Fine-Grained, Multi-Domain Network Resource Abstraction as a Fundamental Primitive to Enable High-Performance, Collaborative Data Sciences

Qiao Xiang
Tongji/Yale

J. Jensen Zhang
Tongji

X. Tony Wang
Tongji

Y. Jace Liu
Tongji

Chin Guok
LBNL

Franck Le
IBM

John MacAuley
LBNL

Harvey Newman
Caltech

Y. Richard Yang
Tongji/Yale

ABSTRACT

Recently, a number of multi-domain network resource information and reservation systems have been developed and deployed, driven by the demand and substantial benefits of providing predictable network resources. A major lacking of such systems, however, is that they are based on coarse-grained or localized information, resulting in substantial inefficiencies. In this paper, we present Explorer, a simple, novel, highly efficient multi-domain network resource discovery system to provide fine-grained, global network resource information, to support high-performance, collaborative data sciences. The core component of Explorer is the use of linear inequalities, referred to as resource state abstraction (ReSA), as a compact, unifying representation of multi-domain network available bandwidth, which simplifies applications without exposing network details. We develop a ReSA obfuscating protocol and a proactive full-mesh ReSA discovery mechanism to ensure the privacy-preserving and scalability of Explorer. We fully implement Explorer and demonstrate its efficiency and efficacy through extensive experiments using real network topologies and traces.

CCS CONCEPTS

• **Networks** → **Network protocol design**; **Signaling protocols**; **Network privacy and anonymity**; *Network resources allocation*;

KEYWORDS

Interdomain, Collaborative Data Sciences, Resource Discovery, Resource State Abstraction

SIGCOMM '18, August 20–25, 2018, Budapest, Hungary

ACM ISBN 978-1-4503-5915-3/18/08…$15.00
https://doi.org/10.1145/3234200.3234208

ACM Reference Format:
Qiao Xiang, J. Jensen Zhang, X. Tony Wang, Y. Jace Liu, Chin Guok, Franck Le, John MacAuley, Harvey Newman, and Y. Richard Yang. 2018. Fine-Grained, Multi-Domain Network Resource Abstraction as a Fundamental Primitive to Enable High-Performance, Collaborative Data Sciences. In *SIGCOMM Posters and Demos '18: ACM SIGCOMM 2018 Conference Posters and Demos, August 20–25, 2018, Budapest, Hungary.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3234200.3234208

1 INTRODUCTION

Many emerging large-scale data science projects (*e.g.*, the Large Hadron Collider experiment [1]), are based on a collaborative, distributed-networks design, where massive datasets have to be moved from storage facilities to large computing clusters distributed at multiple autonomous member networks, and analyzed by multi-stage distributed systems. Such large, correlated and parallel flows (*e.g.*, petabytes of data) have evolved to dominate science networks' traffic. To ensure the completion of those transfers within the application time constraints, users require the ability to reserve and guarantee bandwidth across networks. As such, a number of on-demand circuits reservation systems have been developed and deployed (*e.g.*, the OSCARS system [2] in LHC).

However, such existing systems are based on localized design and hence can suffer poor performance for correlated and concurrent flows across multiple ASes. For privacy reasons, existing multi-domain reservation systems treat each AS as a black box. They probe their available resource by submitting varied circuit reservation requests, and receive Boolean responses. In other words, current solutions perform a depth-first search on all ASes, and rely on a trial and error approach: to reserve bandwidth, repeated, and varied attempts may have to be submitted until success. In addition to requiring a large number of search attempts, this approach may obtain a bandwidth allocation that is far from optimal (*e.g.*, max-min fairness).

This paper presents Explorer, a system designed to optimize large, multi-domain transfers, and address the limitations of current reservation systems through three main components. The first and core component of Explorer is

the use of linear inequalities, referred to as *resource state abstraction* (ReSA), as a compact, unifying representation of multi-domain network available bandwidth. Second, Explorer introduces a ReSA obfuscating protocol to ensure that ASes and other external parties cannot associate an inequality with a corresponding AS. Finally, Explorer includes a proactive full-mesh ReSA discovery component for scalability purposes. This component improves the latency of resource discovery via AS-level ReSA pre-computation and projection. We implement Explorer. Extensive experiments using real network topologies and traces show that Explorer (1) efficiently discovers available networking resources in collaborative networks on average 2 orders of magnitude faster, and allows fairer allocations of network resources, (2) preserves the private information of ASes with little overhead; and (3) scales to a collaborative network with 200 ASes.

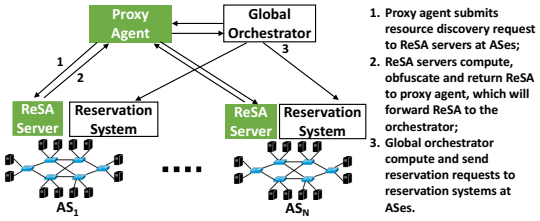


Figure 1: The architecture and workflow of Explorer.

2 EXPLORER OVERVIEW

Architecture: Explorer introduces a resource discovery proxy agent, and a ReSA server in each AS (Figure 1). The resource discovery proxy agent is the main interface for the orchestrator to submit the resource discovery requests and provides a unified view of the resources across different the ASes. The proxy agent has BGP sessions to all participating ASes, and given a request for a set of circuits, can thus infer the AS paths for each circuit in the request. It also has connections to the ReSA servers in each AS, which upon receiving requests provided from the proxy agent, compute the ASes' ReSA abstractions.

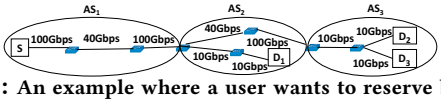
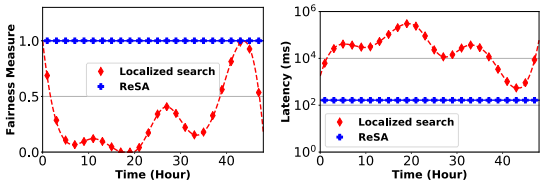


Figure 2: An example where a user wants to reserve bandwidth for three source-destination pairs: (S, D_1), (S, D_2) and (S, D_3).

Resource state abstraction (ReSA): ReSA is a unifying representation that captures the properties (e.g., available bandwidth) of resources shared – within and between ASes – by a set of requested circuits. It relies on linear inequalities to express the available bandwidth of shared resources for a set of requested circuits to be reserved, and can also support more complex traffic engineering policies (*e.g.*, multi-path routing and multicast), with the use of auxiliary variables and additional inequalities. As an example, consider a collaboration network composed of three ASes, where a user wants to reserve bandwidth for three circuits, from source host S to destination hosts D_1, D_2 and D_3 (Figure 2). The ReSA abstraction captures all constraints from all networks using linear inequalities as follows:



(a) Max-min fairness of resource utilization. **(b) Resource discovery latency.**

Figure 3: Comparison of performance between Explorer and localized search (*e.g.*, OSCARS).

$$\begin{array}{ccc} \text{AS}_1: & \text{AS}_2: & \text{AS}_3: \\ x_1 + x_2 + x_3 \le 100, & x_2 + x_3 \le 40,\ x_1 \le 10, & x_2 + x_3 \le 10, \\ x_1 + x_2 + x_3 \le 40, & x_2 + x_3 \le 100,\ x_1 \le 10, & x_2 \le 10, \\ x_1 + x_2 + x_3 \le 100, & & x_3 \le 10, \end{array}$$

where x_1, x_2 and x_3 each represents the available bandwidth that can be reserved for each circuit. For example, $x_1 + x_2 + x_3 \le 100$ means that three circuits share a common resource and the sum of their bandwidths cannot exceed 100 Gbps.

ReSA obfuscating protocol: The key idea of this protocol is to have each AS obfuscate its own set of linear inequalities as a set of linear equations through a private random matrix and a couple of random matrices shared with few other ASes, and send the obfuscated equations to the proxy agent. In this way, the proxy agent can reconstruct the original bandwidth feasible region for the circuits across the ASes, but cannot associate any linear inequality with its corresponding AS.

Proactive full-mesh ReSA discovery: The main idea of this component consists in having the proxy agent periodically query ReSA servers to discover inter-domain ReSA between every pair of source and destination ASes. As such, when a user submits a resource discovery request, the proxy agent does not need to send any query to the ReSA servers. Instead, using the AS-level ReSA information, the proxy agent can immediately perform projection operations to get the ReSA for the request. This mechanism substantially improves the scalability of Explorer.

3 PERFORMANCE EVALUATION

We implement Explorer and evaluate its performance on the topology from LHCONE, a science network with 78 ASes. We replay an actual trace from the CMS experiment [3]. We compare the performance of Explorer with that of existing resource reservation systems (*e.g.*, OSCARS), which we refer as *localized search*. Figure 3(a) shows that Explorer can always compute the optimal max-min fairness allocation, while that of the localized search based solution is 0.37 on average and even drop to 0 at times. Figure 3(b) shows that Explorer can reduce the total time to discover network resources by 2 orders of magnitude on average.

ACKNOWLEDGMENTS

This research is supported in part by NSFC grants #61672385, #61472213 and #61702373; China Postdoctoral Science Foundation #2017-M611618; NSF grant CC-IIE #1440745; NSF award #1246133, ANSE; NSF award #1341024, CHOPIN; NSF award #1120138, US CMS Tier2; NSF award #1659403, SANDIE; DOE award #DE-AC02-07CH11359, SENSE; DOE/ASCR project #000219898; Google Research Award, and the U.S. Army Research Laboratory and the U.K. Ministry of Defence under Agreement Number W911NF-16-3-0001.

REFERENCES

[1] The large hadron collider. https://home.cern/topics/large-hadron-collider.
[2] Oscars: On-demand secure circuits and advance reservation system. https://www.es.net/engineering-services/oscars/.
[3] CMS Task Monitoring. http://dashb-cms-job.cern.ch/.