ALTO WG
Internet-Draft
Intended status: Standards Track
Expires: June 1, 2020

S. Yang
L. Cui
Shenzhen University
M. Xu
Tsinghua University
H. Shen
China Telecom
L. Chen
China Mobile

November 29, 2019

Delivering Functions over Networks: Traffic and Performance Optimization for Edge Computing using ALTO
draft-yang-alto-deliver-functions-over-networks-00

Abstract

With development of Internet of Thing (IoT), artificial intelligence, huge amount of data are generated and need to be processed. To satisfy the user demands, service providers are deploying edge computing across lots of data centers, which are closer to users. In order to achieve better performances, computing functions need to be scheduled properly over networks. However, it is challenging to deploy functions to the distributed edge servers efficiently due to the lack of network traffic information. [RFC5693] and [RFC7285] introduce and define the Application-Layer Traffic Optimization, or ALTO, to compute and provide the network information for the distributed applications using the ALTO protocol. In this document, we employ the ALTO protocol to deliver functions in edge computing platform, where the protocol will provide the network information for the distributed edge computing servers and guide the delivery process. The usage of ALTO will improve the efficiency of function delivering in edge computing.

time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on June 1, 2020.

Table of Contents

1. Introduction

Internet of Things (IoT), artificial intelligence, virtual reality and augmented reality (VR/AR) are developing rapidly and promising in the future. The new applications are generating huge amount of data that need to be processed efficiently. The emergence of edge computing improves the performance by deploying servers at the edge,

such that the selected servers would be closer to users, and the latency/bandwidth between users and edge servers would be guaranteed.

Function as a service (FaaS) is becoming more and more popular among cloud computing providers, e.g., Amazon Lambda and IBM Openwhisk. The current FaaS platform can schedule computing resources efficiently in a computing cluster. However, deploying functions over distributed networks is more challenging due to the lack of network states and information, including network traffic, topology, and other cost metrics, etc. In this document, we will deliver functions over the edge computing networks, to utilize the computing and network resource more efficiently.

We use the ALTO (Application-Layer Traffic Optimization) [RFC7285] to optimize the network traffic and performance in delivering functions over the edge computing network. ALTO can provide global network information and network traffic for the distributed applications, while the information can not be retrieved or computed by the applications themselves [RFC5693]. Generally, ALTO protocol will collect and compute the network information for the distributed edge clusters, including link delay, network traffic, and other cost metrics, and help guide the deliver decision process in edge computing. Finally, the edge computing system will deliver the functions to the most appropriate edge clusters according to the information by ALTO protocol.

For brevity, in this document, we will use the terminologies introduced in [RFC7285] and [I-D.ietf-alto-unified-props-new].

2. Conventions and Terminology

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [RFC2119].

3. Background

3.1. Edge computing

The proposal of edge computing improves the edge network performance in terms of latency, security, bandwidth, etc. In edge computing infrastructure, servers are deployed at the edge, where the network performance between servers and users are better. Users can submit their tasks to the edge servers, which will process the tasks and return the computational results to the users. Compared with traditional centralized computing, the latency, bandwidth and network traffic performance of edge computing is better. Nowadays, edge

computing is used in different areas, e.g., latency-sensitive applications such as IoT, artificial intelligence, 5G, etc.

The FaaS technology allows network resources to be dynamically allocated to computing clusters. Users can apply for function-based computation services (including object detection, big data analysis, etc.) from FaaS providers, and avoid the complicated environment configuration and resource management process. Developers can focus on their business and codes rather than environment management, which will increase the efficiency of application development and save costs for individuals and companies.

To improve the network performance, we will deliver functions over edge computing, such that computing functions can be dynamically scheduled in distributed edge computing network. However, when deploying functions to edge servers, network traffic, topology and other metrics will influence the performance in terms of latency and throughput. Therefore, we SHOULD consider the network traffic, and try to optimize the network performance of the platform.

3.2. Benefits of ALTO protocol

Application-Layer Traffic Optimization (ALTO) [RFC7285] is designed to provide network information for the distributed applications. More specifically, the ALTO server will offer necessary network states and information and guide the resource scheduling process for distributed applications that can not retrieve the information by themselves. The ALTO protocol will provide the essential network information, including network traffic, cost map, and cost metrics, which are necessary in the resource selection process. In this case, the distributed applications are allowed to manage the network traffic, and select a better path with low delay to access the network and process the computation tasks.

In edge computing, since the edge computing clusters are distributed in the network, they have different network states, including the link delay and network traffic. When delivering functions, the delivery decision SHOULD be adaptive to the network states in order to achieve a better latency. Therefore, the ALTO protocol can help manage the network information and traffic, such that the function can be delivered to a proper edge computing cluster with low latency and users can enjoy a better edge computing service.

4. Scenario of delivering function

Suppose a scenario in Internet of Things (IoT), where the surveillance cameras are distributed, connected via the Internet and applying for object detection computing service. When a camera

submit a task, the objection detection function will be delivered to an edge server that handles the task and returns the results to the camera. The system will request and retrieve the network information, including link delay and other cost metrics, by the ALTO protocols from ALTO servers and clients. According to the information provided by ALTO, the function and task will be delivered to the most appropriate edge server that has the best performance from the cameras. The infrastructure is demonstrated in Figure 1.

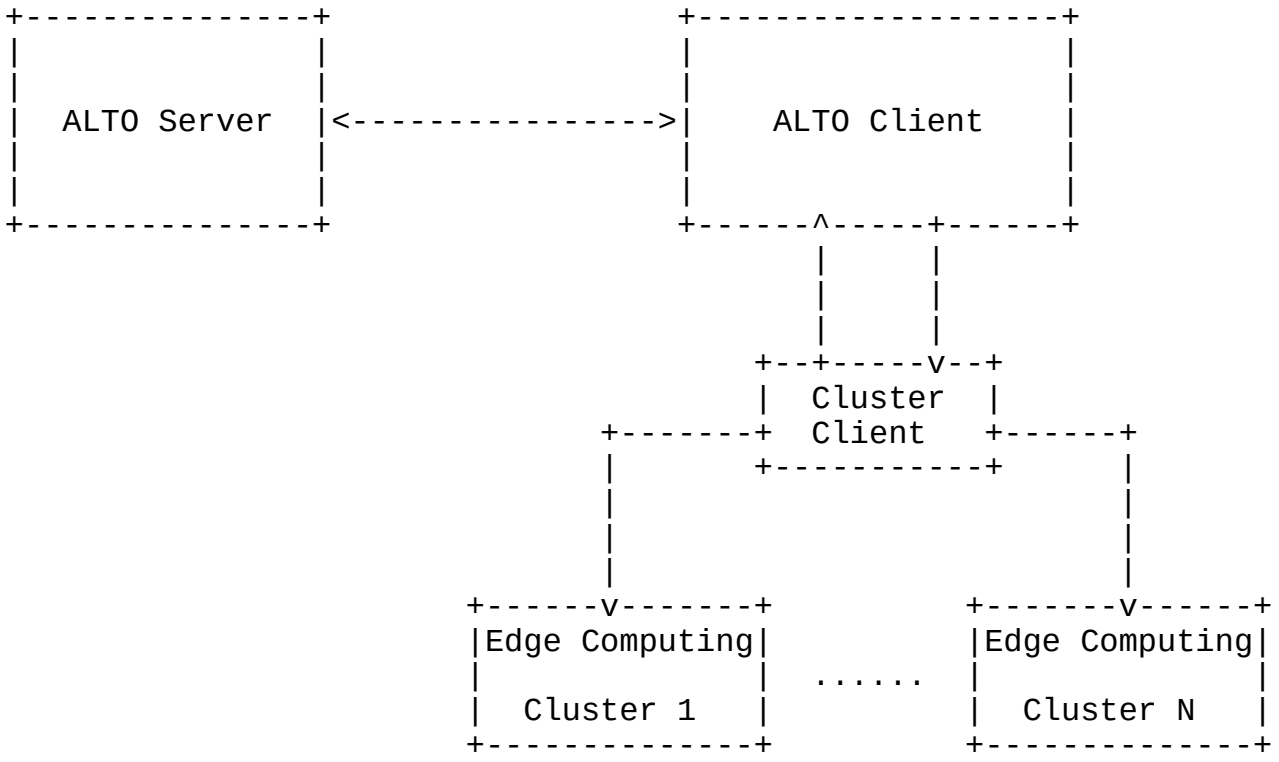

Figure 1. Scenario of delivering function over edge network in IoT

5. Delivering functions over edge computing with ALTO protocol

In edge computing platform, since lots of edge clusters and servers are distributing in the network, the system MUST handle the huge amount of edge devices and their corresponding network traffic. A cluster client is employed to manage the connectivity and traffic information of the distributed edge clusters. The ALTO client will communicate with the cluster client and provide the necessary network information. The usage of ALTO is to optimize the network traffic and guide the function delivering process in edge computing. It will provide the overall network states and information for the distributed edge clusters, and decide the appropriate edge cluster to deploy the functions.

More specifically, the ALTO server will collect and compute the network cost metrics, including the link delay, availability, network traffic, bandwidth, etc. Then the information will be sent to the ALTO client. The ALTO client will select the target appropriate edge clusters to deploy the target function. Finally, the system will connect and deploy the function to the target servers, such that users can submit their computation task to the selected edge clusters.

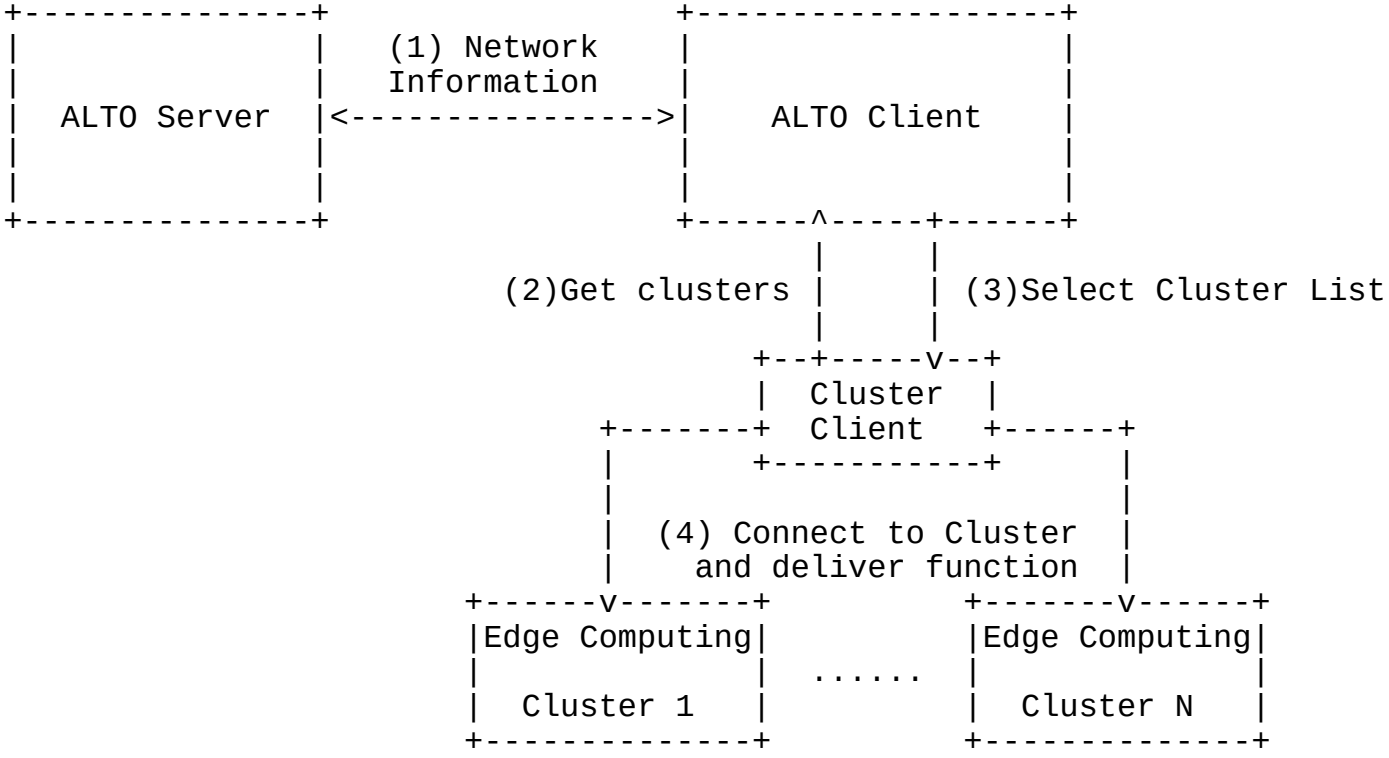

Figure 2. Delivering process in edge computing platform with ALTO

Figure 2 illustrates the infrastructure and function delivering process of the edge computing platform.

1. The ALTO client requests the information, such as network map and cost map of distributed edge clusters from the ALTO server by using ALTO protocol.

2. The Cluster Client requests edge cluster list of the network.

3. The ALTO Client returns the edge cluster list and corresponding resource information about the clusters computed by ALTO servers according to the network state.

4. The Cluster Client connects and delivers function to the corresponding edge computing cluster according to the information, and the cluster will process and return the computation results to users.

Note that the data transfer process is using the ALTO protocol described in [RFC7285] to guarantee the efficiency and security of the delivering process. In this case, the edge computing clusters are allowed to retrieve the network information, such that the function can be delivered to the proper ones to achieve a better performance in terms of latency, throughput, etc.

6. Implementation and Deployment

6.1. Implementation

We are inspired by the concept of Serverless Computing, which is a new computing paradigm providing function-based computing service, and utilize the containerization technology to run the functions. The container, including the running code, library, and data dependencies, will be deployed and orchestrated to target edge servers and clusters by container orchestrator Kubernetes (or K8S). The container orchestration scheme will be computed according to the network information provided by ALTO.

We use IBM OpenWhisk as the FaaS platform in edge clusters, where the resources are managed by K8S. Using the containerization technology, functions can be flexibly delivered to the target edge server, When a user request for function-based edge computing services, its request will be redirected to the edge server for better performance.

6.2. Deployment

We have implemented a prototype, and are deploying it in real networks across different service providers (T.B.D).

6.3. ALTO Integration

T.B.D.

7. Security Considerations

T.B.D.

8. IANA Considerations

This document includes no requests to IANA.

9. References

9.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", March 1997.

[RFC5693] Seedorf, J. and E. Burger, "Application-Layer Traffic Optimization (ALTO) Problem Statement", RFC 5693, DOI 10.17487/RFC5693, October 2009, <https://www.rfc-editor.org/info/rfc5693>.

[RFC7285] Alimi, R., Ed., Penno, R., Ed., Yang, Y., Ed., Kiesel, S., Previdi, S., Roome, W., Shalunov, S., and R. Woundy, "Application-Layer Traffic Optimization (ALTO) Protocol", RFC 7285, DOI 10.17487/RFC7285, September 2014, <https://www.rfc-editor.org/info/rfc7285>.

9.2. Informative References

[I-D.ietf-alto-unified-props-new] Roome, W., Randriamasy, S., Yang, Y., Zhang, J., and K. Gao, "Unified Properties for the ALTO Protocol", draft-ietf-alto-unified-props-new-09 (work in progress), September 2019.

Authors' Addresses

Shu Yang
Shenzhen University
South Campus, Shenzhen University
Shenzhen 518060
P.R. China

Phone: +86-755-2653-4078
Email: yang.shu@szu.edu.cn

Laizhong Cui
Shenzhen University
South Campus, Shenzhen University
Shenzhen 518060
P.R. China

Phone: +86-755-8695-6280
Email: cuilz@szu.edu.cn

Mingwei Xu
Tsinghua University
Department of Computer Science, Tsinghua University
Beijing 100084
P.R. China

Phone: +86-10-6278-5822
Email: xumw@tsinghua.edu.cn

Hongfei Shen
China Telecom
5055, Yitan Road
Shenzhen 518000
P.R. China

Email: 13360090006@189.cn

Lu Chen
China Mobile
19, Jiefang East Road
Hangzhou 310016
P.R. China

Email: chenglu@zj.chinamobile.com