Sizing Router Buffers for Large-Scale TCP/IP Networks

Hiroyuki Hisamatsu

Osaka Electro-Communication University

Transmission Control Protocol

- TCP performance is largely affected by
  - Round-trip time (RTT) and packet loss ratio
- Router buffer size affects RTT and packet loss ratio
  - Packets can be accumulated at the buffer and cause a queuing delay and packet losses
- Utilizing a large buffer
  - Packet loss ratio can be reduced
  - However, can cause a large queuing delay
    - Because a large # of packets are accumulated

Traditional Buffer Sizing

bandwidth-delay product

- Buffer size is traditionally determined by $C \times \overline{RTT}$

link bandwidth | average round-trip time

- Constructing a router buffer based on traditional buffer sizing is difficult
  - Reason: Hardware limitations following tremendous increase of link bandwidth
  - e.x.: Average RTT = 250 [ms] & link bandwidth = 10 [Gbit/s] -> traditional buffer size = 2.5 [Gbit]
- Building a router buffer for future high-speed network based on traditional buffer sizing is extremely difficult

Small Buffer Sizing [1]

of TCP connections pass through the output link

- Buffer sizing discipline for Internet core routers
- Buffer size can be reduced to $C \times \overline{RTT} / \sqrt{n}$
  - e.x.: n = 10,000,small buffer size = 1/100 of traditional buffer size
- Link utilization achieves over 98%
  - Assume # of TCP connection is sufficiently large and TCP connections are desynchronized

[1] G. Appenzeller, I. Keslassy, and N. McKeown, "Sizing router buffers," in Proceedings of ACM SIGCOMM, pp. 281–292, Sept. 2004.

Related Work

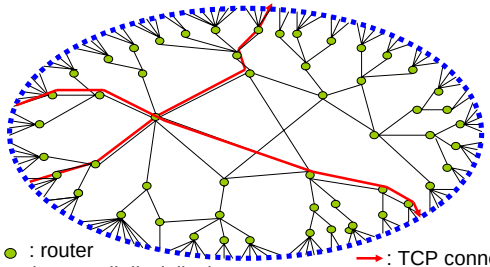
- Focus almost exclusively on link utilization
  - Increase in packet loss ratio is not considered
  - Influence on TCP performance is not considered
- Focus on TCP throughput, however,
  - Network topology is quite simple
  - Small TCP connections (400) pass through the router
    - Small buffer sizing needs at least 500 TCP [1]

Effect of small buffer sizing is evaluated in the small-scale network

Objective

- Investigate the validity of small buffer size in a large-scale network including core and edge networks
  1. Devise a method of analyzing the average TCP behavior in the network with over 100/1,000/10,000 routers/endhosts/links
  2. Apply the method to the Abilene-inspired network
     - Abilene-inspired network: based on the actual router-level topology
  3. Dispute the influence of the small buffer size on the whole network and on TCP connections

Network Model



: router (Drop Tail discipline) : endhost : TCP connection (TCP Reno)

7

Traffic Model

- Assume network traffic is generated from the edge router
  - Edge router: endhosts are connected directly
- Amount of network traffic is determined using gravity model
  - Amount of traffic injected into/leaving from the edge router is proportional to # of endhosts connected to the edge router
- # of TCP connections between edge routers is determined to be proportional to the network traffic

8

TCP Behavior Model

- Focus on the average behavior of TCP
- Discrete time system with time slot Δ
- Inputs: packet loss ratio $d(k)$, number of packets in the buffer $q(k)$
- Output: Congestion window size of TCP $w(k)$

$$w(k+1)=w(k)+\frac{w(k)}{r(k)}\Delta\left\{(1-d(k)\frac{1}{w(k)}-d(k)(1-d_{TO})\frac{2w(k)}{3}-d(k)d_{TO}\left(\frac{4w(k)}{3}\right)\right\}$$

Probability TCP detects packet loss by timeout mechanism

Round-trip time

9

$\lambda(k)$: TCP throughput

b: Buffer size

μ: Link bandwidth

$Erf()$: Error function

Network Link Model

- Assume each router has separate output buffer for each outgoing link
- Model the buffer as a FIFO queue
- Discrete time system with time slot Δ
- Input: window size of TCP $w(k)$
- Outputs: # of packets in the buffer $q(k)$, packet loss ratio $d(k)$

$$q(k+1) = q(k) + \Delta\left(\sum \lambda(k) - \mu\right)$$

$$d(k) = 1 - \frac{1}{2} Erf\left(\frac{b - q(k)}{\sigma(q(k))}\right)$$

10

Connecting Systems & Analysis

- $w_\chi(k)$: window size of TCP connection χ in time slot k
- w_χ^* : window size of TCP connection χ in steady state
- $q_{(v,w)}(k)$: # of packets in buffer (v,w) in time slot k
- $q_{(v,w)}^*$: # of packets in buffer (v,w) in steady state
- Obtain w_χ^* and $q_{(v,w)}^*$ by solving equations:

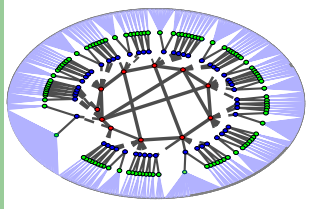
$$w_\chi(k+1) \equiv w_\chi(k) \equiv w_\chi^*$$

$$q_{(v,w)}(k+1) \equiv q_{(v,w)}(k) \equiv q_{(v,w)}^*$$

11

[6] D. Alderson, L. Li, W. Willinger, and J. C. Doyle, "Understanding Internet topology: principles, models, and validation," *IEEE/ACM Transactions on Networking*, vol. 13, no. 6, pp. 1205–1218, Dec. 2005.

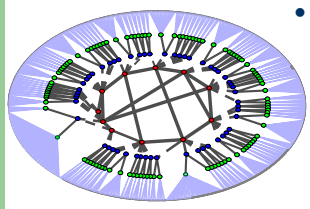
Abilene-inspired Network [1] (1/2)



- Designed based on the characteristics of the actual router topology
  - Core routers have small # of link with higher bandwidth
  - Edge routers have large # of link with slower bandwidth

12

Abilene-inspired Network (2/2)



- The topology consists:
  - 11 core routers
  - 54 middle routers
  - 106 edge routers
  - 812 endhosts
  - 901 bidirectional link

●:core router ●:edge router ●:middle router ●:endhost —: l_{cc} ---: l_{cm} —: l_{me} —: l_{ee}

Slow Edge Network: Parameters

- Link Property:

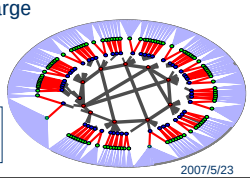
Link	Bandwidth	Prop. Delay
l_{cc}	10 [Gbit/s] (OC192)	0.01 [ms]
l_{cm}	2.5 [Gbit/s] (OC48)	0.1 [ms]
l_{me}	1 [Gbit/s] (GE)	1 [ms]
l_{ee}	5, 30, 100, 1000 [Mbit/s]	1 [ms]

- # of TCP connections in the network: 13,974
  - # of TCP connection is determined by the gravity model
- Average two-way propagation delay of TCP: 4.828 [ms]

Slow Edge Network: Results

- Can not observe significant differences in between traditional and small buffers
- Reason: Bottleneck of Network is l_{me}
  - Traffic injected into l_{cc} is limited
  - Utilization of l_{cc} is not so large to make differences
    - Utilization of l_{cc} is smaller than 0.3

—: l_{me}



No advantage in small buffer size in slow edge network

Fast Edge Network: Parameters

- Link Property:

Link	Bandwidth	Prop. Delay
l_{cc}	10 [Gbit/s] (OC192)	0.01 [ms]
l_{cm}	10 [Gbit/s] (OC192)	0.1 [ms]
l_{me}	10 [Gbit/s] (OC192)	0.1 [ms]
l_{ee}	1 [Gbit/s] (GE)	1 [ms]

- # of TCP connections in the network: 52,394
- Average two-way propagation delay of TCP: 4.829 [ms]

Fast Edge Network: Link Utilization & Packet Loss Ratio

Bufer	Link	Link Utilization	Packet Loss Ratio
Small	l_{cc}	0.638	0.001973
Traditional	l_{cc}	0.968	0.004195

- Link utilization and packet loss ratio of l_{cc} with small buffer are smaller than those with traditional buffer

Reason:

- Reducing the buffer size of l_{cc} causes high packet loss ratio of l_{cc}
- TCP (that traverses l_{cc}) throughput ↘
- TCP (that dose not traverse l_{cc}) throughput ↗

Averages of all TCP in the network

Fast Edge Network: Throughput & Round-trip time

Bufer	TCP throughput [Mbit/s]	Round-trip time [ms]
Small	9.29	16.72
Traditional	9.23	23.93

- No difference in TCP throughput
- Traditional buffer's RTT is lager than small buffer's one
  - Reason1: traditional buffer size is much larger than small buffer size
  - Reason2: link utilization of l_{cc} is high

Small buffer's throughput is larger than traditional buffer's one

Small buffer's throughput is less than traditional buffer's one

Fast Edge Network: TCP throughput detail

- TCP throughput for different # of link hops

Don't traverse the core network | Traverse the core network

buffer	4 hops	6 hops	7 hops	8 hops	9 hops
small	160.32	67.33	3.29	2.41	2.15
traditional	141.94	47.04	7.10	3.33	2.32

- TCP that don't traverse the core network dispossess of throughput of TCP that traverse the core network in the edge network

Buffer size of the core router should not be decreased

Conclusion & Future Work

- Conclusion
  - Proposed an analysis method for a large-scale network
    - Over 100/1,000/10,000 routers/endhosts/links
  - Investigated the effectiveness of small buffer size of the core router
    - Small buffer size of the core router has almost no merit
- Future work
  - Further investigate small buffer size in the network where streaming services exist