Dynamic QoS for High Quality SD-WAN Overlays

*Pham Tran Anh Quang, *Jérémie Leguay, *Feng Zeng, †Jianqiang Hou, †Boyuan Yu, ‡Davide Restivo
*Huawei Technologies Ltd., Paris Research Center
†Huawei Technologies Ltd., Nanjing R&D Center
‡Swisscom

***Abstract*—In SD-WAN networks, the traffic issued from multiple high capacity hubs towards low capacity spokes can create congestions in the underlay and degrade unnecessarily the Quality of Service (QoS). To mitigate this issue, shaping policies can be dynamically controlled to adapt to WAN performance and traffic. While existing solutions work for a single hub, at tunnel level, and in a reactive manner, we propose a more advanced solution for multiple hubs at application level. Our Dynamic QoS solution also takes proactive actions to prevent congestions and protect high priority traffic. It ensures a fast convergence towards an optimal rate allocation. This demonstration presents the design and implementation of this feature in AR8140 devices. It presents a performance evaluation in a testbed and simulation.**

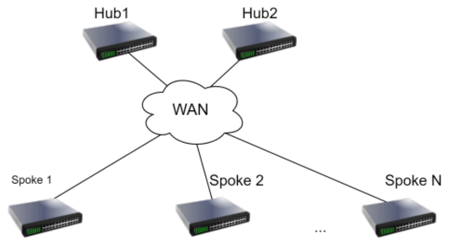

Fig. 1: Multi-hub SD-WAN scenario.

I. Introduction

SD-WAN [1] technologies are now widely employed by enterprises to interconnect remote sites without the need for dedicated and costly network infrastructures. Access Routers (ARs) are used to connect different sites, e.g. branches, headquarters or data centers, over multiple transport networks, e.g. broadband Internet, leased private lines. In this context, self-created congestions can happen in the underlay network when multiple sites, e.g. high capacity hubs, send too much traffic to the same destination site, e.g. a limited capacity spoke. Consequently, packets are randomly dropped in the underlay without proper differentiation among applications, e.g. between high and low priority flows. To mitigate this issue, a basic solution could be to reduce drastically the traffic sent by each site and apply a static shaping based on the ingress bandwidth at destinations. However, as traffic fluctuates over time, dynamic shaping solutions are a must to 1) maximize the utilization of the overlay and 2) ensure proper traffic differentiation.

To reduce WAN congestions, several solutions have been proposed to optimize load balancing [2]. However, after the transport network has been assigned to traffic flows thanks to load balancing, congestion can still happen in the underlay and the sending rate of sites must be controlled. In a previous work, we proposed [3] a centralized solution to globally optimize shaping policies. However, it requires frequent message exchanges between the centralized controller and ARs, which may not be desirable for scalability, fault-tolerance, or deployment reasons. Recently, commercial solutions [4], [5] based on rules have emerged to dynamically tune shaping parameters using WAN measurements and link utilization. However, these solutions operate only at tunnel level and for a single hub. Furthermore, these methods are reactive as they reduce shaping rates after a congestion has happened.

This demonstration showcases a new *Dynamic QoS* feature for ARs to operate at *flow group level* (i.e. flows from a set of applications) in *multi-hub* scenarios. The solution is *proactive* as it detects early congestions from latency variations and it is able to quickly protect high priority flow groups while maximizing a fair rate allocation for low priority ones.

Multi-hub scenario. As illustrated in Fig. 1, we consider a scenario where 2 hubs (headquarter and data center) are connected to multiple spokes (branches) using one transport network. Flows from applications are aggregated in *flow groups* that correspond to traffic classes with different SLA requirements (e.g. Production, VoIP, Office, Bulk). At egress ports, a hierarchical QoS scheduler based on Class-Based Queuing (CBQ) handles high priority flows with Priority Queuing (PQ) and low priority flows with Weighted Fair Queuing (WFQ). Hubs have a much higher access capacity than spokes (e.g. 1 Gbps vs. 100 Mbps) and most of the traffic goes from hubs to spokes. Therefore, when hubs are sending too much traffic, congestion happens in the underlay, e.g. at ingress ports of spokes. Moreover, even if hubs could be aware of the ingress capacity at spokes to shape traffic, the spoke capacity 1) is shared between multiple hubs and 2) it can be much greater than the actual WAN capacity (unknown to ARs).

Challenge: To ease deployment, the challenge is to develop a dynamic shaping solution, at flow group level to improve differentiation, that only operate at hubs and do not require any explicit exchange of information between hubs. The goal is to 1) protect high priority flow groups while 2) maximizing the rates allocated to low priority flow groups.

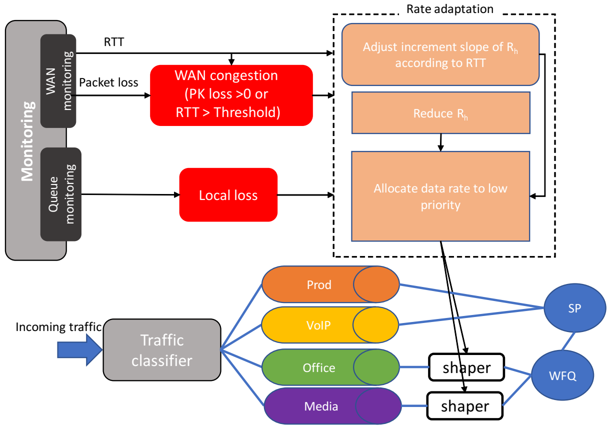

Fig. 2: Dynamic QoS solution at hubs.

II. Dynamic QoS solution

Overall architecture. Fig. 2 presents the architecture of the Dynamic QoS agent which controls, at hubs, the sending rate of low priority flow groups based on measurements about the WAN quality and the queue occupancy. Every 1s, the agent collects 1) WAN statistics about delay and packet loss, and 2) queue statistics about incoming and dropped traffic for each flow group. Based on this information, it decides about shaping rates following a two stage decision process. First, it determines the total rate that the hub is allowed to send to each spoke, called hub rate R_h. Then, after deducting the incoming traffic of high priority flow groups, the remaining rate is allocated to all low priority flow groups based on their priority (e.g. based on their traffic demands).

Global rate adaptation. When a congestion is detected, i.e. if packet loss or delay cross given thresholds, the hub rate is decreased. In case of WAN loss, it is cut by half to protect high priority flows. After the rate is decreased, no further decrease is allowed in the next 3 seconds as it takes some time to solve congestion and update monitoring. In case of high WAN delay but no WAN loss is observed, the hub rate is reduced by a small amount (i.e. 1% of R_h^*). The goal is to proactively decrease low priority traffic to avoid WAN loss. When both packet loss and delay are below thresholds, the agent supposes that there is no congestion and it starts looking at the queue occupancy. If packets are dropped in low priority queues, the hub rate is increased as the current traffic demand is not satisfied. Otherwise, the hub rate stays unchanged. To increase the hub rate efficiently is crucial, as the agent needs to rapidly converge toward a maximum hub rate once congestion has been resolved. In our implementation, we developed an advanced function based on delay measurements, which rapidly increases the rate when delay is small and slows down to avoid congestion. Unlike BBR [6] or Cubic [7], it has to operate with measurements every 1 s (instead of per RTT) and control an aggregate of transport sessions (instead of one).

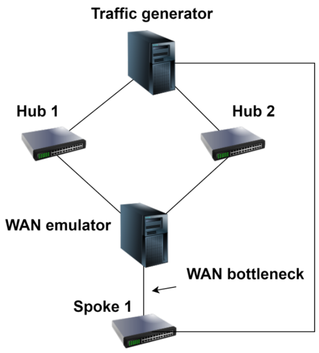

Fig. 3: Real testbed topology with AR8140 routers.

Rate allocation. Whenever the hub rate is adjusted in the above step, the shaping rate r_i for low priority flow groups are updated. For each flow group, the incoming traffic and its priority are denoted d_i and w_i, respectively. Recall that high priority traffic is not shaped. The shaping rates are defined by solving the following proportional fair optimization problem

$$\min \quad \sum_{i\in\mathbb{Q}_L} w_i \log r_i \tag{1}$$

$$\text{s.t.} \quad \sum_{i\in\mathbb{Q}_L} r_i \leq R_h - \sum_{i\in\mathbb{Q}_H} d_i, \tag{2}$$

$\mathbb{Q}_\mathbb{H}$ and $\mathbb{Q}_\mathbb{L}$ are the sets of high and low priority flow groups, respectively. The traffic demand of high priority flow groups, $\sum_{i\in\mathbb{Q}_H} d_i$, are subtracted from the rate hub R_h in Eq. 2.

III. Demonstration

To demonstrate the efficiency of Dynamic QoS, we conducted a comprehensive evaluation in both a real-world testbed and a simulation environment. Our evaluation focuses on three critical dimensions: (i) high-priority traffic protection, (ii) adaptability to changes in the Wide Area Network (WAN), and (iii) rapid convergence to an optimal rate allocation.

Testbed and implementation. In the real demonstration we will consider a sequence of traffic or WAN capacity variations in a simple scenario that well demonstrates the effectiveness of Dynamic QoS. The testbed consists in three AR8140 routers [8], where two ARs function as hubs (hub 1 and hub 2) and one AR operates as a spoke (spoke 1). The Dynamic QoS agent is deployed at hubs. High-priority traffic is sent from hub 1 to spoke 1, while low-priority traffic is sent from hub 2 to spoke 1. A Linux server with a traffic control application [9] emulates the WAN network, with a variable capacity ranging from 10 Mbps to 12 Mbps. The WAN's propagation delay is set at 10 ms. Fig. 4 presents the outcome on the receiving rates for the two flows (high and low priority flow groups coming each from a different hub).

At the beginning (t_0), high-priority traffic from hub 1 is set at 6 Mbps, and low-priority traffic from hub 2 is at 2 Mbps. Given the 10 Mbps bottleneck capacity, no congestion is observed. Then, at time t_1, low-priority traffic increases to 6 Mbps and the total traffic exceeds the WAN capacity.

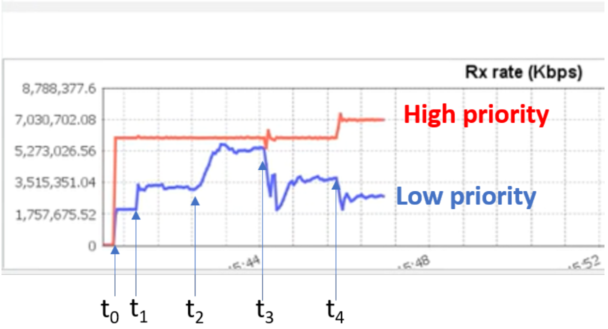

Fig. 4: Received data rates over time.

	Production	VoIP
Packet loss	0%	2%
Delay (ms)	40	60

TABLE I: SLA requirements of high priority flow groups.

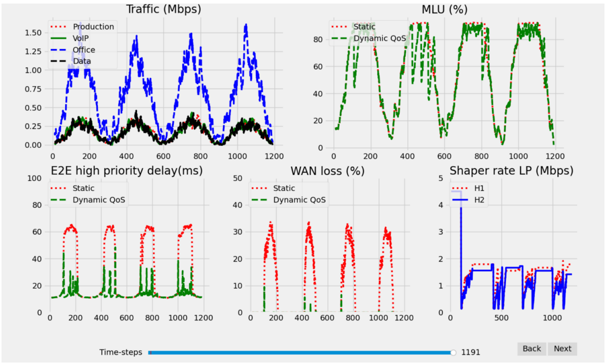

Fig. 5: Visualization of simulation results.

	Production	VoIP
Packet loss satisfaction	98.92%	99.83%
Delay satisfaction	98.65%	100%
Average packet loss	0.0086%	0.011%
95^{th} percentile packet loss	0%	0%

TABLE II: SLA satisfaction for high-priority flow groups.

Despite this, Dynamic QoS effectively adjust low-priority traffic, ensuring that the total traffic remains below the WAN capacity. No packet loss occurs in the WAN, preserving high-priority traffic integrity. At t_2, the WAN capacity increases to 12 Mbps, prompting the dynamic shaping mechanism to scale up the shaper for low-priority traffic. Consequently, the receiving rate of low-priority traffic increase to 5.2 Mbps. Conversely, at t_3, when the WAN capacity reverts to 10 Mbps, dynamic shaping promptly limits low-priority traffic upon detecting a critical RTT increase. This proactive action prevents WAN congestion and potential packet loss for high-priority traffic. We determined that the critical RTT increase should be 1.3 times the lowest measured RTT (RTTlow) using queuing theory. At t_4, high-priority traffic increases to 7 Mbps. The Dynamic QoS agent adjusts the shaping rate to approximately 2.8 Mbps, ensuring zero loss in the WAN. As we will observe, the shaping rates of low priority flows are efficiently adapted under varying network conditions. The full video of this demo can be found here http://tinyurl.com/5yfyzrmb.

Large scale scenario. Beside the real testbed with 3 ARs, we also implemented Dynamic QoS in the NS3 simulator [10] to validate performance in larger scale scenarios. We will present results with 2 hubs with 8 spokes. Traffic goes from the hubs towards the spokes. There are 4 flow groups in which 2 flow groups, i.e. production and voice, are high priority and 2 flow groups, i.e. office and data, are low priority. Low priority flow groups have no SLA requirements, while high priority flow groups have the SLA requirements shown in Tab. I. The WAN capacity is 5 Mbps (corresponding to spoke capacity). Dynamic QoS is compared with a static shaping solution where the spoke capacity is applied at hubs.

Fig. 5 shows a graphical interface to observe what happened during simulation. The top left plot presents the traffic demand for the different flow groups from one hub to one spoke. Each hub-spoke pair has identical traffic. Office traffic is the largest while other traffic patterns are similar. The top right plot is the Maximum Link Utilization (MLU). As it is shown, the MLU of the static mechanism is maximum and leads to congestions in the WAN, while Dynamic QoS offers a lower MLU which leads to almost no congestion. The bottom left and middle plots correspond to the delay and WAN loss of high priority traffic, respectively. The delay for Dynamic QoS is 4 times lower than for the static mechanism. The WAN loss for Dynamic QoS is almost 0% (except 4 small spikes), while for the static mechanism a high WAN loss episodes during congestions (e.g. over 30%) are observed. Tab. II presents the SLA results of Dynamic QoS. The satisfaction rate of a metric (packet loss or delay) is computed as the ratio of the total time in which the metric is lower than the requirement and the simulation time. In Dynamic QoS, the satisfaction rate of all high priority traffic is over 98%. The video replaying simulation results is available here: http://tinyurl.com/2y7y2dyn

References

[1] Z. Yang, Y. Cui, B. Li, Y. Liu, and Y. Xu, "Software-Defined Wide Area Network (SD-WAN): Architecture, Advances and Opportunities," in *IEEE ICCCN*, 2019, pp. 1–9.

[2] P. T. A. Quang, S. Martin, J. Leguay, X. Gong, and X. Huiying, "Intent-Based Routing Policy Optimization in SD-WAN," in *IEEE ICC*, 2022, pp. 4914–4919.

[3] P. T. A. Quang, J. Leguay, X. Gong, and X. Huiying, "Global QoS Policy Optimization in SD-WAN," in *IEEE NetSoft*, 2023, pp. 202–206.

[4] *Cisco Catalyst SD-WAN Forwarding and QoS Configuration Guide, Cisco IOS XE Catalyst SD-WAN Release 17.x.* Cisco Systems Inc.

[5] A. V. Networks, "Dynamic Congestion Management - Adaptive Shaping," Academy Versa Networks, Tech. Rep., Jun. 2020.

[6] N. Cardwell, Y. Cheng, C. S. Gunn, S. H. Yeganeh, and V. Jacobson, "Bbr: Congestion-based congestion control," *Communications of the ACM*, vol. 60, pp. 58–66, 2017.

[7] S. Ha, I. Rhee, and L. Xu, "CUBIC: A new TCP-friendly high-speed TCP variant," *ACM SIGOPS Operating Systems Review*, vol. 42, no. 5, pp. 64–74, Jul. 2008.

[8] "NetEngine AR8000 Series Enterprise Routers," https://e.huawei.com/en/products/routers/netengine-ar8000.

[9] "Traffic Control HOWTO," https://tldp.org/HOWTO/Traffic-Control-HOWTO/intro.html.

[10] G. F. Riley and T. R. Henderson, *The ns-3 Network Simulator*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 15–34.