An inter-domain multi-path flow transfer mechanism based on SDN and multi-domain collaboration

LU You, LI Wei, LUO Junzhou, JIANG Jian, XIA Nu
School of Computer Science and Engineering
Southeast University
Nanjing, P.R. China
luyou,xchlw,jluo,jiangjian,xia_nu@seu.edu.cn

***Abstract*—Increasing numbers of service providers have tended to use tens of geographically dispersed datacenters in recent years. Thus, a major unmet challenge is efficiently transferring data among multiple datacenters in different domains. Traditional single-path transfer mechanism based on BGP has limited reliability and low link utilization; therefore, multi-path inter-domain flow transfer mechanisms have become a research focus. However, the existing multi-path transfer mechanisms have some shortcomings, such as poor compatibility with existing network architecture and difficulty in selecting routes according to the demands of applications and the network status. This paper proposes a novel inter-domain multi-path flow transfer mechanism based on SDN and multi-domain cooperation. First, a hieratical iterative detection method is proposed to find diversified multi-paths based on the analysis of BGP notification and inter-domain collaboration. Second, an information exchange method is devised to exchange and maintain the network status (e.g., inter-domain topology updates, link load). Finally, a decision and deployment method is designed. The experimental results indicate that this mechanism has advantages in ensuring the success rate of data transfer task and improving network throughput.**

Keywords- SDN; multi-path flow transfer; inter-domain;

I. INTRODUCTION

With the rapid growth of cloud and big data services, major service providers now use tens of geographically dispersed datacenters. An unmet challenge in leveraging these datacenters is efficiently transferring data among different domains with relatively strict requirements (e.g., throughput, path reliability). To address these demands, existing BGP-based transfer mechanisms can only utilize one path (although Internet network topology has high inter-domain connectivity[1]). Thus, they have low link utilization, poor reliability and relatively weak load balancing support[2].

To mediate these drawbacks, many studies have been conducted on inter-domain multi-path transfer mechanisms [3-4]. Some studies have proposed the design of new inter-domain route protocols, such as NIRA[2,4] and feedback based routing[2,5]. Some studies have aimed to improve the BGP protocol to support multi-path route, such as the BGP protocol with ADD_PATH function[6] and the R-BGP mechanism, to provide additional alternate paths[2,7]. Moreover, some researchers have proposed multi-path solutions based on BGP notification without changing the existing BGP protocol, such as Cisco's multi-path selection mechanism based on BGP [8], a route deflection mechanism based on labels[9], and the multipath BGP (MBGP) mechanism based on active multi-path detection[10]. However, these mechanisms have some significant deficiencies: (1) the routing decision is made by the router, so the capability is limited; (2) the overhead of most mechanisms is high, and some have problems such as convergence and loopback; and (3) the network status or transfer task requirements have not been considered in the process of path selection.

To overcome the aforementioned deficiencies, this paper proposes a novel inter-domain multi-path flow transfer mechanism based on SDN and multi-domain collaboration. First, we design a routing detection method based on hierarchical iteration. On the one hand, we take advantage of the SDN controller's strong and fine-grained computing abilities to improve the detection capability (the source domain controller analyzes BGP notifications to find all possible paths). On the other hand, we take advantage of inter-domain collaboration to relieve the overhead (the source domain controller sends hierarchical iteration requests to collaborative domains to find more accessible paths according to the adjacent sequence). Thus, we can build diverse multi-path routes, avoid loopbacks and ensure convergence. Second, we design the inter-domain network status information exchange method, by which collaborative domains can measure and exchange network status information. Finally, we design the decision and deployment method: the transfer paths and flow distribution proportion of each path should be calculated according to the network status and application requirements, and all transfer policies should be deployed to involved domains. The experimental results indicate that our method has advantages in ensuring the success rate of transfer tasks and improving the network throughput.

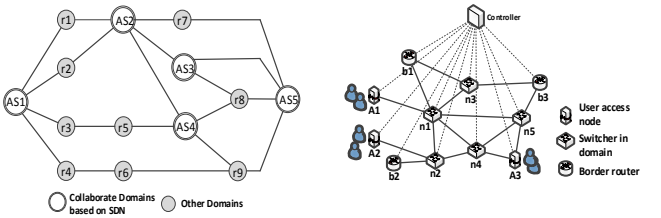
The remainder of the paper is organized as follows. The relevant concepts and model are presented in Section II. In section III, we introduce our mechanism, including multi-path route detection, inter-domain information exchanging, and selection and deployment of the flow transfer policies. In section IV, we introduce the experiment and the analysis of results, and in section V, we draw our conclusion and describe our related future work.

II. FLOW TRANSFER MODELS

A. Description of the problem and relevant concepts

Fig. 1(a) describes the network environment of inter-domain flow transfer problem in this paper. $AS1$- $AS5$ are datacenter domains of the same owner based on SDN network; in addition, $r1$ - $r7$ are normal domains. Assume that the

source of the flow transfer is $AS1$, and the destination is $AS5$, so the problem is as follows: search for all possible multi-path routes from $AS1$ to $AS5$, select some appropriate paths from them, and calculate the flow distributions of each path based on task requirements and network statuses. Figure 1b describes the SDN-based network environment of $AS1$. $A1$-$A3$ are user access nodes; $n1$-$n5$ are intra-domain switching nodes, and $b1$-$b3$ are border routers. The controller is the domain's control center.



(a)inter-domain network (b) SDN based domain network

Fig. 1 Description of the problem

To better describe our mechanism, the following basic concepts are defined.

Definition 1: collaborative domain, **BGP domain**. All domains of geographically dispersed datacenters are collaborative domains. They are based on SDN and have the same owner. The remaining domains are normal BGP domains.

Definition 2: adjacency and **adjacent path segments**. If domain AS_s receives BGP notification $p:\{r1, r2, ..., AS_t\}$, where $r1, r2, ...$are the BGP domains, then AS_s and AS_t are the adjacencies through path p, and p is the adjacent path segment of AS_s to AS_t . In this paper, vectors $< label, \{AS_s, r1, r2, ..., AS_t\} >$ are used to represent p, where $label$ is a unique name for multi-path route detection, information exchange and path selection.

B. Multi-path flow transfer model

To solve the inter-domain flow transfer problem, our mechanism takes advantage of the SDN controller's powerful abilities including decision-making, storage and network views, as well as the collaboration of other domains. In every collaborative domain, the controller should collect and analyze all received BGP notifications, ask other collaborative domains hierarchically and iteratively to detect different accessible paths, and construct the multi-path route. Moreover, they should also measure and exchange network status information such as updating the topology, bandwidth, load, and other information. Finally, when the transfer task arrives, the controller should select appropriate paths from the multi-path route, calculate the flow distribution according to the application requirements and network status, and deploy the policies to the involved collaborative domain. Therefore, this paper proposes an inter-domain multi-path flow transfer model including 4 modules, and its structure is shown in Figure 2.

BGP information collection module: collects all BGP notifications received by the domain's border routers.

Multi-path route detection module: based on the BGP notifications, this module constructs the multi-path route.

Network information exchange module: measures and exchanges network status information with the other collaborative domains.

Transfer decision and deployment module: makes and deploys the transfer decisions.

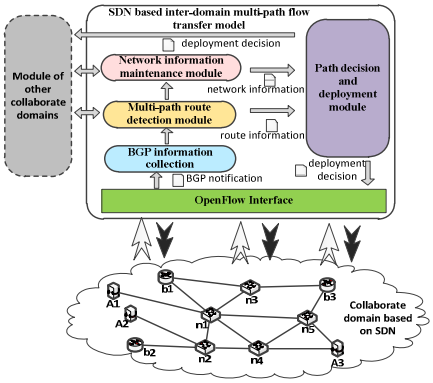


Fig. 2 Multi-path flow transfer model

III. MECHANISM OF MULTI-PATH FLOW TRANSFER

A. Multi-path route detection

The primary concept of our multi-path route detection is that path detection and data transmission are independent, and detection should be performed beforehand. Then, in the detection process, the source domain should ask other collaborative domains to detect further paths. Moreover, this iterative algorithm should ensure convergence and avoid loopback. According to these concepts, detection methods can be divided into the following four steps.

(1) **BGP notification collection**: The source domain controller collects BGP notifications from all border routers.

(2) **Multi-path iterative detection of source domain**: First, the source domain searches all paths accessible to the destination according to BGP notifications. Second, the source domain controller adds all paths including no collaborative domain (the source domain and destination domain are the adjacencies through these paths) into the final results set. Finally, the source domain requests collaborative domains to detect more accessible results based on the remaining paths and repeat this process until all of the collaborative domains are completely handled. To avoid loopback, we design the **marked domain list mechanism** in the iterative process. The request process includes the following steps: (a) the source domain obtains all adjacent cooperative domains according to the remaining paths, adds them and itself as first-level **marked domain**s to the **marked domain list**, and meanwhile saves these adjacent path segments for the multi-path construction process; (b) the source domain requests each adjacent cooperative domain with the **marked domain list**; (c) the collaborative domain receives the request messages, analyzes their BGP notifications and returns results (new adjacent cooperative domains and adjacent path segments, except the domains or segments related to any domain **in the marked domain list)**; (d) the source domain receives all returned results and adds the new collaborative domain to the

marked domain list as a new next-level marked domain. Steps (a) to (c) are repeated until the returned results do not include new collaborative domains. Moreover, to ensure convergence, we set the upper count of the iterations beforehand. The algorithm is described below.

Algorithm 1 Iterative detection algorithm of the source domain.	
input	BGP notification list $list$, AS_S,AS_T, $List\_T$, list of domains on the next level $List\_next$, upper count of iterations Max
output	The adjacent path segments set $Result$ from AS_S toAS_T, temporary path result set $Result\_temp$
1	Initialization: $Result\_temp$ = null, $List\_next$=null, $length$=0
2	***For*** each notification $x \in list$
3	***If*** x only contains AS_T Generate $label$, build path vector r=$< label, x >$
4	$Result$=$Result$ + r
5	***Else*** intercept path x' (from AS_S to the closest collaborative domain AS_temp in x), form $label'$ and r=$< label', x >$
6	$Result\_temp$=$Result\_temp$ +x' $List\_next$= $List\_next$ + AS_temp $List\_T$=$List\_T$+ AS_temp
8	***While*** $List\_next \neq null$ and $length < Max$
9	***For*** each domain $as \in List\_next$ Send request to as; as invokes cooperative detection algorithm (presented later) with parametersAS_S, AS_T, $List\_T$ AS_S and obtains the results: $Result'\_temp$ of the returned temporary path and new cooperative domain $List'\_next$ make new marked domain list $List'\_T$
10	***If*** $List\_next$ is not null $Result\_temp$=$Result\_temp \cup Result'\_temp$ $List\_T$ = $List\_T \cup List'\_T$ $List'\_next$= $List'\_next$+as
11	$length$++
12	***Return*** $Result$, $Result\_temp$

(3) **Coordinated detection of collaborative domains**. The detection process of collaborative domains is as follows: (a) analyze BGP notifications (but do not include any domain in the marked domain list); (b) return all adjacency domains (with corresponding adjacent path segments); or (c) return null if unable to find any adjacency domains.

(4) **Construction of multi-path route for the source domain**. After the iteration, the source domain gathers all adjacent path segments in the final result set to form a multi-path route. The algorithm is described below.

Algorithm 2. Algorithm of multi-path route construction	
input	Adjacent path segments set $Result\_temp$,adjacent path set $Result$
output	Multi-path route G_{AS_S,AS_T} from AS_S toAS_T
1	Initialization, $Temp\_list$=null, G_{AS_S,AS_T}=null
2	Take out all of the path segments starting from AS_S in $Result\_temp$, constitute a forked tree, and add into $Temp\_list$
3	while $Result\_temp$ is not null
4	For each leaf $route\_t \in Temp\_list$
5	Take out all of the path segments starting from $route\_t$ in $Result\_temp$ and join up node $route\_t$
6	add all paths in $Result$ to the forked tree $Temp\_list$
7	Return $Temp\_list$ as G_{AS_S,AS_T}

B. Inter-domain network information exchange

The constantly changing status (such as delay and available bandwidth) of links and domains are very important for transfer decisions, and this information cannot be obtained by the source domain itself. Thus, we design the information exchange method based on the following concepts: (a) each collaborative domain measures the network status of itself and its adjacent path segment's link; (b) each domain sends its information to all other collaborative domains periodically; (c) each domain receives, updates and stores the network status for itself. The network status information of the path segments and domain can be set as vectors as follows:

$$I_{link} : \{lable, state, QVal_1, QVal_2, ...\} \quad (1)$$

$$I_{AS} : \{< AS >, QVal_1, QVal_2, ...\} \quad (2)$$

In the vector of link status and domain, the value of $state$ is one of $\{live, conget, break, new\}$, where $break$ means link interruption and new means finding new paths. $QVal_1, QVal_2, ...$ are QoS performance indicators.

C. Selection and deployment of the flow transfer path

After multi-path detection and information exchanging, the following next problems is how to select appropriate paths according to the application requirements and network status, and how to distribute flow into different paths to ensure balance. The decision process of this decision is as follows: (a) the source domain controller selects transfer paths based on task's requirements and network status of related link or domain; (b) the controller calculates the flow distribution of different paths using a minimum overhead maximum flow algorithm; (c) the source domain sends notification of transfer policies to corresponding collaborative domains and ensures that policies are deployed successfully. According to the SDN network architecture, we use the SDN's flowtable item to construct the deployment notification.

IV. EXPERIMENTS AND ANALYSIS OF RESULTS

A. Environment of experiments

Because it is difficult to evaluate our mechanism in the Internet environment, we built an experimental environment to simulate Internet topology according to CAIDA project's AS connection topology data (released in March 2014) in a campus network. Its topology is shown in Fig. 3.

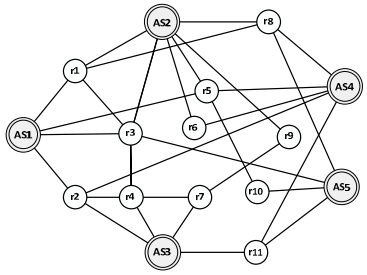


Fig. 3 inter-domain network topology of experiment

In our experiment, five nodes ($AS1$-$AS5$, simulated by five separate subnets based on SDN architecture) are randomly selected from the topology as collaborative domains. Other nodes are regarded as BGP domains ($r1$-$r12$). Our

transfer mechanism is deployed to the collaborative domains' controller. The link capacities are set randomly according to the normal distribution (mean value 1GB, variance 0.5).

B. Experimental design and analysis of results

We compare our mechanism with three state-of-the-art methods regarding the following aspects: success rate, network throughput and link utilization. The state-of-the-art methods are as follows: method 1 uses the MBGP-based multi-path transfer mechanism; method 2 uses a mechanism based on Cisco (using two different paths); and method 3 uses a single-path BGP based mechanism.

Experiment 1: To evaluate the success rate of transfer tasks. Each collaborative domain sets a certain number of transfer tasks (bandwidth demands set as 512 MB, transfer rate as 512 MB/s). The size of the flow was randomly set according to the normal distribution (mean value is 1 GB, variance is 1). The upper limit of the task delay is 5 s. The experimental process is as follows: there are two experimental phase and we simulate a lighter network load in the first phase (1-50 time intervals, every interval is 10 s). Within each interval, each collaborative domain generates transfer tasks according to the Poisson distribution (λ = 2). In the second phase (51-200 intervals), we simulate a heavy network load, so the Poisson distribution uses λ= 6. The results are shown in Table 1.

Table 1 Success rate results

method	success rate	
	first phase	second phase
our method	99.6%	91.2%
MBGP	99.7%	75.8%
two-path	99.8%	85.8%
single-path	96.4%	58.5%

Table 2 Throughput results

method	throughput	
	first phase	second phase
our method	2.57 GB	5.64 GB
MBGP	2.53 GB	5.03 GB
two-path	2.49 GB	4.27 GB
single-path	2.45 GB	2.86 GB

In the first phase, all methods can accomplish tasks relatively smoothly, and the success rate is above 95%. However, upon entering the second stage, the single-path-based method 3 encounters frequent congestion, and there is a substantial decline (below 60%); the success rates of methods 1 and 2, which utilize more paths, are higher, but method 1 uses dispatch flows on average and is affected by the load imbalance, so its success rate (below 80%) is lower than the success rate of method 2 (higher than 80%), which uses two different paths to avoid imbalance. However, our method could make use of all possible paths and dispatch flows according to the link's load, so its success rate is the highest.

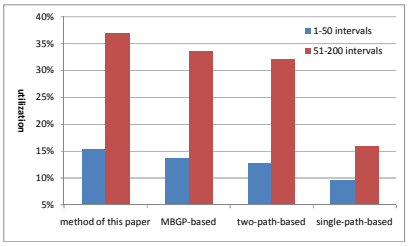


Fig. 4 Link utilization results

Experiment 2 verifies the link utilization and network throughput. The configuration is the same as in experiment 1. We record the link utilization and successful transfer task throughputs of the overall network. The experimental results are shown in Fig. 4 and Table 2. In the first stage, the utilization and throughputs are close. In the second stage, the single-path-based method 3 has low link utilization and throughputs. Method 2 can take advantage of two different paths, so its link utilization and network throughputs are higher than in method 3 but lower than in method 1, which can use more paths but ignores balance. Finally, our method is able to use more possible paths and considers the balance, so it achieves the highest network throughput and link utilization

V. CONCLUSION

How to efficiently transfer data among different domains is an urgent issue. In this paper, we have designed a novel inter-domain multi-path flow transfer mechanism based on SDN and multi-domain cooperation. The experimental results show that our mechanism can ensure the transmission success rate and increase network throughput and resource utilization. Our future work may introduce a traffic prediction method to improve the effectiveness of inter-domain flow transfer.

Acknowledgment

This work is supported by National Key Basic Research Program of China under Grants No. 2010CB328104,National Natural Science Foundation of China under Grants No. 61320106007, National High-tech R&D Program of China (863 Program) under Grants No. 2013AA013503, China Specialized Research Fund for the Doctoral Program of Higher Education under Grants No. 20110092130002, Prospective Research Project on Future Networks of Jiangsu Future Networks Innovation Institute under Grants BY2013095-2-07, Jiangsu Provincial Key Laboratory of Network and Information Security under Grants No. BM2003201, and Key Laboratory of Computer Network and Information Integration of Ministry of Education of China under Grants No. 93K-9.

References

[1] CIDR Report, October 2013: http://www.cidr-report.org/

[2] Su J S, Dai B, Liu Y J, et al. Inter-Domain Multipath Routing Protocols[J]. Ruanjian Xuebao/Journal of Software, 2012, 23(1): 65-81.

[3] He JY, Rexford J. Toward Internet-wide multipath routing. IEEE Network Magazine, 2008,22(2):16-21.

[4] Khoury J S, Abdallah C T. A Survey of Novel Internetwork (and Naming) Architectures[M], Internet Naming and Discovery. Springer London, 2013: 13-33.

[5] Hasegawa T. A Survey of the Research on Future Internet and Network Architectures[J]. IEICE transactions on communications, 2013, 96(6): 1385-1401.

[6] Walton D, Retana A, Chen E, Scudder J. Advertisement of multiple paths in BGP. Internet Draft, 2009.

[7] Liu K. Multipath Routing and Load Sharing Using Game Theory[D]. George Mason University, 2013.

[8] Cisco Inc. BGP best path selection algorithm. 2006. http://www.cisco.com/image/gif/paws/13753/25.pdf

[9] Yang XW, Wetherall D. Source selectable path diversity via routing deflections. In: Proc. of the 2006 Conf. on Applications, Technologies, Architectures, and Protocols for Computer Communications. New York: ACM Press, 2006. 159-170.

[10] Fujinoki H. Multi-Path BGP (MBGP): A solution for improving network bandwidth utilization and defense against link failures in inter-domain routing. In: Proc. of the IEEE Int'l Conf. on Networks. 2008. 1-6.