A Heuristic Algorithm Based on Resource Requirements Forecasting for Server Placement in Edge Computing

Kaile Xiao, Zhipeng Gao*, Qian Wang, Yang Yang
State Key Laboratory of Networking and Switching Technology
Beijing University of Posts and Telecommunications
Beijing, China
Email: {xiaokaile, gaozhipeng, wangqian1991, yyang}@bupt.edu.cn*

***Abstract*—The placement of edge computing server is the key to the rapid development of edge computing. We propose prediction-mapping-optimization heuristic based on resource requirements forecasting for server placement in edge computing. Through this algorithm, we divide the task into multiple subtasks, and then realize the mapping of subtask-location of server, and finish the information interaction between the servers and the data source through the data naming mechanism proposed by us. With the goal of the lowest cost of service providers, we propose a cross-region resource optimization model and obtained the final server placement strategy.**

***Keywords*-edge computing; server placement; NDN; resource optimization**

I. Introduction

In an edge-cloud environment, placement of edge computing server is required, in addition to intelligent terminals that can be used as edge computing nodes. By depending on a priori experience and intuition alone, it does not achieve the lowest cost for the server providers and the best user experience. And the placement strategy of existing service devices (such as sensors, base stations, etc.) cannot be used for placement of edge computing server. In the edge computing environment, data source (such as mobile phones, laptops, cars, buses, etc.) that generate massive data is likely to be in the mobile state, and latency of request is sensitive and the amount of computing is large[1]. We predict the resource requirements of the data source in the edge computing (including computing resources, storage resources, etc.), then propose a heuristic algorithm to develop the placement strategy of the edge computing server.

II. Overview

In this paper, we propose the strategy of edge computing server placement by resource requirements forecasting which is shown in Fig 1. This placement strategy is more advantageous for service providers than traditional solutions.

First, we propose a data naming mechanism for the servers and the data source. They can interact information by this mechanism which including information of location, current time, and so on. On this basis, we introduce the non-homogeneous Markov model to predict the next destination

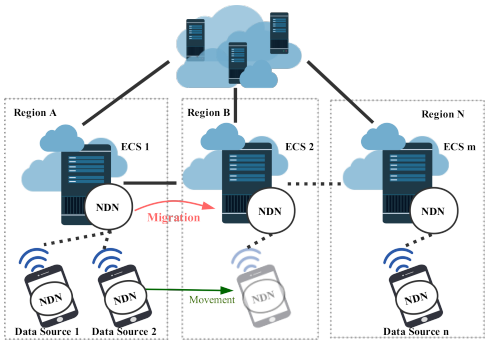

Figure 1. Edge computing servers (ECS) placement in edge-cloud environment

of data source. There are many servers in each region, and how to choose the server in the certain region is also a big challenge. To address the challenge of server selection, we propose a heuristic algorithm to map the subtasks to candidate server locations. At last, we propose a cross-region resource optimization algorithm to minimize the cost of the service providers. In this step, We choose a collection of servers for subtask-location mapping. The location of the server in the collection can be in this region or in adjacent regions. Resource scheduling allows subtasks to be processed cross-regions, which can minimize the cost of service providers by reducing the number of server.

III. Algorithms

A. Data Naming Mechanism for Information Interaction

The edge computing server needs to predict the resource requirements of data source. The prediction is based on the information which is from data source, such as time in a certain place. The data source sends information to the server, which requires a more standard mode, including what content is sent, when it is sent, and how each element of the specific content needs to be sorted.

Implementing the above functions requires a well-defined data naming mechanism, which needs to be adhered to by the data source and servers. In addition, the data source

requires a certain time interval to send information to the server. The time interval should not be too small or too large, which affects the accuracy of data. According to the mobility characteristics of the data source, the data naming specification based NDN[2] is defined as: [current timestamp, location, transition probability, common application, random number].

B. Predict Next Destination of Data Source

Based on the information interaction between the servers and the data source, we introduce the non-homogeneous Markov chain to predict the next destination of the data source. A simple Kolmogorovs equation is obtained if $\mathbf{Q}(t)$ is constant in the period:

$$\mathbf{P}\left(t, t+T\right) = e^{\mathbf{Q}(t)T}\mathbf{P}(t,t) = e^{\mathbf{Q}(t)T} \tag{1}$$

Where $\mathbf{P}$ is the matrix of state transition probability, $\mathbf{Q}(t)$ means the matrix of transition intensity, and $\sum_{k=1}^{N} q_{jk}(t) = 0$ $(q_{jk}(t) \in \mathbf{Q}(t))$.

After determining the next destination, the problem we faced was how to choose the right server within the destination partition. We want to find a set of location of server $Set_1 = \{Sr_1, \cdots, Sr_i, Sr_j, \cdots Sr_n\}$. The servers in this collection need to meet the following conditions:

$$bandwidth_{Sr_n} \geq \frac{data_i}{\tau_{sub_i}} \tag{2}$$

$$\sum\nolimits_{i=1}^{m} \left(t\_i_{Sr_i,Sr_j} + t\_i_{Sr_j} + t\_i_{Sr_j,ds_n}\right) \leq \tau_{total} \tag{3}$$

where the amount of data and time of processing time for the entire task are $data_{total} = \sum_{i=1}^{m} data_i$ and $\tau_{total} = \sum_{i=1}^{m} \omega_i \tau_{sub_i}$. $t\_i_{Sr_i,Sr_j}$ means the time when the i-th subtask is migrated from the server i to the server j, $t\_i_{Sr_j}$ means the time that the i-th subtask to be processed on the server j, $t\_i_{Sr_j,ds_n}$ means the time that the completion result of i-th subtask is transmitted from the server j to the data source.

C. The Heuristic Algorithm for Subtask-Location Mapping

In this paper, we map the subtask-location by using a heuristic algorithm. In order to simplify the model, we only consider the lowest of the total consumption cost of data source(user) in the step. So our goal is to get a collection of server placement $Set_2 = \{Sr_1, \cdots, Sr_i, Sr_j, \cdots Sr_n\}$ that are designed to minimize data source cost. The optimization problem can be formalized as follows:

$$\min \mathrm{Cost}_{ds_{total}} = \sum\nolimits_{i=1}^{m} \left(\mathrm{Cost}\_i_{ds_{Sr_i,Sr_j}} + \mathrm{Cost}\_i_{ds_{Sr_j}}\right) \tag{4}$$

And the server needs to meet:

$$t_{total} = \sum\nolimits_{i=1}^{m} \left(t\_i_{Sr_i,Sr_j} + t\_i_{Sr_j} + t\_i_{Sr_j,ds_n}\right) \leq \alpha\tau_{total} \tag{5}$$

Where $\mathrm{Cost}_{ds_{total}}$ means the total cost of consumption for the data source, $\mathrm{Cost}\_i_{ds_{Sr_i,Sr_j}}$ means the migration cost of the i-th subtask from the server i to the server j. $\mathrm{Cost}\_i_{ds_{Sr_j}}$ means the calculation cost of the i-th subtask processing on the server j. α is the adjustment parameter and the maximum task processing time is allowed when $\alpha = 1$.

D. Cross-region Resource Optimization for The Lowest Cost of Service Provider

We propose an improved algorithm to minimize cost of the service providers. That is, optimization is performed based on the average bandwidth of the server in the collection of server candidate(Set_2). We add a relaxation factor ζ_i in the server selection algorithm, and change β and γ in Eq.7 and 8. Our goal is to get a set of server collection $Set_3 = \{Sr_1, \cdots, Sr_i, Sr_j, \cdots Sr_n\}$ to minimize cost of service providers. The location of the server in the collection may be within the region or within the adjacent regions(cross-region). The optimization problem can be formalized as follows:

$$\min \sum\nolimits_{i=1}^{n} \mathrm{Cost}_{sp_{Sr_i}} \tag{6}$$

And the server needs to meet:

$$t_{total} = \sum\nolimits_{i=1}^{m} \left(t\_i_{Sr_i,Sr_j} + t\_i_{Sr_j} + t\_i_{Sr_j,ds_n}\right) \leq \beta\tau_{total} + \zeta_i \tag{7}$$

$$\sum\nolimits_{i=1}^{m} \left(\mathrm{Cost}\_i_{ds_{Sr_i,Sr_j}} + \mathrm{Cost}\_i_{ds_{Sr_j}}\right) \leq \gamma \mathrm{Cost}_{ds_{total}} \tag{8}$$

$\mathrm{Cost}_{sp_{Sr_i}}$ means the cost that the service provider deploying the i-th server.

IV. Conclusion

In this paper, we propose a prediction-mapping-optimization heuristic for the placement of server in edge computing. The range of candidate servers are narrowed down by predicting the next destination of the data source, and the server and the data source exchange information by the data naming mechanism. Then we accumulate the resources of the each candidate server through the subtask-location mapping, and determine the location and number of the server according to the amount of resources. At last, we propose the cross-regional resources optimization to minimize the cost of the service providers.

Acknowledgment

This work is supported by National Key Research and Development Program of China (2016YFE0204500).

References

[1] Chiang M, Zhang T, *Fog and IoT: An Overview of Research Opportunities*, IEEE Internet of Things Journal. 2016, 3(6):854-864.

[2] Weisong Shi, JC, Quan Zhang, Youhuizi Li, Lanyu Xu, *Edge Computing: Vision and Challenges*, IEEE Internet of Things Journal, 2016, 3(5):637-646.