

# The TCPPathChanger: bringing the spatial dimension to congestion control

Your N. Here  
*Your Institution*

Second Name  
*Second Institution*

## Abstract

To be provided

## 1 Introduction

Networks enable hosts to exchange information through a series of switches, routers and links. Some networks are over-provisioned, the network capacity is larger than the amount of data that the attached hosts can send or receive. Given economic constraints, almost all non-trivial networks, in particular the wide area ones, do not have enough capacity to handle all possible traffic patterns. They can carry some traffic patterns without any loss or queueing, but for most of the patterns, the hosts need to adjust their transmission rate to match the available network capacity.

Over the years, a wide range of techniques have been designed to enable the networks and the hosts to dynamically adjust. On the network side, a range of traffic engineering techniques have been proposed to enable network operators to control the amount of incoming traffic or redirect flows on less heavily used links. Example techniques include setting OSPF weights [12], using ATM, optical paths or MPLS [3, 35] or more recently leveraging Software Defined Networking [17] or Segment Routing [11]. These techniques typically operate on large aggregated flows and their reaction time is on the order of minutes or longer.

On a shorter timescale, endhosts have also adopted a range of congestion control techniques to adjust their transmission rate based on feedback received from the network. Many congestion control schemes have been included in the TCP protocol. They adjust the transmission rate based on the observed packet losses [14, 16], delay variations [4], explicit network feedback [29] or a combination of different signals [6, 9]. Other networks have explored different types of feedback such as explicit rate [20] or different forms of credit [24] in ATM networks. All these approaches consider the congestion control problem as a temporal one. All the packets belonging to a given flow follow the same path and they develop

algorithms and heuristics to determine **when** the next packet should be transmitted and delay packets when the network is congested.

Researchers have explored other techniques to better utilize the available network capacity. Theoretical results showing that it is interesting to utilize the different paths that exist between a pair of hosts [21, 22, 33] have motivated the design of Multipath TCP [28] and its associated coupled congestion control scheme [34]. Other multipath protocols have also adopted similar solutions [7, 15].

Although researchers showed the benefits that Multipath TCP could bring in datacenters [27], operators preferred to stick with single path protocols such as TCP. Given that TCP uses a single path, researchers explored the possibility of spraying packets over different paths [8] and techniques to reroute TCP connections when congestion is detected [18].

End-to-end approach have demonstrated their benefits to solve a wide range of network problems [31]. In this paper, we propose to add a spatial dimension to congestion control schemes. More specifically, we enable a TCP sender to use a set of  $n$  paths for each TCP connection. When a congestion event occurs, the sender can react by moving the TCP connection over another path from its set. This enables TCP senders to explore and use the different paths that reach the destination of each connection.

This paper is organised as follows. We first discuss in Section ?? how networks can efficiently expose different paths to endhosts. Due to paper size limitation, we focus on using IPv6 Segment Routing [26] in this paper, but our solutions are applicable to other techniques that enable endhosts to select network paths. Then in Section ??, we leverage eBPF to enable the Linux TCP stack to change the path of a connection upon detection of a congestion event. In Section ?? we evaluate the performance of TCPPathChanger in simple enterprise networks that utilize a few parallel wide-area links. We compare our approach to existing solutions. In Section ?? we use TCPPathChanger in emulated enterprise networks and demonstrate the ability of our solution to fully utilize the network even when only a few hosts communicate.

use cases TODO

Figure 1: Use cases

## 2 Architecture

### 2.1 Use cases

There are many possible use cases for a solution that enables hosts to dynamically reroute flows around congestion. We focus on three very different ones in this paper and provide all our software to enable other researchers to explore different use cases. Our first use case is content provider that manages a small datacenter that contains file, web or video streaming servers. This datacenter is connected to several transit ISPs that enable it to reach endusers. Many of these endusers are reachable via two or more transit ISPs. This is illustrated in the left part of Fig. 1. Our second use case is an enterprise network that has redundant wide area links. In such a network servers are typically installed in a few well-connected sites and they interact with all other sites. Given the cost of the WAN links, it is impossible to overprovision such a network. This is illustrated in the middle part of Fig. 1. Finally our third use case is /ldots

At a high-level, our design needs to answer three basic questions: (i) When should a host decide to reroute a congested flow, (ii) How are the available paths exposed to hosts and (iii) How are the packets belonging to a flow rerouted ?.

### 2.2 Rerouting congested flows

The choice of changing the path is left to the endhost instead of relying on the controller. This enables quicker reactions to congestion and better scalability of the network core. However letting endhost arbitrarily choose their own paths might lead to unstable solutions. Endhosts might decide to change the paths all at the same time and they could all choose the same path.

We first check that we waited a sufficient amount of time before moving the connection. Indeed, we implement an exponential backoff to prevent stability issues. Without this, all connections in the network would switch paths at the same time and this might cause congestion in another part of the network. Then, they would all switch back to previous situation and continue flapping between both paths.

To choose the good paths, we formulate our problem as an Adversarial Bandit Problem. In the Adversarial Bandit Problem [30], a gambler have to choose which of the  $K$  slot machines to play. At each time steps, he can pull the arm of one of the slot machines and he receives a reward (potentially negative). The problem is to find a tradeoff between the exploration of each slot machine and the exploitation of the most rewarding one.

In our path management, the gambler is the TCP connection

Parameters:  $\Gamma \in [0, 1]$ , value of  $w_i(t) \forall i \in \mathcal{P}$  at time step  $t$

- Compute the reward of the current path  $p$ :

$$x_{p_t}(t) \in [0, 1]$$

- Update its weight:

$$w_p(t+1) = w_p(t) \cdot \exp(\Gamma \frac{x_{p_t}(t)}{\text{prob}_p(t) \cdot K})$$

- Set  $\forall i \in \mathcal{P} \text{prob}_i(t+1) = (1 - \Gamma) \frac{w_i(t+1)}{\sum_{j=1}^K w_j} + \frac{\Gamma}{K}$
- Draw the new path  $p_{t+1}$  randomly according to  $\text{prob}_1(t+1), \dots, \text{prob}_K(t+1)$ .

Figure 2: Pseudocode of an EXP3 algorithm step

that wants to maximize its connection quality (i.e., latency, throughput, ...) The slots machines are the paths that we use by setting the corresponding SRH. The time steps are the congestion signals. We don't want to change paths if there is nothing bad happening. At each time steps, we compute the reward of the current path, that is, the connection quality achieved on this path.

We use the EXP3 algorithm [2] to solve our path management issue. The algorithm pseudocode is shown in Figure 2. We associate a weight to each path. When we consider changing the path, we evaluate the current path quality and we update its weight depending of this metric. A large increase of the weight means a good quality while a small increase indicates the opposite. Then, we make a weighted random choice to select the new path. Note that this "new" path could be the old one. In this case, we do not change the path.

There are three design decisions to make to use this algorithm: the initial path weights, the reward computation and the value of  $\Gamma$ . We set the initial weight to the maximum amount of bandwidth available on the path. The controller gives this information with the SRH. The reward is computed based on the last congestion window observed. **TODO: Change to mean/median cwin** The value of  $\Gamma$  can change the emphasis on the path weights. If we set this value to 1, each path has the same likelihood to be selected. If we set this value to 0, paths are selected stricly based on their weight. **TODO: Explain why we don't want value 0?**

### 2.3 Exposing paths to hosts

There are different possible ways to expose paths to hosts in real networks. Some approaches are applicable for servers in a datacenters, others are relevant for managed enterprise

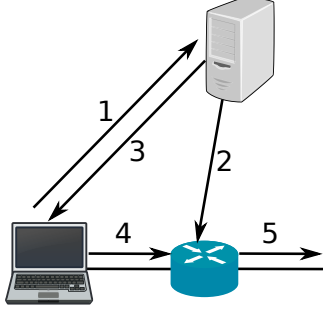


Figure 3: Endhost path query to the controller.

networks.

This paper proposes to integrate TCP with SRv6 to optimise congestion. We reuse a similar controller as in Software Resolved Networks [25]. As shown in Figure 3, an application on the endhost can query (1) the controller for a path to reach a particular destination. The SRN query is implemented by extending DNS requests. The controller computes a path, build its SRH and assign it a binding segment, that we call Path ID. After computation, the mapping between the binding segment and the SRH is inserted in the access router (2). Then, the controller sends the Path-ID inside the DNS reply back to the endhost (3). The application can then build an SRH containing only the Path-ID and insert it in all its packets (4). The access router reads the Path-ID and encapsulate the packet in an outer IPv6 packet with an SRH that steer the packet through the actual path.

While this first prototype shows that we can build a SDN solution with SRv6, this prototype has two issues. First, it requires to modify applications to talk to a controller. This raises deployment issues if you want to control the path taken by all your applications. Second, it did not consider reaction to congestion, only to failures.

In this paper, we offload the choice of the path to the endhost. The TCP stack of endhosts have access to the congestion state of each connection. This state, along with information from the controller enables the endhost to choose between paths that it received from the controller.

We extend the solution by enabling a daemon, the Path Daemon, in endhosts to request for a set of disjoint paths. The paths are disjoint so that we know that a congestion occurring on a path does not influence the others. The endhost can remain agnostic of the actual topology of the network and it will not waste time hitting paths that suffer from congestion on the same link.

We use **TODO: insert algo name ???** from [1] to find these disjoint paths. This algorithm is specific to SRv6 because it compute disjoint paths between two endpoints and checks that these paths can be expressed in fewer segments than a given limit. Some hardware components in their network do not support arbitrarily long SRHs as documented in [?].

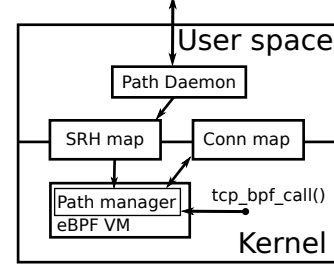


Figure 4: Path management code talks through eBPF maps to the Path Daemon.

The controller computes a set of disjoint paths for every set of pair of source prefix/destination prefix. Each endhost daemon receives a list of Path IDs for pairs matching its access router as source. Along with the Path ID, the controller sends the bandwidth capacity, its current usage and its delay. We scale up the number of controllers with the number of endhost **TODO: inaccurate, it is actually the database but it might be simpler to explain that.**

## 2.4 Enforcing paths for flows

As in the previous sections, several technical solutions are possible to enable hosts to request the utilisation of a specific path through a network. They all require the hosts to provide a hint on the path that it wishes to use by setting a specific field in the packet's header. A first approach could be to use Multi-Topology Routing [?] and indicate with the DSCP field of the IP header the topology selected for each packet. This requires the network administrator to carefully select different topologies that he/she wants to disseminate with the intradomain routing protocol. This solution would work well in a datacenter with a few different border routers. Another approach would be to rely on VLAN tags.

In this work, we build and evaluate our prototype on top of IPv6 Segment Routing (IPv6SR) [?]. IPv6 SR is a modern version of source routing that implements the Segment Routing architecture [11], that is being finalised within the IETF.

## 2.5 Application-independent path management

The previous section describes the path injection to a daemon that runs on the endhost. However this endhost needs to enable other applications to manage these paths without actually modifying them. A solution could be to modify the kernel directly to include this path management. However, this would require operators to modify the kernel of all the endhosts. An alternative is to inject code in the kernel thanks to eBPF hooks in the kernel. Most of these hooks exist since Linux v.**TODO: version ???**. **TODO: Quid eBPF windows?**

In [25], applications had to be modified to inject their Path IDs in their sockets so that they are inserted in the packets. This limits the deployment of the solution. We need a way to inject the Path ID that is application-independent. Another different with [25] is that we want to react from congestion events inside the endhost. We want to inject a path management behavior inside TCP which is quick to react to congestion in the network. Modifying the kernel directly would require operators to update the kernel of all their machines. Therefore, relying on kernel modification is not the best solution.

Since Linux **TODO: version ???**, programmer can inject code to the kernel thanks to the eBPF VM [?]. We can inject code at various locations, called hooks, each calling a `tcp_bpf_call()` function. We use four of the hooks in our prototype. `BPF_SOCK_OPS_TCP_CONNECT_CB` is call before sending the SYN packet. This means we can still choose the SRH of the SYN packet in our path management code. We also use `BPF_SOCK_OPS_STATE_CB` that is called each time the state of TCP changes. We use it to know when the connection is closed to clean the per-connection state that we maintain. `BPF_SOCK_OPS_RTO_CB` is called after the expiration of the retransmission timer which is an important congestion signal. `BPF_SOCK_OPS_ECN_CE` is called before sending the segment setting CWR bit in response to packets with ECN bit set. The hook is the only one that we add to the kernel. We choose to place it when sending the CWR instead of the reception of the actual ECN because the CWR bit is sent only once by RTT. This is better done there than in the path management code.

Figure 4 shows the interactions of the different components inside the endhost. The Path manager code runs in an the isolated eBPF VM inside the kernel space. A verifier [?], packaged with the kernel, checks memory calls and program termination of the code before injection it so that we are guaranteed that the code does not make the kernel crash **TODO: Comments on other verifications ?**. For injected code in the TCP stack, the injected code receives a structure containing the code of the hook that triggered the code, the five tuple, and some variables of the TCP connection, such as the minimum RTT or the congestion window. Moreover, it can read and writes to memory chunks, called eBPF maps, that can be written to or read by a user space application. In this architecture, our daemon will fill a first eBPF map, the SRH map with the Path IDs received from the controller. This map is a hashmap that maps the destination prefix to a list of structures containing a Path ID, the path bandwidth, its latency,... (i.e., all the pieces of information transmitted by the controller). **TODO: Discuss how the eBPF map can support multiple size of prefixes?** Since eBPF injected code does not have global variable or heap, we use a second eBPF map also oragnized as hashmap, the connection map. The key is the five tuple and it maps to a structure containing data about the connection.

**TODO: Explain cgroups?**

```

1 int handle_sockop(struct bpf_sock_ops *skops)
2 {
3     struct five_tuple five_tuple;
4     get_five_tuple(&five_tuple, skops);
5
6     struct conn *conn = bpf_map_lookup_elem(&
7     conn_map, &five_tuple);
8     if (!conn) { // New connection
9         struct connection new_conn;
10        memset(new_conn, 0, sizeof(new_conn));
11        new_conn.srh = get_best_path(srh_map, &
12        five_tuple.remote);
13        bpf_setsockopt(skops, SOL_IPV6, IPV6_RTHDR
14        , new_conn.srh, sizeof(*new_conn.srh));
15        init_exponential_backoff(&new_conn);
16        bpf_map_update_elem(conn_map, &five_tuple,
17        &new_conn, BPF_ANY);
18        *conn = new_conn;
19    }
20
21    switch (skops->op) {
22        case BPF_SOCK_OPS_STATE_CB:
23            if (skops->args[1] == BPF_TCP_CLOSE)
24                bpf_map_delete_elem(conn_map, &
25                five_tuple);
26            break;
27        case BPF_SOCK_OPS_ECN_CE:
28            // TCP segment with CWR bit set
29            react_to_congestion(five_tuple, conn);
30            break;
31        case BPF_SOCK_OPS_RTO_CB:
32            // Retransmission timer expiration
33            react_to_congestion(five_tuple, conn);
34            break;
35    }
36    socks->reply = rv;
37    return 0;
38 }

```

Listing 1: Path management

Listing 1 shows the pseudo code of the path management. When an application starts a connection to a server, the eBPF hook ??? is triggered. We start by retrieving the five tuple and check if an entry for this connection exists. Since it is the new connection, it won't find it. Therefore, we create this entry, choose the best path for the destination and change set the SRH with the Path ID in the socket so that it is inserted in every packet sent. After that, we identify the hook that triggered the call. When the TCP state changes, and when the connection is closed, we clean the connection state of the eBPF map. When we receive a congestion signal, either triggered by ECN or by the retransmission timer expiration, we call another routine to choose the new path.

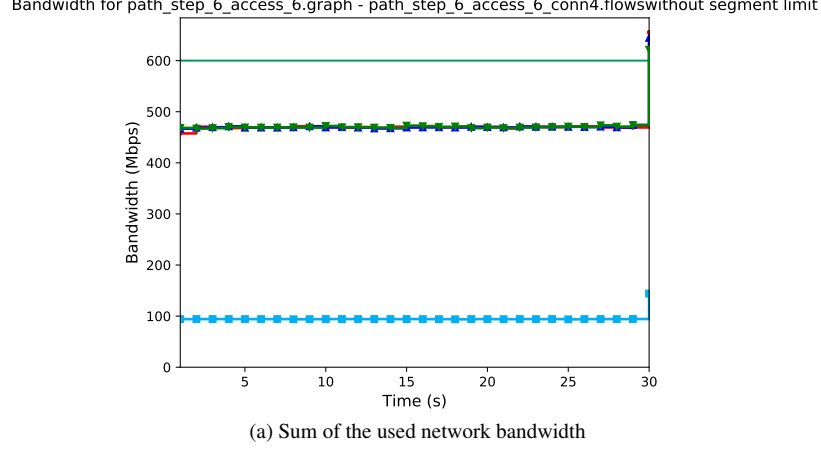


Figure 5: Simple example with 6 disjoint paths and 4 clients communicating with 4 servers

## 2.6 Stable path management with EXP3

## 3 Evaluation

### 3.1 Simple example

We use the topology shown in Figure ?? . This simple network only has links with the same bandwidth (100 Mbps) and latency (1ms). We prevent ECMP from working by setting arbitrary IGP weights on the links. **TODO: Test with ECMP enabled** Each endhost starts 4 connections using iperf3 with the server in front of it the network. We emulate the network with Mininet [?] on a Linux kernel 5.3 running on machine with 20 CPU and 16 GB of RAM.

Figure 5a shows the network bandwidth used by all the connections through time. In theory, this badnwidth could go as high as 600 Mbps according to the Maximum Flow approximation [?] since there are six disjoint paths available. In practice, the total bandwidth is slightly below that number but since this number is above 500Mbps, we can deduce that all of the six paths are used. Regular TCP connections, without ECMP, cannot achive more than 100Mbps because they can only use one path. **TODO: Test with ECMP enabled** We do not see significant changes with different values of GAMMA. **TODO: Discuss with Schapira**

### 3.2 Topology Zoo set

We used 233 topologies from the Topology Zoo [23]. We modified these topologies to remove part of the graphs that lack link redundancy... trees. The resulting topologies have only routers with at least two links. Communication between routers in a tree can only use a single path. Such trees cannot see any improvement from using Segment Routing. Moreover, such trees are likely artifacts of the topology collection because they cannot preserve connectivity even with a single link failure. Figure 6a and 6b shows the CDF of the number of

routers and links in the resulting topologies. Figure 6c shows the distributions of routers percentage with a given number of attached links.

The set of topologies does not include which routers are actually linked to endhosts and which are core routers. We assume that routers with two links are edge routers and the others are core routers because they are more connected.

### 3.3 Maximum improvement

This sections evaluation the theoritical improvement of using multipath transport protocols and Segment Routing on the endhosts instead of a single path protocol with only shortest paths.

We used a linear programming formulation of the Max Flow problem. Given a set of paths linking a set of the edge routers, we compute the maximum amount of traffic that can go through the topology without exceeding the link capacity.

$$\begin{aligned} & \max_f \quad \sum_{p \in \mathcal{P}} f_p \\ & \text{s.t.} \quad \sum_{p \in \mathcal{P}} r(l, p) \cdot f_p \leq c(l) \quad \forall l \in \mathcal{L} \end{aligned}$$

$\mathcal{P}$  represents the set of sr-paths, encoded as a list of intermediate routers. As for IPv6 SR, The real path taken between each intermediate router is the shortest one.  $r(l, p)$  represents the ratio of flow that goes through a particular link. This ratio is used to model the ECMP splitting of the traffic.

To model a single-path transport protocol without SR we insert all the shortest sr-paths between each pair of edge router to  $\mathcal{P}$ . We assume that we have a high number of connections for each path and therefore that hash-based ECMP splits the traffic evenly. [5] shows that it is not the case in practise.

Multipath transport protocols with SR also uses shortest path routing. However, they can establish a big number of



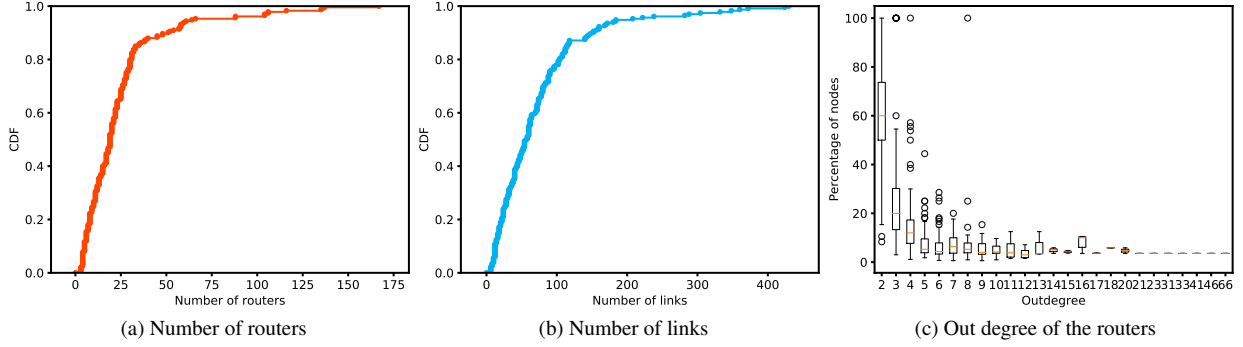


Figure 6: Features of the evaluation topologies

subflows for each connection and hope that they will cover all or a portion of the shortest paths. This hijacks ECMP that sees subflows as independent connections. Flow Bender [19] and SEMPER [13] used this solution to leverage some of the path diversity of a topology. We will assume for simplicity that this method works perfectly and that multipath protocols can always cover all the shortest paths. We insert a sr-path for each shortest path between each pair of edge routers instead of giving one sr-path representing all the shortest paths between a given pair.

Introducing Segment Routing means adding more paths to  $\mathcal{P}$ . Note that generating the set of all possible paths is in  $O(|\mathcal{R}|^k)$  with  $\mathcal{R}$  being the set of routers and  $k$  the maximum number of intermediate nodes. This is not practicle for big topologies. Therefore, we reuse the approach of [?] that iteratively adds relevant paths to the set until no interesting paths can be added to  $\mathcal{P}$ . When it reaches this point, computing the model on this restricted set or on the complete one gives the same solution. **TODO: Explain how the model is different from infocom, i.e., demands become unbounded or is it too detailed ?**

**Adding IPv6 SR can significantly improve the maximum throughput.** Figure 7a shows the relative maximum flow improvement of using single path or multiple path with SR. There is a boxplot for a given percentage of edge router pairs that have heavy hitters to route **TODO: Give the rational behind the choice of the values.** The number of heavy hitters has influence on the relative improvement. It is easier to unused alternative paths with fewer heavy hitters. With 1% of heavy hitter pairs, for half of the topologies, we can improve the maximum flow by 71%. For 10% of heavy hitter pairs, we can still improve it by 30%. Figure 7b shows the improvement with multiple path protocol without SR. The improvement is about 5% lower. This justifies the use of SR even with multipath protocols.

### 3.4 Largest set of disjoint paths

Finding the largest set of disjoint paths between two routers can be done with Edmonds–Karp algorithm [10] in  $O(|\mathcal{R}||\mathcal{L}|^2)$ . Even faster strategies can imply applying Dijkstra algorithm iteratively. Each time that the algorithm finds a path, we remove its links for the next computations. We stop it when it cannot find another path. This is solved in  $O((|\mathcal{L}| + |\mathcal{R}|)n \log |\mathcal{R}|)$  with  $n$  being the number of iterations of dijkstra algorithm.  $O(n)$  cannot be bigger than  $O(|\mathcal{L}|)$  because at each iteration, if a path is found, the links if this path will be removed. In practise, source and destinations are behind routers with exactly two outgoing links. There will be at most two disjoint paths. Therefore, applying the Dijkstra algorithm iteratively is faster than using the Edmonds-Karp.

Figure 8a shows a CDF of the mean number of disjoint paths by pair of access router in each topology. We consider routers with exactly two links as access routers. This means that there are at least one path between two routers but at most two since there are only two outgoing links. Edmonds-Karp computes the largest set of disjoint paths that can be reached. We can reach the same sets for 70% of topologies by applying iteratively Dijkstra algorithm.

SRv6 is limited in the number of segments that it can apply in a packet. This limit can be due to hardware limitations of IPv6 SR implementations [32]. So we need to use algorithms that prevent the number of segments to grow beyond a given limit. Always finding the largest set of disjoint paths within the limit of segment is exponential and therefore not feasible in practise. [1] defines the SPSTODO: **Check with Francois for the name** algorithm that computes the shortest path between two routers under a given limit of segments. Each time that the algorithm finds a path, they remove all the links for the next computations. They stop when no more path can be found. Figure 8b shows that the number of disjoint paths increases with the number of authorized segments. With only one segment, we can only have one path because only the destination can be encoded in the segment list. Increasing to 4 segments, (i.e., 3 intermediate routers), signigicantly im-

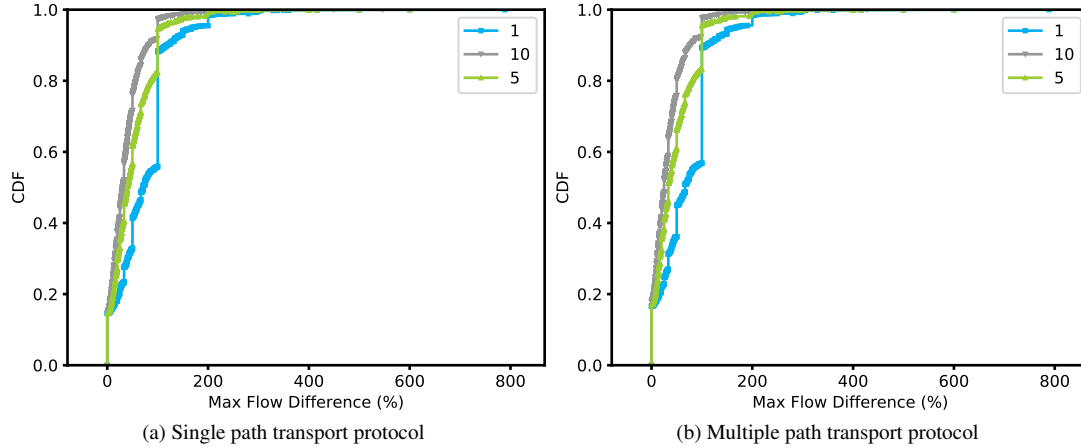


Figure 7: Relative Maxflow improvement of adding IPv6 SR

proves the number of disjoint paths. We see that increasing the limit beyond this does not improve significantly the number of disjoint paths.

#### TODO: Conclusion

#### References

- [1] François Aubry, David Lebrun, Yves Deville, and Olivier Bonaventure. Traffic duplication through segmentable disjoint paths. In *2015 IFIP Networking Conference (IFIP Networking)*, pages 1–9. IEEE, 2015.
- [2] Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002.
- [3] Daniel Awduche, Angela Chiu, Anwar Elwalid, Indra Widjaja, and XiPeng Xiao. Overview and principles of internet traffic engineering. Technical report, RFC 3272, may, 2002.
- [4] Lawrence S. Brakmo and Larry L. Peterson. Tcp vegas: End to end congestion avoidance on a global internet. *IEEE Journal on selected Areas in communications*, 13(8):1465–1480, 1995.
- [5] Zhiruo Cao, Zheng Wang, and Ellen Zegura. Performance of hashing-based schemes for internet load balancing. In *Proceedings IEEE INFOCOM 2000. Conference on Computer Communications. Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies (Cat. No. 00CH37064)*, volume 1, pages 332–341. IEEE, 2000.
- [6] Neal Cardwell, Yuchung Cheng, C Stephen Gunn, Soheil Hassas Yeganeh, and Van Jacobson. Bbr: congestion-based congestion control. *Communications of the ACM*, 60(2):58–66, 2017.
- [7] Quentin De Coninck and Olivier Bonaventure. Multi-path quic: Design and evaluation. In *Proceedings of the 13th international conference on emerging networking experiments and technologies*, pages 160–166, 2017.
- [8] Advait Dixit, Pawan Prakash, Y Charlie Hu, and Ramana Rao Kompella. On the impact of packet spraying in data center networks. In *2013 Proceedings IEEE INFOCOM*, pages 2130–2138. IEEE, 2013.
- [9] Mo Dong, Qingxi Li, Doron Zarchy, P Brighten Godfrey, and Michael Schapira. {PCC}: Re-architecting congestion control for consistent high performance. In *12th {USENIX} Symposium on Networked Systems Design and Implementation ({NSDI} 15)*, pages 395–408, 2015.
- [10] Jack Edmonds and Richard M Karp. Theoretical improvements in algorithmic efficiency for network flow problems. *Journal of the ACM (JACM)*, 19(2):248–264, 1972.
- [11] Clarence Filsfils, Nagendra Kumar Nainar, Carlos Pignataro, Juan Camilo Cardona, and Pierre Francois. The segment routing architecture. In *2015 IEEE Global Communications Conference (GLOBECOM)*, pages 1–6. IEEE, 2015.
- [12] Bernard Fortz and Mikkel Thorup. Internet traffic engineering by optimizing ospf weights. In *Proceedings IEEE INFOCOM 2000. Conference on Computer Communications. Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies (Cat. No. 00CH37064)*, volume 2, pages 519–528. IEEE, 2000.

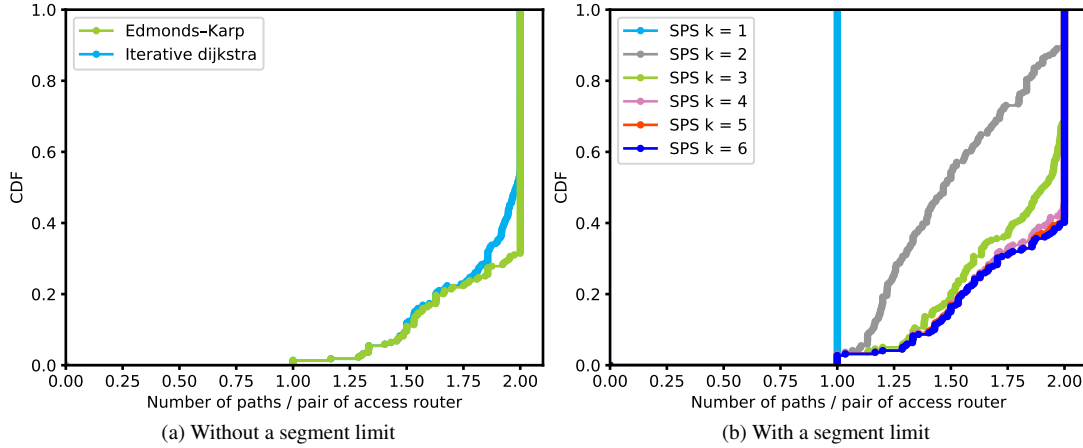


Figure 8: Number of disjoint paths found by pair of access routers

- [13] Gines Garcia-Aviles, Marco Gramaglia, Pablo Serrano, Marc Portoles, Albert Banchs, and Fabio Maino. Semp: a stateless traffic engineering solution for wan based on mp-tcp. In *2018 IEEE International Conference on Communications (ICC)*, pages 1–6. IEEE, 2018.
- [14] Sangtae Ha, Injong Rhee, and Lisong Xu. Cubic: a new tcp-friendly high-speed tcp variant. *ACM SIGOPS operating systems review*, 42(5):64–74, 2008.
- [15] Janardhan R Iyengar, Paul D Amer, and Randall Stewart. Concurrent multipath transfer using sctp multihoming over independent end-to-end paths. *IEEE/ACM Transactions on networking*, 14(5):951–964, 2006.
- [16] Van Jacobson. Congestion avoidance and control. *ACM SIGCOMM computer communication review*, 18(4):314–329, 1988.
- [17] Sushant Jain, Alok Kumar, Subhasree Mandal, Joon Ong, Leon Poutievski, Arjun Singh, Subbaiah Venkata, Jim Wanderer, Junlan Zhou, Min Zhu, et al. B4: Experience with a globally-deployed software defined wan. *ACM SIGCOMM Computer Communication Review*, 43(4):3–14, 2013.
- [18] Abdul Kabbani, Balajee Vamanan, Jahangir Hasan, and Fabien Duchene. Flowbender: Flow-level adaptive routing for improved latency and throughput in datacenter networks. In *Proceedings of the 10th ACM International on Conference on emerging Networking Experiments and Technologies*, pages 149–160, 2014.
- [19] Abdul Kabbani, Balajee Vamanan, Jahangir Hasan, and Fabien Duchene. Flowbender: Flow-level adaptive routing for improved latency and throughput in datacenter networks. In *Proceedings of the 10th ACM International on Conference on emerging Networking Experiments and Technologies*, pages 149–160. ACM, 2014.
- [20] Shivkumar Kalyanaraman, Raj Jain, Sonia Fahmy, Rohit Goyal, and Bobby Vandalore. The erica switch algorithm for abr traffic management in atm networks. *IEEE/ACM Transactions on networking*, 8(1):87–98, 2000.
- [21] Frank Kelly and Thomas Voice. Stability of end-to-end algorithms for joint routing and rate control. *ACM SIGCOMM Computer Communication Review*, 35(2):5–12, 2005.
- [22] Peter Key, Laurent Massoulié, and Don Towsley. Multipath routing, congestion control and dynamic load balancing. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP’07*, volume 4, pages IV–1341. IEEE, 2007.
- [23] Simon Knight, Hung X Nguyen, Nickolas Falkner, Rhys Bowden, and Matthew Roughan. The internet topology zoo. *IEEE Journal on Selected Areas in Communications*, 29(9):1765–1775, 2011.
- [24] NT Kung and Robert Morris. Credit-based flow control for atm networks. *IEEE network*, 9(2):40–48, 1995.
- [25] David Lebrun, Mathieu Jadin, François Clad, Clarence Filsfils, and Olivier Bonaventure. Software resolved networks: Rethinking enterprise networks with ipv6 segment routing. In *Proceedings of the Symposium on SDN Research*, page 6. ACM, 2018.
- [26] Stefano Previdi, Clarence Filsfils, Brian Field, Ida Leung, J Linkova, Ebben Aries, Tomoya Kosugi, Eric Vyncke, and David Lebrun. Ipv6 segment routing header (srh). *Internet Engineering Task Force, Internet-Draft draft-ietf-6man-segment-routing-header-08*, 2017.
- [27] Costin Raiciu, Sebastien Barre, Christopher Pluntke, Adam Greenhalgh, Damon Wischik, and Mark Handley.



- Improving datacenter performance and robustness with multipath tcp. *ACM SIGCOMM Computer Communication Review*, 41(4):266–277, 2011.
- [28] Costin Raiciu, Christoph Paasch, Sebastien Barre, Alan Ford, Michio Honda, Fabien Duchene, Olivier Bonaventure, and Mark Handley. How hard can it be? designing and implementing a deployable multipath {TCP}. In *9th {USENIX} Symposium on Networked Systems Design and Implementation ({NSDI} 12)*, pages 399–412, 2012.
  - [29] Kadangode Ramakrishnan and Sally Floyd. A proposal to add explicit congestion notification (ecn) to ip. Technical report, RFC 2481, January, 1999.
  - [30] Herbert Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–535, 1952.
  - [31] Jerome H Saltzer, David P Reed, and David D Clark. End-to-end arguments in system design. *ACM Transactions on Computer Systems (TOCS)*, 2(4):277–288, 1984.
  - [32] J. Tantsura. The critical role of Maximum SID Depth (MSD) hardware limitations in Segment Routing ecosystem and how to work around those. In *NANOG71*, 2017.
  - [33] Damon Wischik, Mark Handley, and Marcelo Bagnulo Braun. The resource pooling principle. *ACM SIGCOMM Computer Communication Review*, 38(5):47–52, 2008.
  - [34] Damon Wischik, Costin Raiciu, Adam Greenhalgh, and Mark Handley. Design, implementation and evaluation of congestion control for multipath tcp. In *NSDI*, volume 11, pages 8–8, 2011.
  - [35] Xipeng Xiao, Alan Hannan, Brook Bailey, and Lionel M Ni. Traffic engineering with mpls in the internet. *IEEE network*, 14(2):28–33, 2000.