k to the data center.

4: The data center collects request video streaming x_n^{k+1} and u_n^k from all FCN and updatesz^{k+1} by using z-update in (10). Then, the data center sends response video streaming x_n^{k+1} with z_n^{k+1} to FCN_n for all n.

5: Each FCN_n updates dual value u_n^{k+1} by using u-update in (11).

6: *end while*

Numerical results: The numerical parameters are set as follows: $c = 19 \cdot 10^{-6}$, $r = 562$, $PUE = 1.5$, $P_{\{peak\}} = 200$, $P_{\{idle\}} = 100$ W, $\alpha_n = \text{Uniform}(0{,}1000)$ · Fig. 2 plots the convergence of objective values. Since the number of iterations is small while the number of user 100, it suggests that our algorithm can solve a large-scale problem effectively.

5. Conclusion

We study the joint resource allocation and minimizing carbon footprint problem for streaming service in Fog computing. We formulated the problem as a general convex optimization, where the location diversity of requested video streaming utility and costs are modeled. We developed an efficient distributed algorithm based on the proximal algorithm to decompose the large scale global problem into many sub-problems, each of which can be quickly solved. The numerical results are conducted to evaluate the algorithm's performance

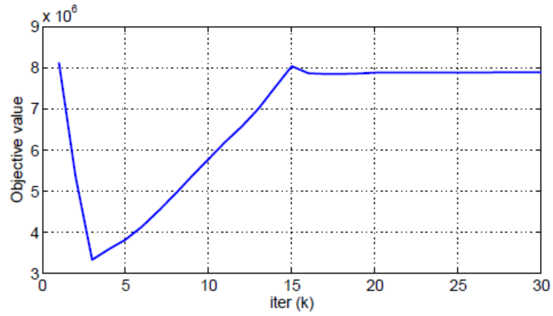

Fig.2.Convergence of Algorithm 1.

Acknowledgement

This work was supported by the ICT R&D program of MSIP/IITP, Republic of Korea. (2014-044-011-003, Open control based on distributed mobile core network) Dr. CS Hong is the corresponding author.

Reference

[1] F. Bonomi, R. Milito, J. Zhu, and S. Addepalli, “Fog computing and its role in the internet of things,” in Proceedings of the first edition of the MCC workshop on Mobile cloud computing. ACM, 2012, pp. 13-16.

[2] Fog computing is a new concept of data distribution. [Online]. Available:http://www.cloudtweaks.com/2013/12/fogcomputingis-a-new-concept-of-datadistribution/

[3]Z. Liu, M. Lin, A. Wierman, S. H. Low, and L. L. Andrew, “Greening geographical load balancing,” in Proceedings of the ACM SIGMETRICS joint international conference on Measurement and modeling of computer systems. ACM, 2011, pp. 233-244.

[4]S. Boyd, L. Xiao, and A. Mutapcic, “Subgradient methods,” lecture notes of EE392o, Stanford University, Autumn Quarter, vol. 2004, 2003.

[5]D. Niu and B. Li, “An efficient distributed algorithm for resource allocation in large-scale coupled systems,” in INFOCOM, 2013, pp. 1501-1509.

[6]N. Parikh and S. Boyd, “Proximal algorithms,” Foundations and Trends in Optimization, vol. 1, no. 3, pp. 123-231, 2013.

[7]D. Xie, N. Ding, Y. C. Hu, and R. Kompella, “The only constant is change: incorporating time-varying network reservations in data centers,” ACM SIGCOMM Computer Communication Review, vol. 42, no. 4, pp. 199-210, 2012.

[8]S. Narayana, J. W. Jiang, J. Rexford, and M. Chiang, “To coordinate or not to coordinate? wide-area traffic management for data centers,” Technical report, Princeton University, Tech. Rep., 2012.

[9]X. Fan, W.-D. Weber, and L. A. Barroso, “Power provisioning for a warehouse-sized computer,” in ACM SIGARCH Computer Architecture News, vol. 35, no. 2. ACM, 2007, pp. 13-23.