Bandwidth Constrained Distributed Inter-domain Path Selection (DIPS)

Rui Meng, Bin Da, Zhe Chen, Chuang Wang, Sheng Jiang
NGIP, Beijing Huawei Digital Technologies Co., Ltd.
No.156 Beiqing Street, Haidian District, Beijing, P.R.China, 100095
Email: {mengrui, dabin, chenzhe17, wangchuang, jiangsheng}@huawei.com

***Abstract*—In this paper, bandwidth constrained path selection technologies are reviewed, which may utilize distributed routing protocols with appropriate extensions or centralized mechanisms based on Software Defined Network (SDN) to calculate end-to-end bandwidth constrained paths over Internet without using links with depleted bandwidth. In these traditional schemes, the resource reservation may sometimes fail due to inadequate bandwidth over running network links. Because of the requirements of scalability and privacy for routing policies, SDN-based technologies, such as B4, KANDOO and Control eXchange Point (CXP), are not perfect to solve this problem faced by path selection. In addition, distributed routing protocols with extensions, e.g., OSPF-TE, can only help calculate paths inside one particular autonomous system, which cannot satisfy the need of end-to-end path selection over whole internet. Based on these observations, the Distributed Inter-domain Path Selection (DIPS) scheme is then proposed in this paper. In which, all links without enough bandwidth for allocation are excluded from the network resource reservation pool. This DIPS scheme can generate bandwidth constrained end-to-end paths, which serve as a distributed solution for inter-domain path calculation, considering centralized or distributed intra-domain technologies. Finally, the proposed scheme is compared with existing solutions on bandwidth utilization and computational complexity.**

***Index Terms*—path selection, resource reservation, routing information base, OSPF, BGP.**

I. INTRODUCTION

End-to-end Quality of Service (QoS) guarantee is normally the pre-requisite of many network services, such as high-definition end-to-end real-time video streaming and tele-medical applications [1]. The path selection is one main enabler to guarantee end-to-end connectivity under variety of QoS requirements. Both distributed and centralized path selection methods have been designed in the literature.

Distributed Path Selection (DPS) is a method, in which all nodes cooperatively, or edge nodes individually, compute constrained end-to-end paths, based on available bandwidth information of the whole network, via synchronization among all network elements, in a distributed manner.

For Intra-AS (Autonomous System) paths computation, as studied in [2], TEDB (Traffic Engineer Database) is generated through synchronizing bandwidth status within one specified area, by using Type 10 Link State Advertisement (LSA). Computing component of MPLS-TE can calculate intra-area paths for access requests by running CSPF (Constrained Shortest Path First) algorithm based on up-to-date TEDB. However, such MPLS-TE is not capable of handling inter-domain path selection. In addition, in the BGP-TE scheme [3], there exists one special attribute being used to carry TE information. However, this attribute is merely for VPN reachability [4], which cannot be used for end-to-end paths with QoS guarantee.

Centralized Path Selection (CPS) uses SDN controller to maintain the overall topology of whole network, along with link status. This information is then utilized by the controller to calculate constrained end-to-end paths.

As embodied in [5], the B4 solution is devised for a private WAN, connecting Google's data centers worldwide which is not involved in multi-ISP interworking. Path computation for massive concurrent accessing flows is avoided by B4 through flow aggregation, which could still be a challenge for general purposed solution.

In the scheme of CXPs, controllers are set in Internet eXchange Points (IXPs) for computing paths among ISPs through IXP but direct and customer-provider links between ISPs are not considered, thus it cannot provide end-to-end connectivity across the whole Internet. Kandoo solution has been proposed in [6], which is a hierarchical SDN controller scheme that cannot help control applications requiring network-wide status, because they believe such applications are intrinsically hard to scale.

Furthermore, HyperFlow [7] is known as a distributed SDN control plane that can handle a few thousand events per second, but any processing beyond is considered out of scalability limitation. In Onix [8], a distributed control plane is proposed, in combination of a distributed file system and a distributed hash table under logical centralization. However, the obvious limitation of such schemes is to request a consistent network-wide view over all controllers, which will significantly reduce the scalability of the overall system.

So far, there is no solution that can select bandwidth constrained end-to-end paths, considering internet-wide links, either in a centralized way or in a distributed manner. Due to the issue of privacy for inter-domain routing policies [9], an internet-scale SDN solution is not available. Thus, Distributed Inter-domain Path Selection (DIPS) scheme is proposed here.

The rest is as follows: Section II describes the details of the proposed DIPS scheme. Then, one simple embodiment with potential applications is described in Section III. Then, the analysis is provided in Section IV. Finally, Section V concludes this paper.

978-3-903176-14-0 © 2018 IFIP

II. Proposed DIPS Scheme

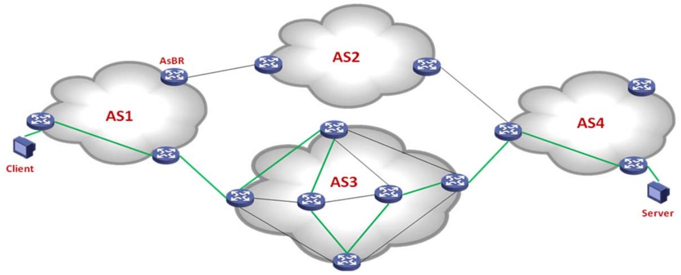

Fig. 1. A Typical Interconnected Network System.

In this section, the proposed DIPS scheme is described in detail, which generally consists of using three forwarding planes. One simplified network example is illustrated in Fig.1.

Particularly, the first plane is known as Best-Effort Forwarding Plane (BEFP), which is a traditional forwarding plane without any modification.

The second plane is newly introduced, which is named as Signaling Forwarding Plane (SFP). Its function is mainly designed to forward signaling packets for enabling resource reservation (e.g., Path message in RSVP [10]) according to specific network requirements. Note that this SFP is constructed based on a subset of overall network topology that removes links with inadequate bandwidth. The removal policy is based on the amount of reserved resource, and is irrelevant to real-time link traffics. The implementation of SFP's link removal and restoration should be manipulated by a pair of thresholds (e.g., percentage of total bandwidth, absolute empirical value of bandwidth). The detailed setup of these thresholds will be discussed in the following analysis section. In addition, SFP should also maintain an independent routing table, which is parallel with the traditional routing table and is further denoted as SFP-RIB (Routing Information Base) hereafter. Since the only purpose of SFP-RIB is used to forward resource reservation signaling, it then can use relatively cheaper software address-lookup instead of expensive dedicated hardware scheme.

Another new plane called Bandwidth Guaranteed Forwarding Plane (BGFP) is also constructed to assume the responsibility of forwarding packets with dedicated purpose on bandwidth guarantee. Normally, critical application packets should be forwarded on BGFP along the paths established by signaling messages mentioned above. The scheduling method could be strict priority queue or any other available ways.

Based on the above three planes, the corresponding protocols should be customized accordingly, which are described as follows.

II.1. Intra-AS Path Selection

As shown in Fig.2, a typical Intra-AS example usually adopts OSPF, in which an AS could be divided into multi areas, therefore intra-AS path calculation is composed of intra-area part and inter-area part.

For the Intra-area part, e.g., OSPF-TE uses Type 10 LSAs that have an area flooding scope to synchronize bandwidth information within one area. Detailed bandwidth information of links is described by link TLV, in which, key information (like link type, maximum bandwidth and unreserved bandwidth) is listed.

For Intra-area part, e.g., MPLS-TE, as one existing solution, can calculate the intra-area paths by running CSPF algorithm based on TEDB gathered from OSPF-TE flooding.

In our proposal, OSPF-TE is used to maintain a new LSDB (Link State Database), denoted by SFP-LSDB that can be further used to formulate a new parallel RIB, given as SFP-IGP-RIB. Such SFP-LSDB may have exact same initiation as the traditional LSDB that is updated by a newly added link or a link failure, while this SFP-LSDB can be modified not only by the events mentioned above but also by the bandwidth usage of links.

Individual links in the system are configured with two thresholds as special attributes. Specifically, the first threshold is set to delete one particular link (e.g., with unreserved bandwidth lower than 1st-threshold), while the second threshold is for adding back one link (e.g., with unreserved bandwidth higher than 2nd-threshold) into the associated SFP-LSDB.

In our proposal, the intra-area routing entries should be computed by running standard OSPF algorithm based on SFP-LSDB within each area.

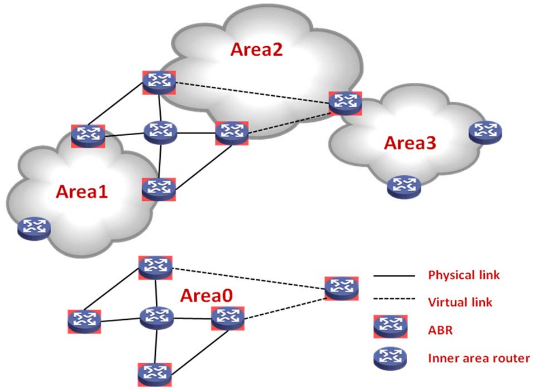

Fig. 2. Intra-and-Inter Area Network Connections.

For inter-area part, if using MPLS-TE, there usually have two solutions for paths selection: (*i*), Automatic way, due to inner-area flooding, TEDB of the whole AS cannot be gathered, the only way is to set one area in one AS at one time. (*ii*), Manual way, using explicit inter-area paths. However, we can infer that MPLS-TE cannot compute intra-AS end-to-end constrained paths in a fully distributed way.

Our scheme runs standard OSPF algorithm on SFP-LSDB, in which all links with depleted resources have already been excluded. However, virtual links in backbone area, which are not defined in OSPF-TE, cannot be maintained in SFP-LSDB. So the extension in OSPF-TE for virtual links is another contribution of our proposal. Note that, the full-mesh intra-AS tunnels can be created based on SFP-IGP-RIB for simplification of inter-AS paths calculation.

II.2. Inter-AS Path Selection

BGP is the de-facto inter-domain protocol, and its basic procedure is not modified in our scheme. We formulate the inter-domain part of the SFP plane by a simple extension of BGP in which a new Type 36 attribute is defined, which has a flag for showing inadequate bandwidth of the path along which one particular prefix is advertised.

II.2.1. IBGP (Internal BGP)

IBGP is used inside one AS. If an ASBR (AS Boarder Router) receives routing UPDATE messages from external peers, then these routing updates will be transmitted to all its inner-AS peers, the following actions should be performed by all these peers.

Firstly, when the flag in Type 36 attribute is 1, there will be no more processing with our solution.

Secondly, routing iteration should be carried out in the plane of SFP. The originators (ASBRs) of UPDATE messages are used as keys to search SFP-IGP-RIB, so as to establish or update the corresponding inter-AS routing entries. Provided that one lookup is failed, it shows that there does not have a reachable intra-AS path with enough bandwidth to the originator. Then, the binding between intra-AS routing entries and inter-domain routing entries should be appropriately setup on egress ASBRs, for serving further actions in our proposal.

Thirdly, routing iteration can be perform in another way by which mapping relationships between intra-AS full-meshed tunnels and inter-AS routing entries are set. If one lookup of intra-AS tunnel with a specified ASBR as destination is successful, the mapping between one particular tunnel with the targeted prefix is then configured. Otherwise, it means there is no intra-AS path with enough bandwidth to the ASBR. In such situation, the corresponding prefix will be marked with inadequate bandwidth, and the flag is thus set to be 1 for external routing announcement.

II.2.2. EBGP (External BGP)

In addition, for EBGP, the peer routers shall perform the following actions for received inter-domain update messages:

Firstly, check the Type 36 attribute. If its flag is 1, meaning at least one link in the path does not have enough bandwidth already, the prefixes associated with this received message should not be added into SFP-BGP-RIB. The normal BGP process is then run in a traditional manner. Secondly, if the value of flag is 0, then proceed to verify the adequacy of bandwidth of inter-AS link to the external peer, following similar way described for IBGP. When finding the bandwidth of the corresponding link is below the 1st threshold, meaning such link does not have enough bandwidth to allocate. The Type 36 attribute will be set to be 1 in updates sent to inter-AS peers, and the associated prefixes are not added into SFP-BGP-RIB. Otherwise, when the particular link has enough bandwidth, the corresponding SFP-BGP-RIB entry is created.

II.3. Link Status Update

When bandwidth utilization of a link within an AS changes, IGP convergence will be triggered to form a new SFP-IGP-RIB or new full-meshed intra-AS tunnels. On the other hand, the IBGP iteration should also be performed periodically, so as to renew inter-AS RIB. If bandwidth utilization of one inter-AS link changes, the downstream ASBR may re-advertise the received prefixes to its IBGP peers. Thus, we extend two flags in BGP-RIB entries, one is for upstream resource, the value of which is copied from Type 36 attribute in the associated UPDATE message. The other one is for local resource that indicates the existence of an intra-AS path with enough bandwidth for intra-AS routers or an inter-AS link with enough bandwidth for ASBRs. The IBGP iteration is implemented periodically. If intra-AS resource status of a prefix changes after iteration, from inadequate or enough, its upstream resource is enough, the prefix is advertised downstream again. In addition, if an inter-AS link resource status changes, the relevant prefixes should be re-advertised if upstream resource is enough.

III. Embodiment and Applications

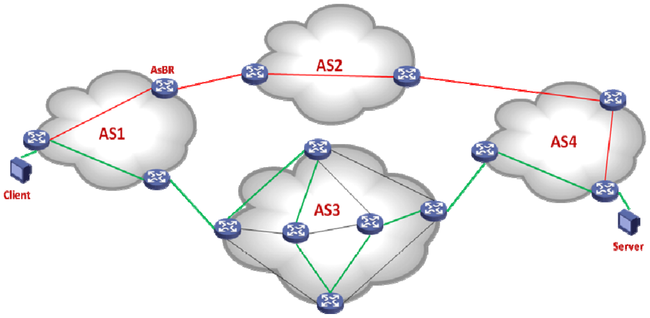

Fig. 3. Embodied Network Setup.

III.1. Embodiment of Our Scheme

As shown in Fig.3, normal packets from server to client are forwarded along the red path which is generated by searching routing table of BEFP hop-by-hop.

In comparison, the green path that is constructed by searching SFP-RIB, hop-by-hop, is the route along which, path messages of RSVP protocol are transmitted from server to client. Subsequent critical application packets will go through the green path along which bandwidth resource has already been reserved by RSVP's Resv messages [10].

III.2. Potential Applications

The potential applications of our proposed scheme is mainly for bandwidth or throughput guaranteed scenarios, which could be for super high-definition end-to-end real-time video streaming, tele-medical application, remote industrial operation, and so forth.

IV. Comparison and Analysis

IV.1 Intra-area Processing Compared with MPLS-TE

MPLS-TE runs CSPF algorithm for each request, while our solution runs only when bandwidth usage exceeds pre-defined thresholds. However, because of the 1st-threshold, bandwidth cannot be fully utilized in our scheme as in MPLS-TE.

There is a trade-off between bandwidth utilization and computational complexity, in which the utilization of bandwidth drops to the 1st-threshold, and the number of calculations is decreased, according to the number of requests.

Assuming the 1st-threshold to be 10% and the 2nd-threshold be 20%, the calculation comparison is given below, while the initial environment is set clean and no bandwidth is allocated.

Note that, with OSPF-TE, a threshold can also be used by MPLS-TE to reduce flooding times. Based on above calculation, our schemes demonstrates better scalability due to low complexity.

	M	*N*
G = 1%	1	90
G = 0.1%	1	900

M: Number of calculations of our scheme.
N: Number of calculations of MPLS-TE.
G: Bandwidth granularity.

IV.2. Analysis

IV.2.1. Bandwidth Granularity

The proposed scheme is aimed to the scenarios when the requested bandwidth is much less than the total available link bandwidth. For example, the total link bandwidth is 100Gbps, and the average bandwidth request might be in the scale of 100Mbps for resource reservation. However, in the case that, the applied bandwidth accounts for a noticeable percentage of overall available bandwidth, such as applying for 10Gbps bandwidth, some challenging issues might occur as follows.

IV.2.1.A, Failure of Resource Reservation

Recalling that there exist two thresholds defined in our scheme, when one particular requested bandwidth is higher than 2nd-threshold, such request will always fail. For a simple instance, the 1st-threshold is assumed to be 50Mbps, and the 2nd-threshold be 150Mbps, and the current available bandwidth is 160Mbps. When the applied bandwidth is 155Mbps, such request fails since the link cannot be removed from SFP-LSDB.

IV.2.1.B, Route Flapping

Under certain scenarios, the routes may flap, especially one particular link obtains and releases its associated resources in a short time, periodically.

Same as previous setup, if one link applies 120Mbps, it will be granted successfully, and then the total available bandwidth goes down to 40Mbps, and the corresponding link is excluded. After short time of usage, this link may release 120Mbps, then the total available bandwidth is recovered to be 160Mbps and this link is added back. However, if such allocation and release happen periodically, the route flaps over the whole network.

Such phenomenon is mainly due to the applied bandwidth is relatively high, comparing to the total available remaining resource. To solve this problem, multiple control plane routing tables could be utilized, so that the two thresholds can be configured flexibly according to different bandwidth granularities of applications. In addition, for large bandwidth requests, static setup might be adopted directly, without using the proposed automatic control.

IV.2.2. Convergence

The events that may trigger the convergence in our scheme, may not lead to the convergence in traditional BGP, e.g., bandwidth changing events. Thus, the number of iterations for convergence in our solution is necessarily greater than that of traditional BGP. For the belated convergence, especially caused by link removal, it may have serious problems, like packet dropping in traditional BGP. Therefore, it requires a fast convergence in traditional BGP, while the average time is usually at the level of minutes [11]. However, it is less severe in our solution, when the convergence is triggered by bandwidth depletion. In addition, if unreserved bandwidth of a link is less than 1st-threshold and the convergence is not initiated timely, the ongoing requesting messages can still try to reserve bandwidth on particular links. Due to the small amount of available bandwidth, these requests can be satisfied for a very short moment before final convergence.

V. Conclusion and Future Work

This paper introduces dedicated planes to forward signaling messages, while extensions of existing routing protocols are adopted for inter-domain end-to-end bandwidth constrained path computation. As a result, resource reservation upon these selected paths can avoid links with inadequate bandwidth in shortest path solutions, and fulfil the expected objective of distrusted inter-domain path selection for bandwidth guaranteed applications. In future, the configuration of optimized thresholds will be experimented, meanwhile, latency and routing convergence will also be studied with new proposal.

References

[1] V. Kotronis et al., “Control exchange points: Providing QoS-enabled end-to-end services via SDN-based inter-domain routing orchestration,” in Presented as Part of the Open Networking Summit 2014 (ONS 2014).

[2] D. Katz, et al., Traffic Engineering (TE) Extensions to OSPF Version 2, RFC 3630, September 2003.

[3] H. Ould-Brahim, D. Fedyk, and Y. Rekhter, BGP Traffic Engineering Attribute, RFC 5543, June 2009.

[4] H. Ould-Brahim, D. Fedyk, and Y. Rekhter, BGP-Based Auto-Discovery for Layer-1 VPNs, RFC 5195, June 2008.

[5] S. Jain et al., “B4: Experience with a globally-deployed software defined WAN,” in ACM SIGCOMM, 2013, pp. 3-14.

[6] S. H. Yeganeh and Y. Ganjali, “Kandoo: a framework for efficient and scalable offloading of control applications,” in Proceedings of ACM SIGCOMM HotSDN, 2012, pp. 19-24.

[7] A. Tootoonchian and Y. Ganjali, “Hyperflow: a distributed control plane for openflow,” in INM/WREN, 2010.

[8] T. Koponen et al., “Onix: a distributed control platform for large-scale production networks,” in OSDI, 2010.

[9] L. Subramanian, M. Caesar, C. Ee, M. Handley, M. Mao, S. Shenker, I. Stoica, “HLP: a next-generation inter-domain routing protocol,” ACM SIGCOMM, August 2005.

[10] R. Braden, et al., Resource ReSerVation Protocol (RSVP) - Version 1, Functional Specification, RFC 2205, September 1997.

[11] R. Oliveira, B. Zhang, D. Pei and L. Zhang, “Quantifying path exploration in the Internet,” IEEE/ACM Transactions on Networking, vol. 17, no. 2, pp. 445-458, 2009.