3
Topology Aware Flow Prioritization

3.1 INTRODUCTION

Data center technology has been growing in the last few years because it can accommodate variety of applications. Some data intensive applications are deadline sensitive and require minimum flow completion time. Applications, such as online gaming require continuous interactions with the users. Users expect the response of the keystroke or mouse click on screen as soon as possible. Such type of applications impose greater pressure on the service provider for meeting Service Level Agreement (SLA). In the literature, there exists a large number of proposals for either reducing average flow completion time or the long-tail of flow completion. The state of art proposal is pFabric [Alizadeh et al., 2013] which uses shortest remaining size first scheduling for reducing flow completion time. It uses remaining size of flow for prioritization and scheduling of a flow.

This chapter discusses Topology Aware pFabric (TAP) and Topology aware Preemptive Shortest Remaining Size First (TP-SRSF) prioritizing schemes for reducing flow completion time. In TAP, the distance[1] information with the remaining size of flow is used for flow prioritization. In addition to distance information, TP-SRSF uses flow duration information for the purpose of flow prioritization. The results show that prioritizing long distance flows and newer flows further reduces average flow completion time and minimizes the total number of timeout events.

3.2 pFabric

pFabric [Alizadeh et al., 2013] is designed with the aim to obtain near to optimal flow completion time for data center traffic. In pFabric, packet's scheduling and dropping is based on priorities. By associating a single priority to the packets and implementing priority queue at switch, pFabric decouples the flow scheduling from rate control. The source node assigns a priority to each packet of a flow based on the remaining size of the flow. At the switch these packets are scheduled for transmission or accepted into the buffer strictly based on the priority. In case of congestion only the lower priority packets are dropped no matter they are newly arrived or not. In other words, a packet arriving at the switch is either dropped if it has lower priority than all the packets in the buffer, or it is accepted by dropping the lowest priority packet from the buffer. For transmission of a packet the switch selects the highest priority packet from buffer. pFabric uses minimal rate control mechanism, where initially all flows start transmission at line rate and decrease their sending rate only if they experience high and persistent loss. pFabric requires a large number of priority levels, which are not supported in present network switching fabric.

3.3 MOTIVATION

A data center is built using multi-rooted tree for gaining high bandwidth. Layer-2 forwarding is cost effective and makes services location independent, but it increases the overhead of broadcast traffic and does not support QoS or filtering. In Layer-3 based data centers, IP addresses of hosts are assigned hierarchically. For any communication, a source always knows destination address of

[1] Distance refers to number of hops

the receiver. All the hosts that are connected to a ToR (Top of Rack) switch could be assigned the same network prefix address [Niranjan Mysore et al., 2009] and length of this prefix address depends on the number of servers connected to the ToR. For example a ToR switch connecting n machines needs r bits where

$$n \leq 2^r$$

and

$$\text{prefix size} = \text{Size of address}(32 \text{ bit}) - r$$

Similarly, all hosts of a row could be assigned the same network prefix address. Prefix address length of a row is smaller compared to that of a rack as a row has multiple racks. This type of address aggregation enables small forwarding tables across all data center switches and reduces the look-up time.

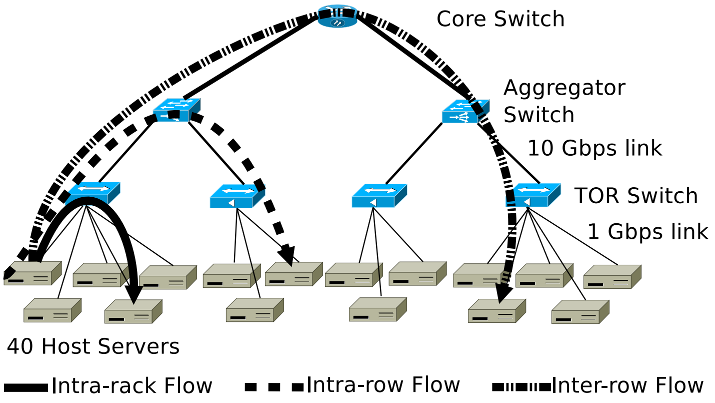


Figure 3.1. : 3-tier data center topology

With such a careful address assignment, any sender in a data center can find the distance to the receiver by maximal address comparison. For example, if last r bits of the IP address of sender and receiver are not matched, they can communicate directly using a single ToR switch. So in this case the distance (hop length) between sender and receiver is 1. If unmatched bits are more than r, then the communication between them is not possible using only ToR. In this case, at least one aggregation switch must be involved. Therefore, the distance between sender and receiver is either 3 or 5. In literature and to best of our knowledge, data centers are at most of 3-tier. Maximum distance between any pair of sender and receiver nodes in 3-tier data center is 5 as shown in Figure 3.1. Similarly for 2-tier data center the maximum distance between any pair of sender and receiver nodes is 3 as shown in Figure 3.2. A sender can have multiple flows at a time with different distances. Flows with distance 1 are known as intra-rack flows.

All to all communication pattern is very common in data center. In case of 3-tier data center the total number of intra-rack connection with distance 1 is

$$R_{row} \times R_{rack} \times n(n-1)$$

where R_{rack} is number of racks, n denotes hosts per rack and R_{row} is number of rows in data center. Similarly total number of intra-row connections with distance 3 is

$$R_{row} \times R_{rack} \times n^2(R_{rack} - 1)$$

and the total number of inter-row connections with distance 5 is

$$R_{row} \times R_{rack}^2 \times n^2(R_{row} - 1)$$

Generally in a data center there are multiple racks and every rack will have multiple machines. For any value of $R_{rack} > 1$, $n > 1$ and $R_{row} > 1$ the number of connections with long distance is more compared to the short distance connections. The analysis given above considers unidirectional connections between the end hosts.

In queuing theory, the expected queue length increases with load. Packets of long distance flows spend more time in queue compared to those of short distance flows. As the packets of long distance flows pass through more number of queues compared to those of short distance flows, the round trip time (RTT) of long distance flows gets further escalated due to this unfair behavior. Congestion window of short distance flows are updated frequently compared to that of long distance flows; this will affect the throughput of long distance flows. This frequent update of congestion window leads to unfairness among long and short distance flows [2]. As already stated, the number of short distance flows is less compared to that of the long distance flows, further the contention for bandwidth for short distance flows is less compared to that of the long distance flows. Therefore, higher preference needs to be given to long distance flows.

3.4 FLOW PRIORITIZATION

Flow prioritization is an important aspect to reduce the latency of flow as it helps to schedule the flows appropriately. This section discusses two prioritization schemes namely TAP and TP-SRSF. TAP is based on previously proposed scheme pFabric and it uses distance of flow with the remaining size of flow. Further TP-SRSF uses the idea of TAP as well as time duration for prioritization of a flow.

3.4.1 TAP

In pFabric, a sender sends the remaining size of flow as a priority in IP header and router schedules the lowest remaining size packet first. It is assumed in pFabric that the sender also knows the flow size.

$$\text{Priority}_{\text{pFabric}} = \text{Remaining Flow Size}$$

However pFabric does not differentiate among long or short distance flows.

In this proposed scheme, it is assumed that a sender knows the distance to the receiver, which is possible by hierarchically IP address assignment as discussed earlier. It is also assumed that the sender knows flow size as in pFabric. A flow is divided into two categories, small flow (mice flow) and large flow (elephant flows) based on its size. All flows with size less than a specific threshold are treated as mice flows and flows with size greater than the threshold are treated as elephant flows. Researchers have different opinions on the question of the clear cut threshold for mice and elephant flows. Some researchers treat all flows less than 100 KB as mice flows and few researchers suggest the threshold for mice and elephant flows should be 200 KB [Munir et al., 2013] [Alizadeh et al., 2013]. TAP gives higher priority to long distance mice flows compared to short distance mice flows. In TAP, priority of mice flows is calculated as

[2] Note that the condition of TCP outcast is drop-tail queue[Prakash et al., 2012], not applicable here

$$\text{Priority}_{\text{TAP}} = \frac{\text{Remaining Flow size}}{\text{Flow distance}}$$

For elephant flows priority calculation is same as in pFabric because elephant flows are not deadline sensitive. In case of same remaining size, the packets of long distance mice flow will be scheduled before those of short distance mice flow.

It is well defined that the number of mice flows are higher compared to the elephant flows [Alizadeh et al., 2010]. In all to all communication as already discussed, the number of long distance flows are more compared to the short distance flows. Therefore, the majority of flows get benefited using TAP that helps to reduce timeout events for mice flows. This reduction of timeout events further reduces average flow completion time.

3.4.2 TP-SRSF

Two preference schemes, namely Milk and Wine are well defined in the literature. The Milk scheme gives higher preference to new one's compared to old one's. In contrary to the Milk scheme Wine gives higher preference to older one's. The Milk scheme (new is better than old) provides a way of preemptive scheduling.

In shortest remaining size first (SRSF) scheduling as a flow progresses, the size of the flow reduces and the priority of the flow increases. Hence for efficient prioritization, flows need to be differentiated based on their duration. With the help of flow duration, one can differentiate old and new flows, and this flow duration information can also be used for the flow preemption purpose, (with respect to their total size). Preferring new flows ensures faster completion of small size flows as older flows may be elephant flows. This idea prevents unnecessarily waiting of small flows.

TP-SRSF uses flow duration and flow size information for scheduling purpose. In TP-SRSF every flow record its Start time, whenever a flow is started. With the help of this start time a flow can find its duration. With the scheme of TAP, TP-SRSF gives higher priority to recent flows and less priority to older flows as

$$\text{Priority}_{\text{TP-SRSF}} = \text{Priority}_{\text{TAP}} {}^{*}(\text{C} + \text{Flow duration})$$

where C is a constant and for preserving the idea of SRSF the value of C should be higher than 1.

The function used in TAP for priority calculation is a decreasing function which reduces the priority by dividing the size with a number of hops. However in case of TP-SRSF priority calculation function increases the priority by multiplying a number greater than one. Due to the increasing nature of priority function, constant (C) should be different for both elephant and mice flow. Further, we can say

$$\text{C}_{\text{elephant}} > \text{C}_{\text{mice}}$$

3.5 EVALUATION

The NS-2 simulator is used for the simulation of TAP. The performance of TAP is compared with that of pFabric which is, as per our knowledge, best flow scheduling scheme for data centers [Alizadeh et al., 2013]. The source code of pFabric was obtained from its authors.

3.5.1 Simulation Setup

Both 2-tier and 3-tier data center topologies are used for evaluating TAP. Our topologies of both 2-tier and 3-tier data center have full bisection bandwidth (non-oversubscribed). Bandwidth of links connecting hosts and ToRs is 10 Gbps. The simulated workload is already observed in the data center. The all to all communication traffic pattern is used in simulation that is mainly replication of web search workload [Alizadeh et al., 2010]. In simulations, flow arrival follows Poisson process and for each flow source and destination are chosen uniformly at random. As well with the help of load, the arrival rate of flows are changed to achieve desired level of load in the network, as in pFabric. The simulated workload is a mix of mice and elephant flows as shown in Table 3.1. In this workload, the size of 30% of flows is 1-20 MB and these flows are accounted for 95% of total bytes transferred in the network. Drop-tail queue with a maximum size of 24 packets (36KB) was used in simulations. Initial congestion window for TCP is 12 segments and minRTO is 45 microseconds. The size of mice or small flow was assumed to be less than 100 packets.

CDF	Number of packets send
0.15	6
0.20	13
0.30	19
0.40	33
0.53	53
0.60	133
0.70	667
0.80	1333
0.90	3333
0.97	6667
1.00	20000

Table 3.1. : Empirical traffic distribution

Two-tier Topology

In the simulations of 2-tier data center, 6 racks with 8 servers per rack is used. All 6 ToR switches are connected with four 2^{nd} level switches as shown in Figure 3.2.

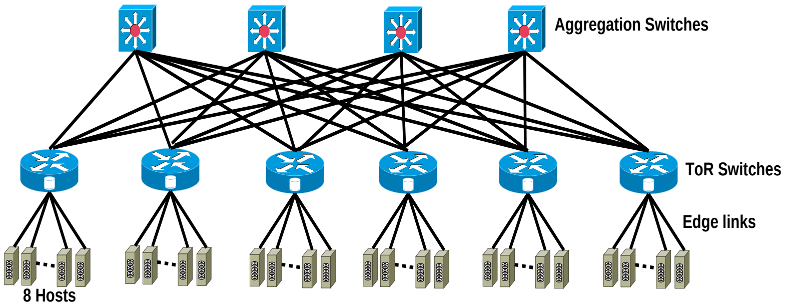


Figure 3.2. : 2-tier simulation topology for TAP

Three-tier Topology

For simulating a 3-tier data center, four such 2-tier topologies that are shown in Figure 3.2 have be combined. All 6 ToR switches are connected with six 2nd level aggregation switches. All 6 aggregation switches are connected with 4 core switches as shown in Figure 3.3. Due to hardware limitations such topology of less number of servers (8) per rack and less number of rack (6) per row and only 4 rows are taken for simulation. As NS-2 is a packet-level simulator, with a high number of servers per rack and the large number of racks and rows, the simulations of 3-tier data center became computationally expensive. However, the performance gain of TAP over pFabric would increase with the size of data center topology.

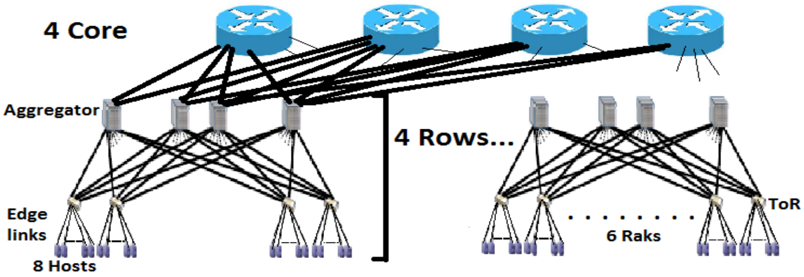


Figure 3.3. : 3-tier simulation topology for TAP

3.5.2 Simulation Results

The number of timeouts, average flow completion time and total time to complete all mice flows are the performance metrics used for validating TAP. Cumulative Distributive Function (CDF) of the completion time of mice flow is also plotted.

Two-tier Topology

Here the performance difference between TAP and pFabric is marginal due to the small size of topology (both servers per rack and the number of racks are few). With such a few servers and racks scenario, inter-rack flows are unable to gain the benefits of the proposed scheme. Hence, the performance of TAP is same as that of pFabric, as can be observed in Figure 3.4. However, TAP shows better performance over pFabric in terms of the number of timeouts and the total flow duration can be observed from Figure 3.5 and Figure 3.6, respectively. In Figure 3.7, one can see that the average throughput of elephant flows is the same for TAP and pFabric. CDF of flow completion time of mice flows are also analyzed. As illustrated in Figure 3.8, as load increases the number of flows completed within the deadline for TAP is more than that of pFabric. In other words, TAP flows are completed sooner than pFabric flows and reduces the long-tail of flow completion time as well.

Three-tier Topology

In a 3-tier data center, the improvement of performance of TAP over pFabric is clearly visible. As load increases, the total number of timeouts for mice flows are reduced, and can be noticed in Figure 3.10. Further the average flow completion time is shown in Figure 3.9 and total time to complete all mice flows illustrated in Figure 3.11 are also reduced. While TAP shows performance gain in terms of reduced flow completion time and reduced number of timeouts, it is seen to exhibit a marginal reduction in throughput for elephant flows at higher load as indicated in Figure 3.12.

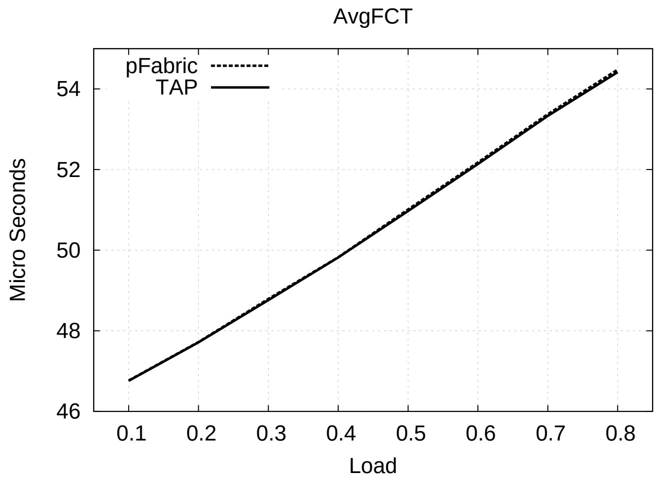


Figure 3.4. : 2-tier: Average flow completion time

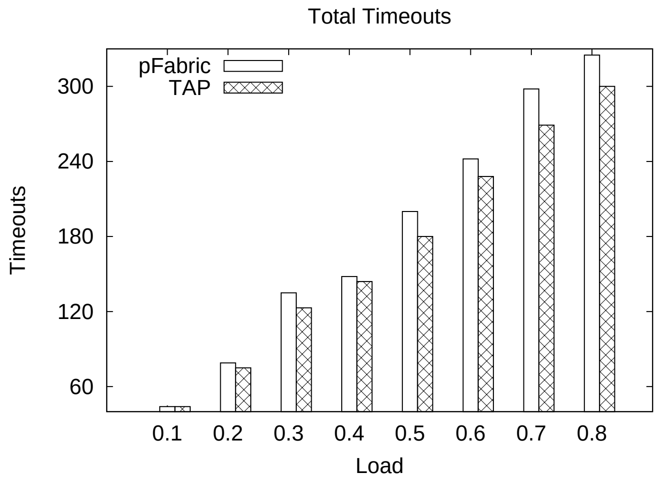


Figure 3.5. : 2-tier: Total timeout events

CDF graphs manifest that TAP is more beneficial in 3-tier compared to the 2-tier data center. Figures 3.13(d) and 3.8(d) confirm significant improvement of 3-tier over 2-tier at load 0.8. Since an increase in the size of a data center (number of hosts, rack, row) causes more long distance flows, this recommends TAP over pFabric.

3.5.3 TP-SRSF

For validating this idea, P-SRSF (Preemptive SRSF), TP-SRSF (Topology aware Preemptive SRSF) are implemented and compared with SRSF (pFabric) and T-SRSF (Topology aware SRSF (TAP)) scheduling. The difference between P-SRSF and TP-SRSF is about topology awareness, P-SRSF calculates priority on top of SRSF, whereas TP-SRSF apply on top of T-SRSF. In these simulations, P-SRSF uses single constant (C) for mice and elephant flows. P-SRSF is simu-

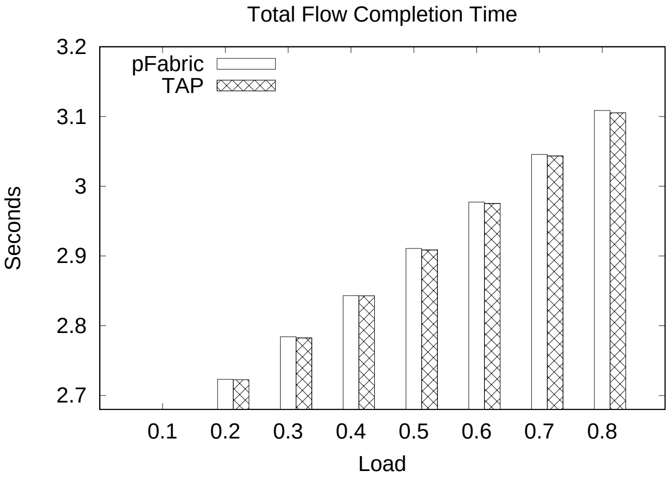


Figure 3.6. : 2-tier: Total flow duration

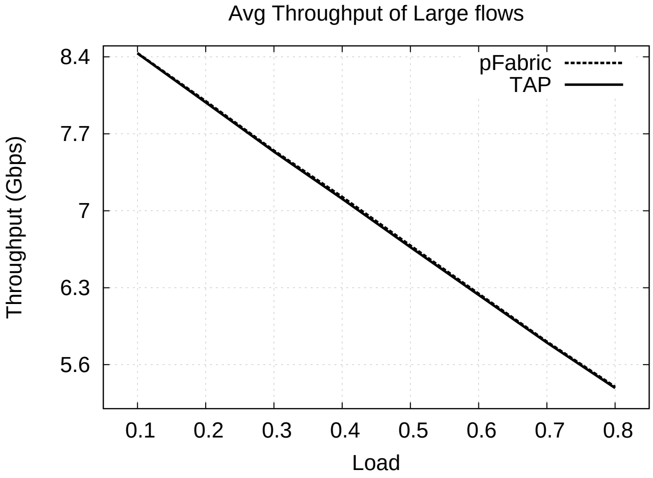
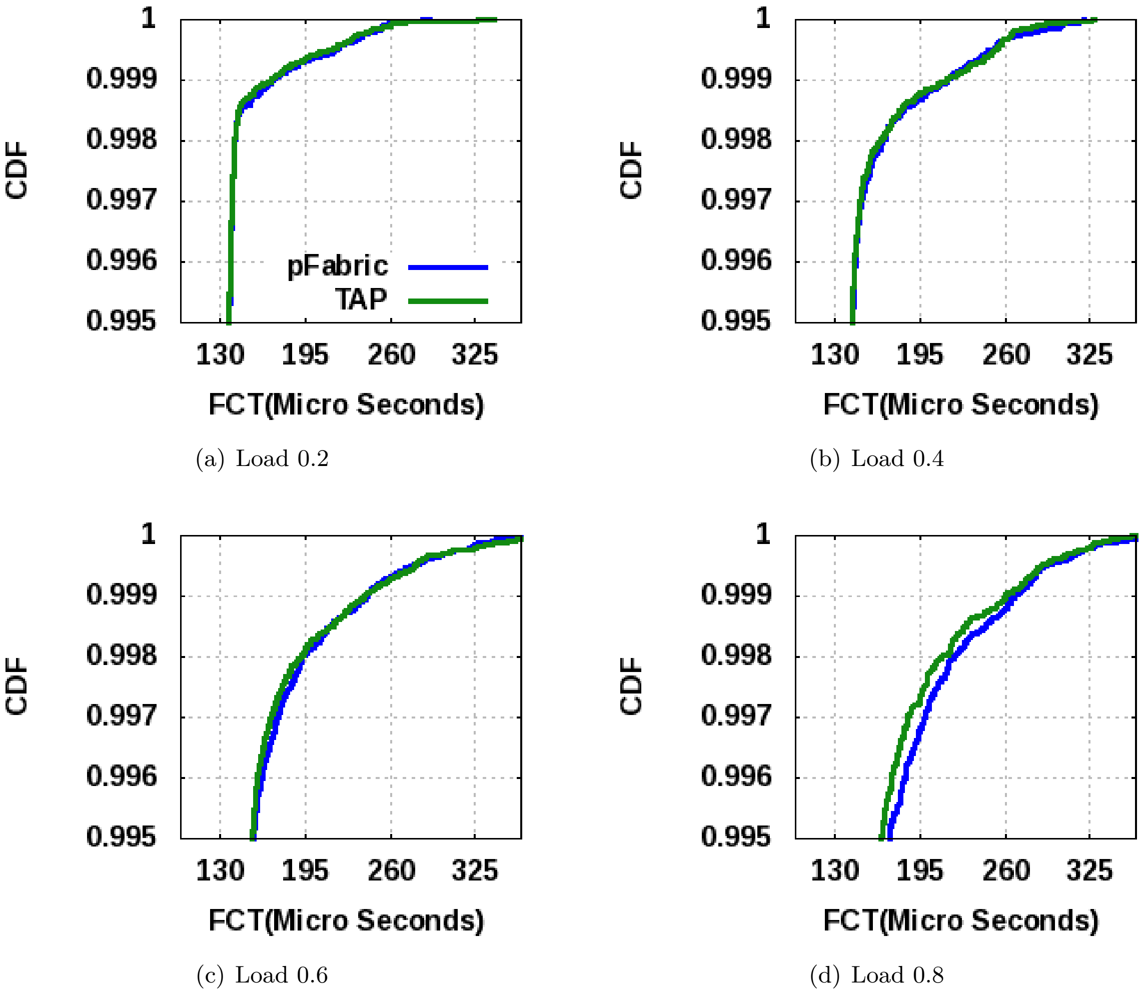


Figure 3.7. : 2-tier: Average throughput of large flows

lated for two different values of C (1,10). On the other side TP-STSF uses two values of constant $C_{mice} = 1$ and $C_{elephant} = 10$.

In the simulated topology for TP-SRSF idea, there are total 144 nodes. This topology uses 9 racks with 16 servers per rack. All 9 ToR switches are connected with 4 Aggregation switches as shown in Figure 3.14. Here the performance difference between SRSF and TP-SRSF is visible. But the gain of P-SRSF is marginal and not clearly visible in Figure 3.15. However, TP-SRSF shows better performance over T-SRSF and SRSF in terms of the number of timeouts and the total flow duration which can be observed from Figure 3.16 and Figure 3.17, respectively. In Figure 3.18, one can see that the average throughput of elephant flows is reduced for TP-SRSF compared to SRSF. As illustrated in Figure 3.20, as load increases the number of flows completed within the deadline for TP-SRSF is more than that of SRSF.



(a) Load 0.2

(b) Load 0.4

(c) Load 0.6

(d) Load 0.8

Figure 3.8. : CDF of 99.5th percentile flow completion time of 2-tier data center

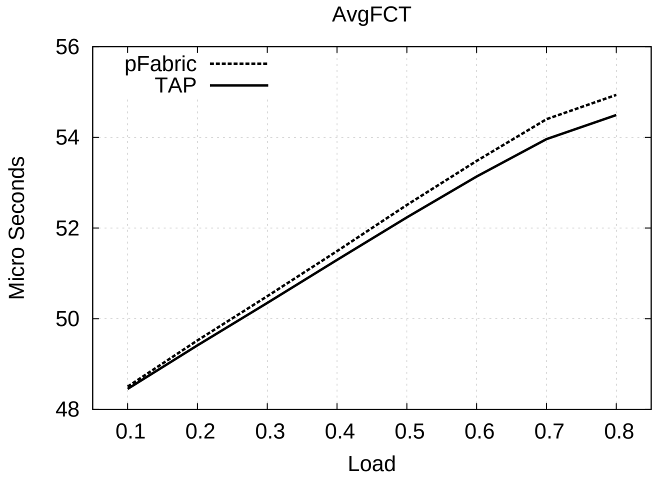


Figure 3.9. : 3-tier: Average flow completion time

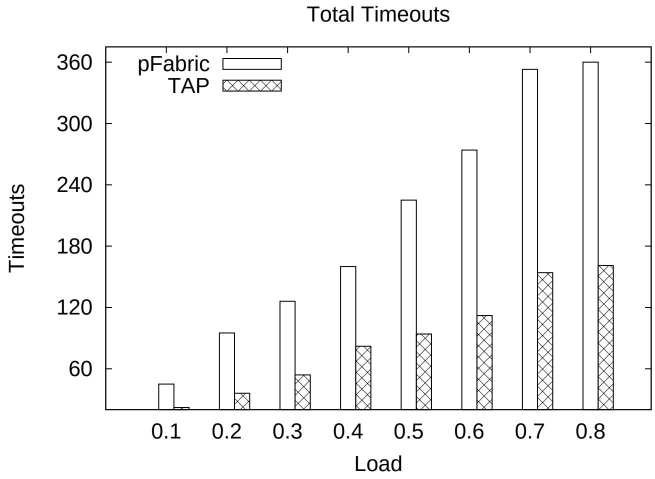


Figure 3.10. : 3-tier: Total timeout events

Figure 3.19 shows the benefit of preferring new flows over old flows. P-SRST achieves a marginal benefit over SRSF by ensuring small flow completion time. CDF graphs manifest that TP-SRSF is more beneficial than SRSF. Figure 3.20(c) confirms significant improvement at load 0.6. Since an increase in the size of a data center (number of hosts, rack, row) causes more long distance flows, this recommends TP-SRSF over SRSF.

3.6 SUMMARY

This chapter proposes TAP and TP-SRSF, which improve the priority scheme of pFabric (SRSF). The improvised priority scheme is not only based on the remaining flow size but also depends on the distance in terms of hop length. In all to all communication scenario, the number of long distance flows is more compared to the short distance flows. This difference grows as the size

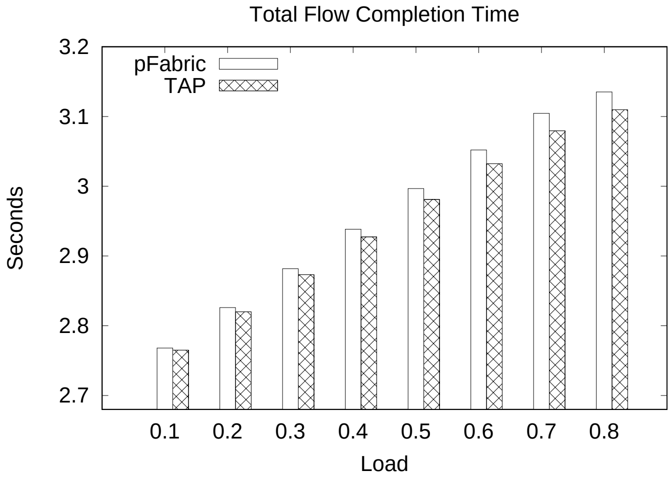


Figure 3.11. : 3-tier: Total flow duration

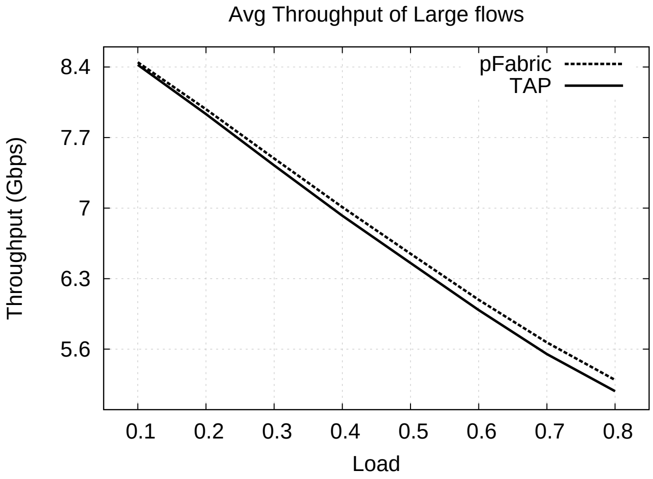
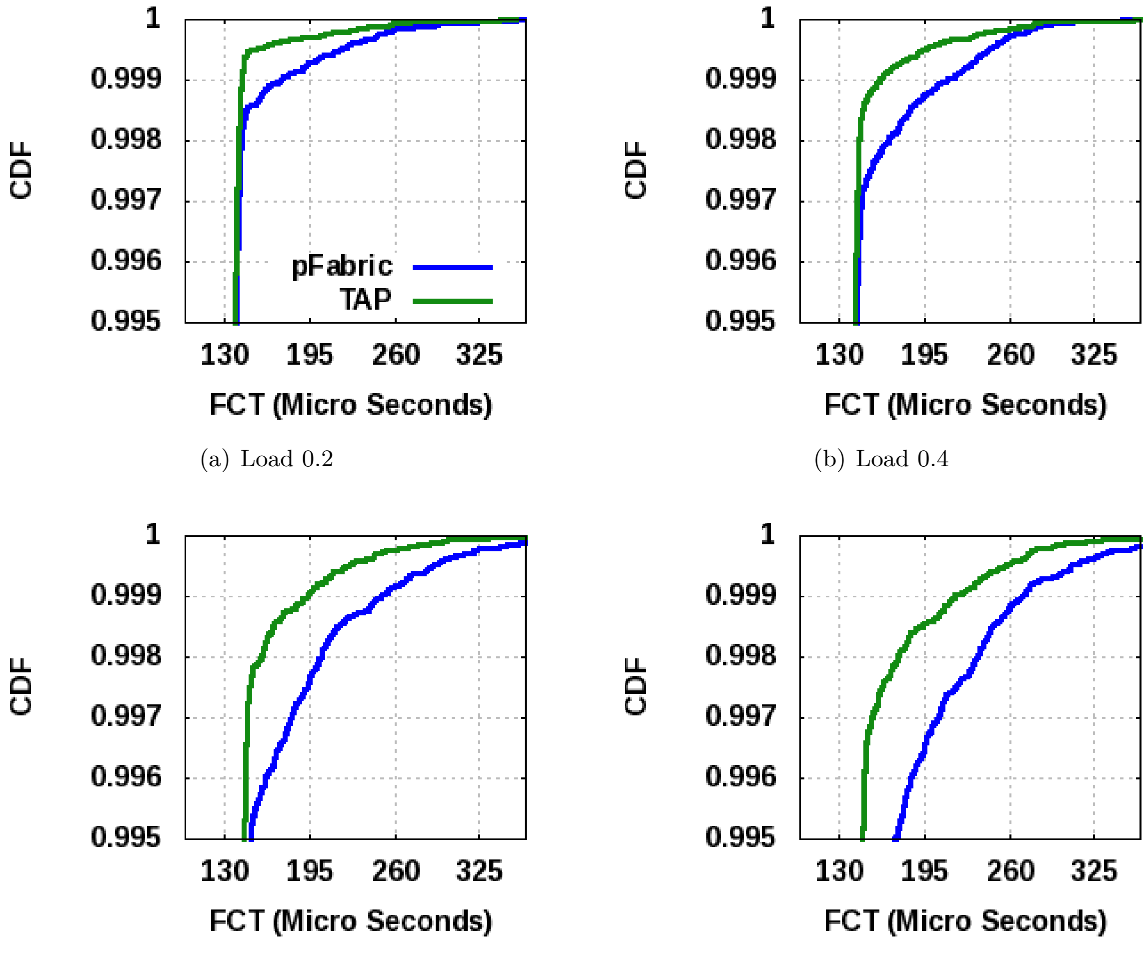


Figure 3.12. : 3-tier: Average throughput of large flows

of data center increases. Since TAP and TP-SRSF are independent of topology structure, which giving high priority to long distance flows and new flows in any structure would ensure that majority of flows get benefited. Simulation results encourage Topology Aware prioritization to improve the performance of data center.

...



(a) Load 0.2

(b) Load 0.4

(c) Load 0.6

(d) Load 0.8

Figure 3.13. : CDF of 99.5th percentile flow completion time of 3-tier data center

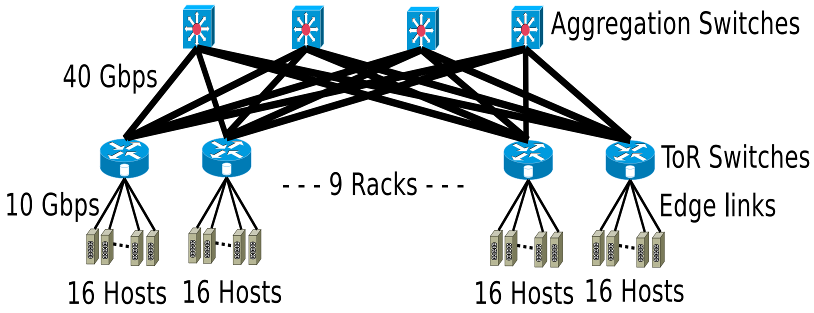


Figure 3.14. : Simulation topology for TP-SRSF

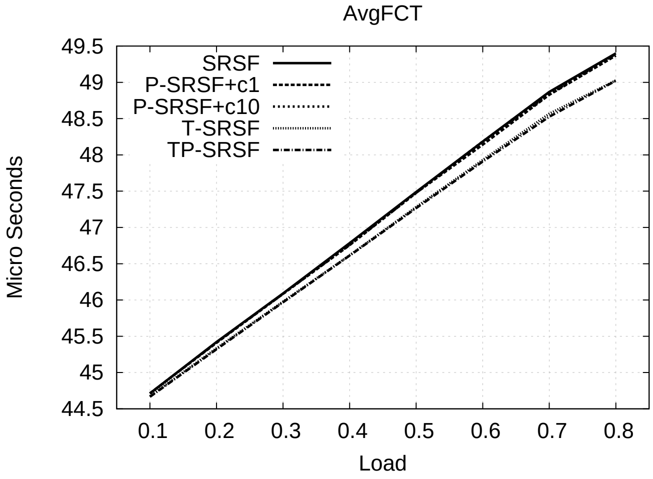


Figure 3.15. : Average flow completion time for TP-SRSF

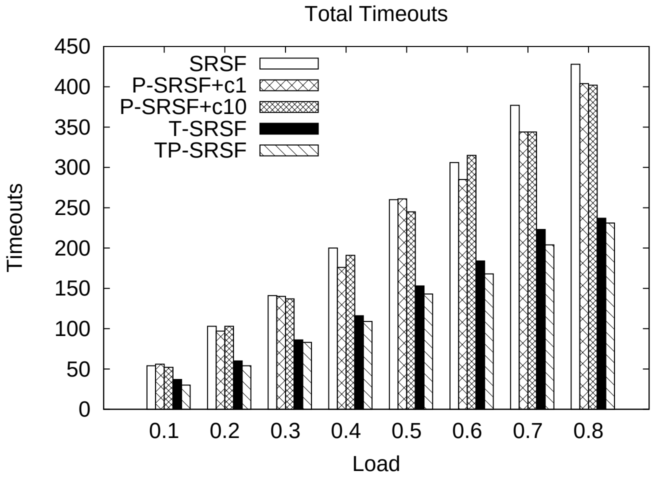


Figure 3.16. : Total timeout events for TP-SRSF

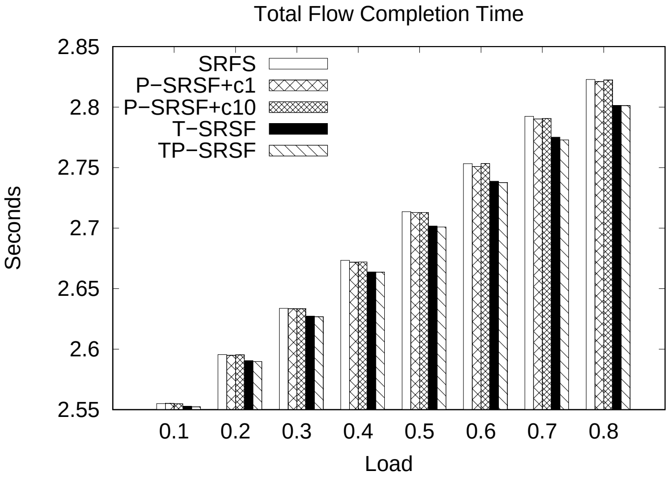


Figure 3.17. : Total flow duration for TP-SRSF

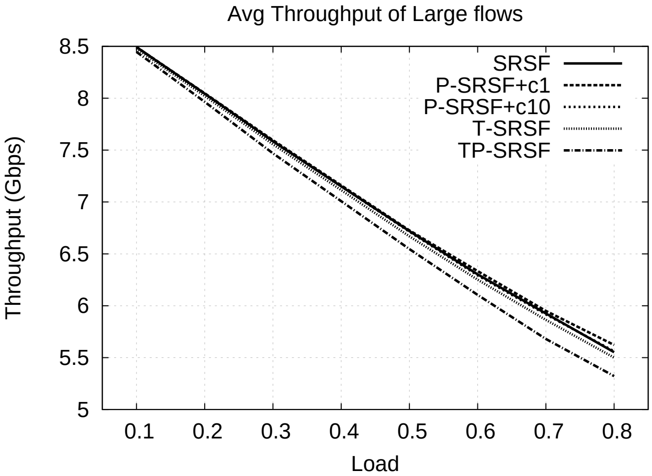


Figure 3.18. : Average throughput of large flows for TP-SRSF

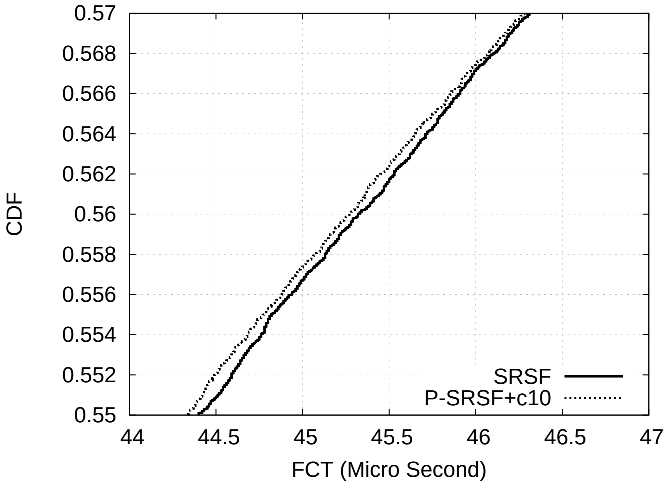
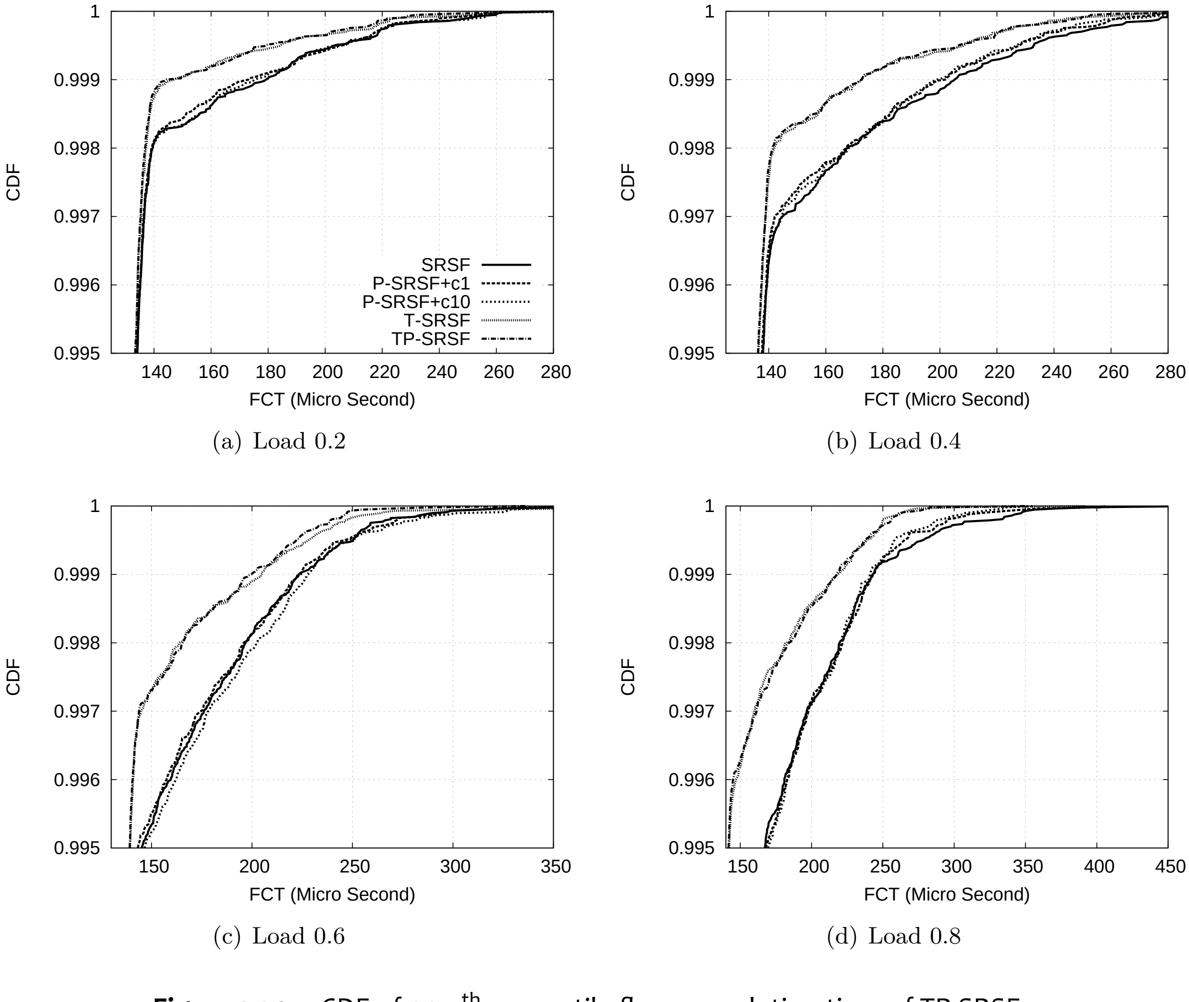


Figure 3.19. : Effect of flow preemption



(a) Load 0.2

(b) Load 0.4

(c) Load 0.6

(d) Load 0.8

Figure 3.20. : CDF of 99.5^{th} percentile flow completion time of TP-SRSF