Additional Appendix

1 Queuing Analysis

Table 1: Important Symbol

Symbol	Description
n	the number of hosts per rack
$K_1 + 1$	the number of ToR switches
K_2	the number of core switches
i	Hop i of a path, indexing from 0
BDP	base BDP of the network
BDP_i	BDP of hop i
r_i	egress port rate
win_i	the initial window
b_i	maximum buffer occupancy of data
d_i	the duration of output port transmission when buffer reaches maximum

In this section, we analyze the maximum buffer occupancy of DCQCN/TIMELY/HPCC with or without Floodgate. When a flow arrives, the sender transmits up to one BDP worth of packets at line rate before congestion control taking effect1. The symbols we used are listed in Table 1. We focus on a $K_1 * n$-scale incast where each host except for which connected to the same ToR with the destination host transmits a BDP-sized flow simultaneously to a destination host. The topology we used is a two-tier non-blocking clos-network topology, as shown in Figure 1. Assume the traffic is well balanced to all equivalent egress ports, *i.e.*, each core switch receives the same amount of data.

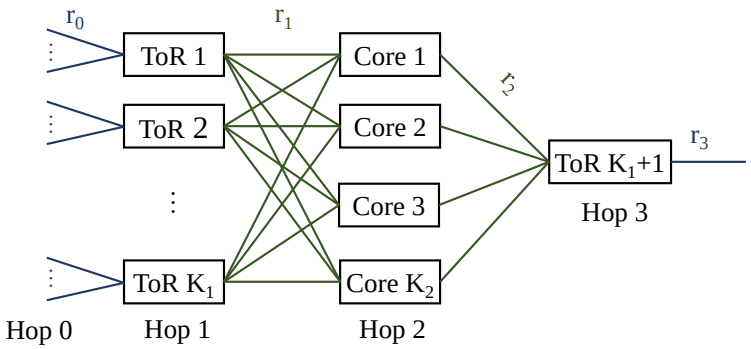

Figure 1: A leaf-spine topology with K_2 core switches and $K_1 + 1$ ToR switches. Each ToR switch connects n hosts. We assume all hosts under ToR 1, ToR 2, $\cdots$, ToR K_1 send incast flows to a host under ToR $K_1 + 1$, thus we expand the topology according to the hop a packet passes by to make a better understanding.

Theorem 1. *The maximum buffer occupancy of DCQCN/ TIMELY/ HPCC is proportional to the number of hosts in network when a $K_1 * n$-scale incast occurs.*

Answer. A number of $K_1 \times n$ hosts send a flow simultaneously at line rate to a single host. Because the topology is non-blocking, packets do not accumulate at the source ToR. Meanwhile, data packets accumulate at core switches and destination ToR. Each core switch receives $K_1 \times BDP \times n/K_2$ packets, at rate $K_1 \times r_1$. The sending rate is $r_2 = r_1$, *i.e.*, the egress port bandwidth of core switches. Because the receiving rate is at least equal to the sending rate when $K_1 \geq 2$, the buffer occupancy of core switches continues increasing until the receiving duration ends. The maximum buffer occupancy of core switch is,

$$b_2 = K_1 \times BDP \times n/K_2(1 - r_2/(K_1 \times r_1)) \quad (1)$$

$$= BDP \times (K_1 - 1)/K_2 \times n \quad (2)$$

For the destination ToR, it receives $K_1 \times BDP \times n$ packets. The destination ToR receives data at rate $K_2 \times r_2$ while

the output rate is r_3. The maximum buffer occupancy of destination ToR is,

$$b_3 = K_1 \times BDP \times n(1 - r_3/(K_2 \times r_2)) \tag{3}$$
$$= BDP \times K_1 \times (n-1) \tag{4}$$

$K_1 \times (n-1)$ is proportional to the number of hosts, *i.e.*, $(K_1+1) \times n$, in network. Therefore, the maximum buffer occupancy is proportional to the number of hosts in network under a $K_1 * n$-scale incast. □

Theorem 2. *The maximum buffer occupancy of DCQCN/ TIMELY/ HPCC + Floodgate is proportional to the number of core switches when a $K_1 * n$-scale incast occurs.*

Answer. The ToR switch connected to the source hosts receives $n \times BDP$ packets after incast flows start. Each ToR switches only transmits win_1 packets to its downstream switch in the first RTT because of the per-dst window. The maximum buffer occupancy of the source ToR is,

$$b_1 = BDP \times n - win_1 \tag{5}$$

For a core switch, it aggregates the data packets from its upstream ToR switches and transmits packets win_2 at rate r_2. The maximum buffer occupancy of the core switch is,

$$b_2 = win_1 \times K_1/K_2 - d_2 \times r_2 \tag{6}$$
$$\leq win_1 \times K_1/K_2 \tag{7}$$

where d_2 denotes the duration of output port transmission time of core switches when the buffer occupancy reaches the maximum value.

$$d_2 = min\{win_1/(K_2 \times r_1), win_2/r_2\} \tag{8}$$
$$= min\{win_1/(K_2 \times r_1), win_2/r_1\} \tag{9}$$

For a destination ToR, it receives the data packets from its upstream core switches and transmits packets at the rate of r_3. The maximum buffer occupancy of the destination ToR switch is,

$$b_3 = win_2 \times K_2(1 - r_3/(K_2 \times r_2)) \tag{10}$$
$$= win_2 \times K_2(1 - 1/n) \tag{11}$$
$$\leq win_2 \times K_2 \tag{12}$$

Rather than proportional to the number of hosts, the maximum buffer occupancy of Floodgate is proportional to,

$$b_i \propto max\{BDP \times n, win_1 \times K_1/K_2, win_2 \times K_2\} \tag{13}$$

Therefore, Floodgate reduces the buffer occupancy by an order of magnitude. When the topology scales up, *i.e.*, $n \ll K_1, k_1/K_2 \ll K_2$,the maximum buffer occupancy of Floodgate is proportional to the number of core switches in network.

□

2 Bandwidth Utilization Analysis

In this section, we focus on the bandwidth utilization of Floodgate. In the beginning, we define *perfect load balance*.

Definition 1. *(Perfect Load Balance). A switch that achieves perfect load balance means that traffic is equally split to the next-hop switch.*

Because traffic is equally split to the next-hop switch, among the same tier switches (except for ToR), buffer occupancy is the same. (ToR switches are not taken into consideration because load balance cannot solve the imbalance of traffic injection.) Therefore, in a perfect load-balanced network, flows sharing the same source and destination pair pass through the same congestion point, *i.e.*, destination ToR, or different points with the same degree of congestion, *i.e.*, core switches.

Theorem 3. *In symmetric topology, with Perfect Load Balance, the bandwidth is fully utilized when switches initialize the per-dst window to one* BDP_i.

Answer. To prove the necessity of perfect load balance, we prove the inverse negative propositions of Theorem 3, *i.e.*, when load balance is not perfect, bandwidth waste can occur. Considering flows arrives at a two-tier non-blocking clos-network, as shown in Figure 1, where $K_1 + 1 = K_2 = 2$, and $r_0 = r_1$. When load balance is not perfect, the buffer occupancy of core switches is $0 \leq b_2(core_0) < b_2(core_1)$, respectively. During every RTT, each core switch transmits BDP_2 data packets to destination ToR switches and then each receives BDP_2 credits. The bandwidth of ToR switches can be fully utilized until $0 = b_2(core_0) < b_2(core_1)$. In the forthcoming time, only one core switch transmits one BDP_2 packets to two destination ToR switches per RTT. Therefore, the bandwidth of ToR switches is under-utilized.

Now we prove the sufficiency of perfect load balance. (i) The network has no buffer occupancy. Floodgate's switch has nothing to do. Apparently, no bandwidth waste occurs. (ii) The network has buffer occupancy. According to the definition of perfect load balance, buffer occupancy on core switches are the same, *i.e.*, $0 < b_2(core_0) = b_2(core_1)$. Therefore, in every moment, the network is fully utilized until $0 = b_2(core_0) = b_2(core_1)$. It degrades to the condition in (i). By proving the necessity and sufficiency of perfect load balance, Theorem 3 is proved. □ In Floodgate, the initial per-dst window is $BDPi + C_{out} * T$, much larger than BDP_i. Therefore, the bandwidth can be fully utilized with Floodgate's per-dst window.