Absorbing Micro-burst Traffic by Enhancing Dynamic Threshold Policy of Data Center Switches

Danfeng Shan, Wanchun Jiang, and Fengyuan Ren

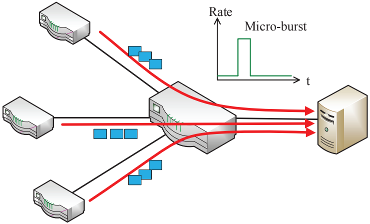
Tsinghua University

INFOCOM 2015

Outline

Micro-burst

- Micro-burst is a common traffic pattern in data center networks.
  - "myths about microbursts," White Paper, Arista.
  - "Efficiently measuring bandwidth at all times scales," NSDI 2011
  - ...
- It usually appears in the switch when packets from multiple concurrent flows are destined to the same output port.



Micro-burst

- Micro-burst is a common traffic pattern in data center networks.
    - "myths about microbursts," White Paper, Arista.
    - "Efficiently measuring bandwidth at all times scales," NSDI 2011
    - ...
- It usually appears in the switch when packets from multiple concurrent flows are destined to the same output port.
- Packet dropping caused by micro-burst is unacceptable
    - Micro-burst is comprised of several delay-sensitive short flows.
    - Timeout triggered by packet dropping extends the flow completion time.

Buffer management policy in switch

- Packet dropping in a switch is directly related to the buffer architecture and buffer management policy.
- Buffer architecture: the majority of switches employ the on-chip shared memory.
    - The on-chip packet buffer is dynamically shared across ports by statistical multiplexing
    - Fairness problem: few output ports could occupy all of the shared buffer, starving other output ports.

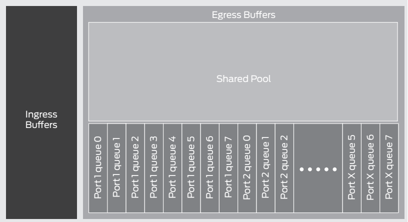


Figure: Shared buffer architecture for Juniper EX2200/EX3200/EX4200 switches

Buffer management policy in switch

- Packet dropping in a switch is directly related to the buffer architecture and buffer management policy.
- Buffer architecture: the majority of switches employ the on-chip shared memory.
  - The on-chip packet buffer is dynamically shared across ports by statistical multiplexing
  - Fairness problem: few output ports could occupy all of the shared buffer, starving other output ports.
- Buffer management policy: *Dynamic Threshold* (DT) has been widely used by switch vendors
  - "Broadcom smart-buffer technology in data center switches for cost-effective performance scaling of cloud applications," White Paper, Broadcom, Apr. 2012.
  - "Congestion management and buffering in data center networks," White Paper, Extreme Networks, Dec. 2013.
  - ......

Dynamic Threshold Policy

- Mechanism of DT
  - The queue length is restricted by a threshold.
  - The threshold is proportional to the current amount of free buffer space.

Formulation

$$T(t) = \alpha \cdot \left(B - \sum_i Q_i(t) \right)$$

$T(t)$: threshold α: a parameter
B: buffer size $Q_i(t)$: queue length of output port i

- Problem of DT
  - When micro-burst occurs in switches employ DT policy, packets from micro-burst traffic are dropped even when there is free buffer space in the switch.

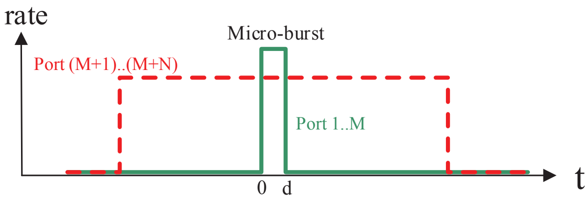
Target

1. Theoretically deduce the sufficient conditions for packet dropping caused by micro-burst traffic
2. Quantitively estimate the corresponding free buffer size in DT switches

Assumptions

Assumptions

1. At time 0, Queue lengths of port $1, \cdots, M$ are empty
2. At time 0, port $(M+1), \cdots, (M+N)$ have reached their steady states
3. At time 0^+, port $1, \cdots, M$ begin to transmit micro-burst traffic



$R_i (i = 1, \cdots, M)$ is constant and $R_1 = R_2 = \cdots = R_M = R$

At time $t = 0^+$,

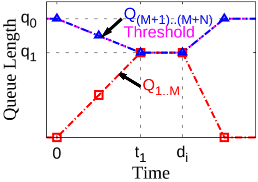
- the micro-burst traffic arrived at port $1, \cdots, M$
- Queue length of port $1, \cdots, M$ will increase
- Meanwhile, the unused buffer is occupied
- The threshold will decrease
- Queue length of port $(M + 1), \cdots, (M + N)$ will decrease

Two cases (Since the maximum decreasing rate of queue length is C):

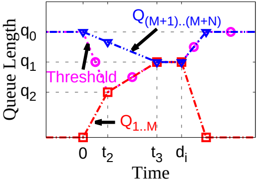
1. Threshold decreases at a rate lower than C
   - Queue length decreases at the same rate as the threshold
2. Threshold decreases at a rate greater than C
   - Queue length decrease at a rate of C

$R_i (i = 1, \cdots, M)$ is constant and $R_1 = R_2 = \cdots = R_M = R$ (Cont.)

- Packet dropping happens at $[t_1, d_i]$ and $[t_2, d_i]$
- $t_1 = \frac{\alpha B}{[1+\alpha(M+N)](R-C)}$, $t_2 = \frac{\alpha B}{(1+\alpha N)[(1+\alpha M)(R-C)-\alpha NC]}$



(a) $R \leqslant C\left(1+\frac{1+\alpha N}{\alpha M}\right)$



(b) $R > C\left(1+\frac{1+\alpha N}{\alpha M}\right)$

Notions

d_i: Duration of micro-burst traffic in i-th port
Q_i: The queue length of port i
R: Arriving rate of micro-burst traffic
α: Parameter
C: Link capacity
B: Buffer size
M and N: Constant value

Sufficient condition and free buffer size in this case

Theorem

When $R_1 = R_2 = \cdots = R_M = R$, the packets from micro-burst traffic will be dropped in port k $(k = 1, 2, \cdots, M)$ if

$$d_k \geqslant \begin{cases} \frac{\alpha B}{[1+\alpha(M+N)](R-C)}, \textit{if } R \leqslant C\left(1 + \frac{1+\alpha N}{\alpha M}\right) \\ \frac{\alpha B}{(1+\alpha N)[(1+\alpha M)(R-C)-\alpha NC]}, \\ \textit{if } R > C\left(1 + \frac{1+\alpha N}{\alpha M}\right) \end{cases}$$

and the free buffer size while packets are dropped is

$$F = \begin{cases} \frac{B}{1+\alpha(M+N)}, \textit{if } R \leqslant C\left(1 + \frac{1+\alpha N}{\alpha M}\right) \\ \frac{(R-C)B}{(1+\alpha N)[(1+\alpha M)(R-C)-\alpha NC]}, \\ \textit{if } R > C\left(1 + \frac{1+\alpha N}{\alpha M}\right) \end{cases}$$

Remark 1: Why micro-burst is easier to cause packet dropping?

$$d_k \geqslant \frac{\alpha B}{[1+\alpha(M+N)](R-C)} \Rightarrow R \cdot d_k - C \cdot d_k \geqslant \frac{\alpha B}{1+\alpha(M+N)}$$

If the micro-burst traffic size (i.e., $R \cdot d_k$) is fixed, then the condition can be easier to be satisfied for smaller d_k or larger R (d_k: duration of micro-burst in port k R: arriving rate of micro-burst traffic)

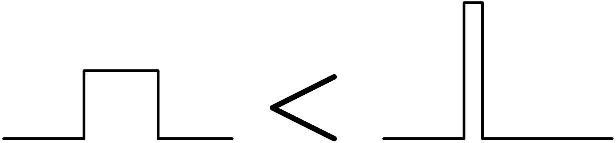


Figure: Micro-burst traffic is easier to cause packet dropping than smooth traffic

Remark 2: The free buffer size when packets are dropped

Free buffer size when $R \leqslant C\left(1+\frac{1+\alpha N}{\alpha M}\right)$: $F=\frac{B}{1+\alpha(M+N)}$

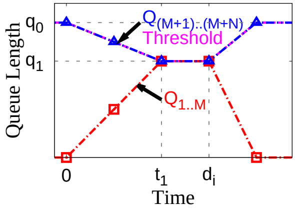
- The free buffer size is negatively related to the number of overloaded ports (i.e., $M+N$)
- When the number of overloaded ports is small, the free buffer size would be very large
  - $M+N=1, \alpha=1$, then free buffer size is $B/2$

Remark 2: The free buffer size when packets are dropped (Cont.)

- Why DT reserve this fraction of buffer?
  - Provides a cushion for newly overloaded ports (prevent starving)
  - Notify DT to change the threshold
- However, this fraction of buffer should be utilized when a port's transmitting micro-burst traffic
  - The time-scale of micro-burst traffic is quit short
  - Reserved buffer are occupied for only a short time
  - This is worthwhile since this can help absorb micro-burst traffic
  - The actions that a packet enter into and departs from the buffer can be used to inform DT of adjusting threshold instead

Remark 3: Fairness constraint of DT

- When packets from micro-burst traffic are dropped?
  - The queue length of newly overloaded ports reach the queue length of other ports
- Why packets are dropped at that time?
  - To ensure fair buffer sharing among overloaded ports
- However,
  - Avoiding packet dropping of micro-burst is of great importance
  - Allocating more buffer for micro-burst traffic has few effects since the micro-burst duration is very short



$R_i (i = 1, \cdots, M)$ is constant and $R_1 \geqslant R_2 \geqslant \cdots \geqslant R_M = R$

Sufficient condition and free buffer size in case 1:

Theorem

When $\sum_{i=1}^{M}(R_i - C) \leqslant \frac{(1+\alpha N)C}{\alpha}$*, packets will be dropped in port* k $(k = 1, 2, \cdots, M)$ *if*

$$d_k \geqslant t_k \quad (1)$$

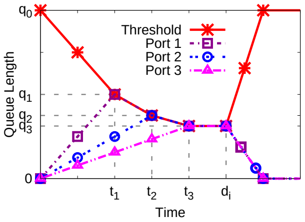
where

$$\begin{cases} t_k = \dfrac{\alpha\left[F_{k-1} + \alpha F_{k-1}(N+k-1) + G_k t_{k-1}\right]}{(R_k - C)[1+\alpha(N+k-1)] + \alpha G_k} \\ F_k = F_{k-1} - \dfrac{G_k(t_k - t_{k-1})}{1+\alpha(N+k-1)} \\ G_k = \sum_{i=k}^{M}(R_i - C) \end{cases} \quad (2)$$

Proof: Using mathematical induction

Basic idea (Details are omitted):

- Basis: proof that the theorem holds for port 1 (i.e., $k = 1$)
- Inductive step
  - Assume the theorem holds for port i (i.e., $k = i$)
  - When port i reaches the threshold, there are $N + i$ ports whose queue lengths are decreasing
  - Let $N_i = N + i$. Then following the same way as *Basis*, we can deduce the packet dropping time for port $i + 1$ and corresponding free buffer size.



$R_i (i = 1, \cdots, M)$ is constant and $R_1 \geqslant R_2 \geqslant \cdots \geqslant R_M = R$

Sufficient condition and free buffer size in case 2:

Theorem

When $\sum_{i=1}^{M}(R_i - C) > \frac{(1+\alpha N)C}{\alpha}$*, packets in port* k $(k = 1, 2, \cdots, L)$ *will be dropped if*

$$d_k \geqslant t_k \quad (3)$$

where

$$\begin{cases} t_k = \dfrac{\alpha\{F_{k-1} + [G_k - (N+k-1)C]t_{k-1}\}}{\alpha[G_k - (N+k-1)C] + R_k - C}, \\ F_k = F_{k-1} - [G_k - (N+k-1)C](t_k - t_{k-1}), \\ G_k = \sum_{i=k}^{M}(R_i - C) \end{cases} \quad (4)$$

L *is the largest* k *such that* $G_k > \frac{(1+\alpha N_k)C}{\alpha}$ *and* $L \leqslant M$*.*

DT can be improved to absorb micro-burst traffic

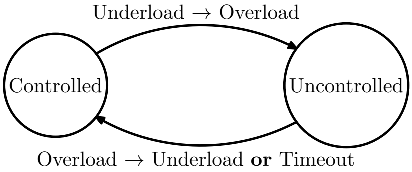
To absorb micro-burst traffic

1. The switch buffer should be fully utilized
2. The fairness constraint of DT should be temporarily relaxed

Overview of EDT (Enhanced Dynamic Threshold) policy

Allows an output port to aggressively occupy buffer in a relatively short interval when the port becomes overloaded

- Each port has two states: Controlled and Uncontrolled
- In the controlled state, the port threshold is determined by DT
- In the uncontrolled state, the port threshold is temporarily set to the buffer size
- Controlled to Uncontrolled state: when the port becomes overloaded
- Uncontrolled to controlled state:
  - Micro-burst traffic: when the port becomes underloaded
  - Long-lived traffic: after a specified time

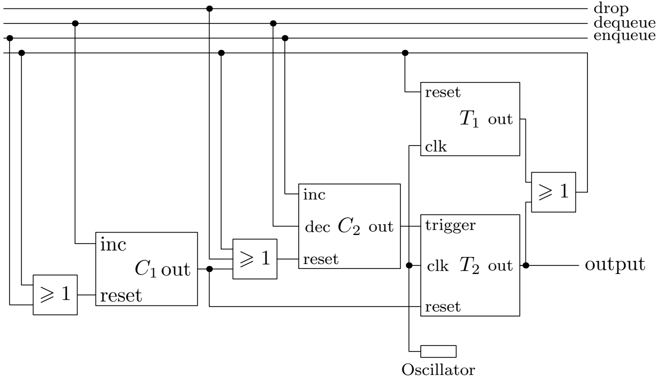


Benefits

1. Packets are dropped only when it is inevitable when micro-burst traffic arrives
2. Buffer can be fairly shared among output ports transmitting long-lived flows
   - The period over which EDT stays in uncontrolled state is short
3. EDT is simple enough to be implemented in high-speed switches
   - It only requires several additional timers and counters

Circuit diagram of EDT

EDT can be implemented by several timers and counters.

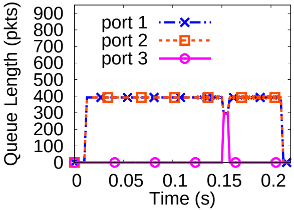


Main components

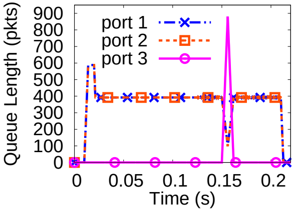
- Counter 1: identifying that the output port returns to the underloaded state (from overloaded state)
- Counter 2: identifying that the output port becomes overloaded
- Timer 1: making sure that the stat transition happens only when bursty traffic arrives
- Timer 2: controlling the period over which EDT stays in uncontrolled state

Evolutions of queue lengths when $N = 2,\ M = 1$

- For DT, packets are dropped immediately after the arriving of micro-burst traffic
- For EDT, the micro-burst traffic are absorbed



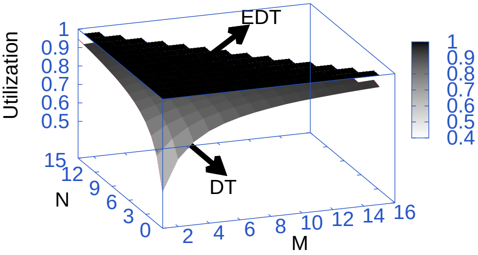
(a) DT



(b) EDT

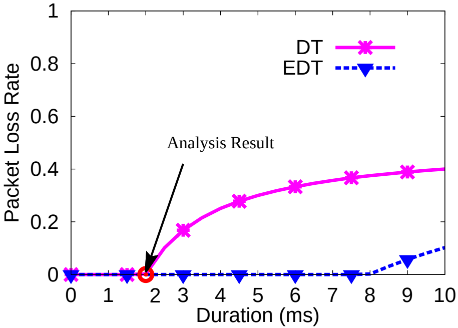
Buffer utilization when packets from micro-burst traffic are dropped

- For DT, buffer utilization is low when M and N is small
- For EDT, buffer is fully used (Packets are dropped only when it is inevitable).



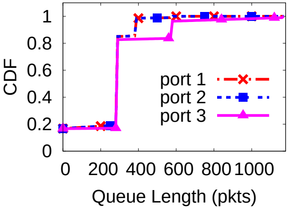
Packet loss rate as a function of its duration

- For DT switches, packets are dropped when duration of micro-burst traffic is 2ms
- For EDT switches, packets are dropped when duration of micro-burst traffic is 8ms

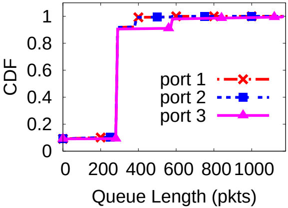


Fairness among ports transmitting long-lived flows

- Scenario: port 1 and port 2 have reached their steady state while port 3 becomes overloaded
- Result: EDT can promise fairness among ports transmitting long-lived flows



(c) duration: 500ms



(d) duration: 1000ms

Figure: Queue length CDFs with different durations of long-lived flows

Conclusion

- In this paper, we
  - theoretically deduce the sufficient conditions for packet dropping caused micro-burst traffic
  - quantitively estimate the corresponding free buffer size
- According to the analysis, we find that to absorb micro-burst traffic
  - the switch buffer should be fully utilized
  - the fairness constraint of DT should be temporarily relaxed
- Therefore, we designed the EDT policy, which can absorb micro-burst traffic as much as possible