Workgroup: ALTO Working Group
Internet-Draft:
draft-yang-alto-multi-domain-01
Published: 13 March 2023
Intended Status: Standards Track
Expires: 14 September 2023
Authors: Y. Yang
Yale University
M. Lassnig
CERN

ALTO Multi-Domain Services

Abstract

Application-Layer Traffic Optimization (ALTO) provides means for network applications to obtain network information. In the definitions of ALTO services ([RFC7285] and existing extensions), there is no requirement on whether the source and the destination endpoints must belong to the same autonomous network, which is a single-domain setting, or they can belong to different autonomous networks, which is a multi-domain setting. This document explains problems of realizing ALTO in multi-domain settings and then presents 3 potential solutions to realize ALTO multi-domain services.

Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at https://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any 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 14 September 2023.

Table of Contents

1. Introduction

Application-Layer Traffic Optimization (ALTO) provides means for network applications to obtain network information. For example, the Endpoint Cost Service (ECS) defined by ALTO in [RFC7285] can provide the network costs of data transmissions from a set of sources to a set of destinations. The costs (called distances) then can be used by Rucio to rank data sources or destinations to make data orchestration decisions, where Rucio is the de facto data orchestration system of CERN experiments.

As another example, to extend FTS, which is the data scheduling system of CERN experiments, to realize resource allocation to multiple experiments sharing the same network link, the ongoing TCN

project need the ALTO path vector service to map each source-destination pair to the links used by the pair. The project then computes the total traffic sent by each activity of each experiment on a given link, where each experiment consists of a set of activities, each activity consists of a set of data transfers, and each data transfer has a given source-destination pair. With the aggregation, TCN computes scheduling of data transfers according to resource allocation policies.

In the definitions of ALTO services ([RFC7285] and existing extensions), there is no requirement on whether the source and the destination endpoints must belong to the same autonomous network, which is a single-domain setting, or they can belong to different autonomous networks, which is a multi-domain setting. The unification of a single interface covering both single-domain and multi-domain settings provides a simple-to-use interface to ALTO clients. However, it leaves standardization gaps in multi-domain settings. Although participating autonomous systems can define private mechanisms to realize ALTO services in multi-domain settings, standard mechanisms allow wider deployment.

This document first specifies the issues that may arise in providing ECS in multi-domain settings. It then provides initial designs, based on current implementation experiences to start the design conversation. To be concrete, this document is based on the basic ALTO ECS service. Additional complexities such as network maps and cost maps will be discussed in the next iteration.

2. Multi-domain Problem

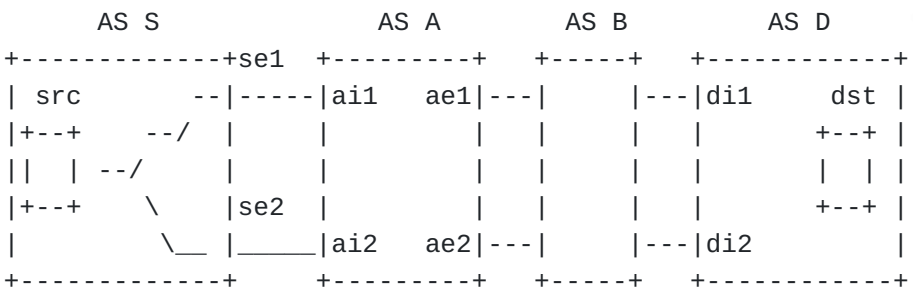
2.1. Use Cases

Consider the following ECS query realizing ALTO ECS for LHCONE to support Rucio. The source is located at CERN and the destination candidates are at multiple locations of the LHCONE network (BNL, Caltech, and KIT, for example).

```
POST /endpointcost/lookup HTTP/1.1
Host: alto.example.com
Content-Length: 248
Content-Type: application/alto-endpointcostparams+json
Accept:
   application/alto-endpointcost+json,application/alto-error+json

{
  "cost-type": {"cost-mode" : "numerical",
                "cost-metric" : "routingcost"},
  "endpoints" : {
    "srcs": [ "ipv4:128.141.201.74" ],
    "dsts": [
      "ipv4:130.199.4.27",
      "ipv4:104.18.24.74",
      "ipv4:141.3.128.6"
    ]
  }
}


HTTP/1.1 200 OK
Content-Length: 274
Content-Type: application/alto-endpointcost+json

{
  "meta" : {
    "cost-type": {"cost-mode" : "numerical",
                  "cost-metric" : "routingcost"
    }
  },
  "endpoint-cost-map" : {
    "ipv4:128.141.201.74": {
      "ipv4:130.199.4.27" : 20,
      "ipv4:104.18.24.74" : 30,
      "ipv4:141.3.128.6"  : 10
    }
  }
}
```

It is straightforward to change the query to be the ALTO path vector service, to support TCN: the value is a vector of network links (e.g., [link-1, link-2, ...]), not the numerical routingcost (e.g., 20).

The use cases provide examples of multi-domain settings, which the figure below shows. We choose one of the destinations as an example. For such a query, the path from src to dst spans multiple autonomous networks.



2.2. Challenge: Distributed Information

In the Internet setting which we consider, the network information of the path from the src to the dst spreads into multiple autonomous networks: 4 autonomous networks (AS A, B, and D) in the example. BGP collects information from multiple autonomous networks through back propagation from the destination, but the information is coarse-grained, and incomplete.

Source: The BGP router at AS S knows that the path from src to dst consists of the AS-PATH [S A B D]. Combining BGP and intradomain routing, AS S will also know which one of the two egress routers (se1, se2) that it will use to forward traffic to dst. However, AS S does not know more details downstream: for example, it does not know whether the packet will use ae1 or ae2 as the egress router at AS A to enter AS B; neither does it know the internal routing inside AS A. Hence, an ALTO server provided by AS S cannot provide all of the information for the example ECS query.

Non-Source AS: A non-source AS knows the AS-PATH starting from itself to dst. But it may not know the ingress point. For example, AS A does not know whether the packet will come in from ai1 or ai2. Hence, an ALTO server provided by AS A may consider the example ECS query as an ambiguous query (because it gives only source (src) and destination (dst), but it does not in general know the ingress point).

2.3. Challenge: Partial Deployment

It is possible to design protocol extensions to collect the aforementioned distributed information to provide complete information (see below), but one challenge is that the deployment may be only incremental and hence is partially deployed during the process. Consider the example, assume that AS B will run only standard protocols (also no traceroute) and will not provide extended ALTO, then the ingress point to D will be ambiguous.

3. Candidate Solutions

During the process of integrating ALTO into Rucio and FTS, multiple solution candidates are discussed and below we enumerate each of them.

3.1. Candidate Solution: Routing Layer Design

This is a type of solution that makes it possible to collect all needed network information at a single autonomous network, and then use an ALTO server at the source network to abstract and expose the information. One natural candidate is to modify the routing control plane itself: BGP extensions, which can be extended to collect needed information and propagate upstream. For example, when a BGP router at AS A (e.g., ai1) propagates BGP info to its peer at AS S (se1), it includes not only the AS-PATH [A, B, D], but also additional information so that the upstream can construct the complete path cost (distance) metrics. The upside of this design is that it integrates with routing system and hence may even extend routing capabilities. However, routing protocol extensions can be complex in deployment. Further, it provides a different trust model: the original ALTO model is a star trust model, with the application (e.g., Rucio/FTS) at the hub and each AS needs to trust the application. The BGP extension model requires the trust of peers and recursive peers (BGP community may be used to impose policies).

3.2. Candidate Solution: Data-Path Sampling/Collection

This is a type of solution that allows data path to collect control plane information. For example, a traceroute based system called PerfSonar is widely deployed. Such a system can collect other network information such as delay and loss naturally as measurements. However, this type of solution typically cannot collect full topology information such as link capacity or handle more complex query such as ALTO Path Vector.

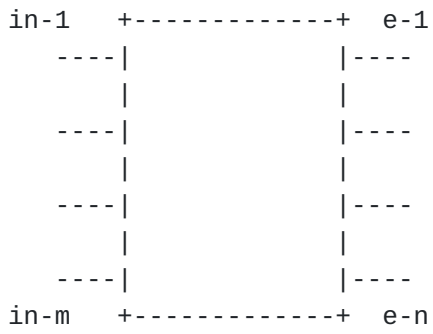
3.3. Candidate Solution: Multi-Domain ALTO Composition Refinement

This is an ALTO based system, and consists of two components: (1) it introduces a new abstraction of each autonomous network and associated query process to allow multi-domain ALTO information composition; and (2) it introduces a generic-path model at ALTO clients so that they can use the acquired information to gradually refine network information.

3.3.1. ALTO Server Multi-Domain Information Model

In the ALTO base model, a network is a container, with endpoints attached to the big switch. In the multi-domain model, each network (represented by an ALTO server) has a set of ingress points (in-1 to

in-m) and a set of egress points (e-1 to e-n). An endpoint belonging to the network will be attached to an ingress point and an egress point. Hence, a single-domain ALTO query will specify ingress and egress directly attached to an ingress point and an egress point. A source network, to a destination that is not in the same network, however, will only return the egress point; a destination network, when the source is from a different network, will need an ingress point. A general transit network will need an ingress point and return egress point. For consistency, the egress point must be a valid ingress point, represented by a unique address, of the peer.



ALTO Server Multi-domain Query Model: Each ECS query, if the src is not in the home domain of the ALTO server, should include an ingress point, where the ingress point is returned by the ALTO server of the previous domain. If the domain of the ALTO server is not the home domain of the destination, the ALTO server should return the egress point of the home domain and the ingress of the network domain.

As an optional feature, the query should allow indication of iterative or recursive queries.

To support incremental deployment, an ALTO server may respond to a query without specifying an ingress point and the source is not in the domain of the ALTO server. In this case, the ALTO server will return the results from each potential ingress points. For each ingress point indicated, the server indicates information of the previous hop (e.g., peer AS number and potential address).

3.3.2. ALTO Client General-Path Model

In particular, it allows the path from a src to a dst to be a directed acyclic graph, with the following components:

A set of nodes, where each node has both a type, and attributes, where the type can be (1) host: such as src/dst, with attributes such as IP address; (2) AS: which is a group of nodes, i.e., subgraph, with attributes including ASN; (3) router, with subtypes such as BGP-router, with attributes such as IP address.

A set of links, where each link has a head and a tail; hence the types of links will be the unique combinations of head-type x tail type. A link can have its attributes as well.

Now, some examples of this representation in our deployment use case:

For the geo-distance ALTO cost derived from geo-ip: the src is a host and the dst is also a host, and the metric is the geo distance;

For CERN looking glass ALTO server, from a src host in CERN to a dst host in another network, say KIT, the src is a host, with two links, one for each of the two looking glass BGP routers from cern; each of these BGP routers links to its BGP peer, and each such BGP peer links to the next AS, in the AS-PATH exposed by CERN.

4. IANA Considerations

Some of the solutions will need IANA registrations.

5. Acknowledgments

The authors of this document would also like to thank many for the reviews and comments.

6. References

6.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/info/rfc2119>.

[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>.

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <https://www.rfc-editor.org/info/rfc8174>.

6.2. Informative References

[RFC7971] Stiemerling, M., Kiesel, S., Scharf, M., Seidel, H., and S. Previdi, "Application-Layer Traffic Optimization (ALTO) Deployment Considerations", RFC 7971, DOI

10.17487/RFC7971, October 2016, <https://www.rfc-editor.org/info/rfc7971>.

Authors' Addresses

Y. Richard Yang
Yale University
51 Prospect St
New Haven, CT 06520
United States of America

Email: yry@cs.yale.edu

Mario Lassnig
CERN
CH-1211 Geneva 23
Switzerland

Email: mario.lassnig@cern.ch