Analysis of Backward Congestion Notification with Delay for Enhanced Ethernet Networks

Wanchun Jiang, Fengyuan Ren, Chuang Lin
Department of Computer Science and Technology
Tsinghua University, Beijing, China 100084
Email: {jiangwc, renfy, chlin}@csnet1.cs.tsinghua.edu.cn

Ivan Stojmenovic
SITE, University of Ottawa
800 King Edward, Ottawa,Ontario K1N 6N5, Canada
Email: ivan@site.uottawa.ca

***Abstract*—Recently, companies and standards organizations are enhancing Ethernet as the unified switch fabric for all of the TCP/IP traffic, the storage traffic and the interprocess communication(IPC) traffic in Data Center Networks(DCNs). Backward Congestion Notification(BCN) is the basic mechanism for the end-to-end congestion management enhancement. To fulfill the special requirements of the unified switch fabric that being lossless and of extremely low latency, BCN should hold the queue length around a target point tightly. Thus, the stability of the control loop and the buffer size are critical to BCN. Currently, the impacts of delay on the performance of BCN are unidentified. When the link capacity increases to $40Gbps$ or $100Gbps$ in the near future, the number of on-the-fly packets becomes the same order with the shallow buffer size of switches. Thus, the impacts of delay on the performance of BCN will become significant. In this paper, we analyze BCN, paying special attention on the delay. Firstly, we model the BCN system with a set of segmented delayed differential equations. Then, the sufficient condition for the uniformly asymptotic stability of the BCN system is deduced. Subsequently, the bound of buffer occupancy under this sufficient condition are estimated, which provides guidelines on setting buffer size. Finally, the numerical analysis and the experiments on the NetFPGA platform verify the theoretical analysis.**

***Index Terms*—Backward Congestion Notification, Data Center Ethernet, Stability and Delay**

I. Introduction

Currently, DCNs tend to deploy a unified switch fabric for all of the TCP/IP traffic, the storage traffic and the IPC traffic. Ethernet is being enhanced by IEEE 802.1 Data Center Bridging (DCB) work group [1] to satisfy the special requirements of the unified switch fabric, such as lossless and controllable low delay. This enhanced Ethernet is called Data Center Ethernet (DCE). In the view of IEEE 802.1 DCB work group, the transient congestion is supposed to be eliminated by the priority-based Pause mechanism developed by IEEE 802.1Qbb work group [2]. Although this Pause mechanism can guarantee DCE is lossless, it results in the saturation tree problem [3], which will severely degrade the performance of Ethernet under the long-lived congestion. In order to eliminate the long-lived congestion, the end-to-end congestion management scheme is developed by IEEE 802.1Qau work group [4]. Aiming to hold the queue length at the target point, the end-to-end congestion management scheme also contributes to the lossless and controllable low delay of DCE.

Nowadays, $10Gbps$ Ethernet switch is appearing in commercial applications. The $40Gbps$ and $100Gbps$ Ethernet standards have been ratified in Jun. 2010 [5]. To adapt to the high speed, the congestion management scheme for DCE should be simple enough for the hardware implementation. More importantly, the delay bandwidth product, which is a key element in the design of the congestion management scheme, changes drastically with the Ethernet speed. Since the propagation delay is in the order of microseconds, the delay bandwidth product is small when the link capacity is $1Gbps$ or $10Gbps$. For example, when the link capacity is $10Gbps$, the number of on-the-fly packets is only about 2 ($\frac{3\times10^{-6}\times10^{10}}{1500\times8} \approx 2.5$) with $3\mu s$ propagation delay (which imply the length of link is about $500m$, when the average packet length is $1500Bytes$). It suggests that the delay can be neglected. But when the Ethernet speed becomes $100Gbps$, the number of on-the-fly packet is about 25, which becomes comparable to the shallow buffer size of switches. With the increase of the Ethernet speed, the delay will become a central element of the congestion management scheme.

Up to now, four proposals for the end-to-end congestion management in DCE have been published and BCN [6] is the basic one, in which the framework of the end-to-end congestion management scheme for DCE is established. Most of the investigations on BCN are based on simulations. Although simulations can provide parameters settings for BCN working on certain environment, these parameters can't adapt to all kinds of environment, especially when the link capacity becomes $40Gbps$ or $100Gbps$. In contrast, theoretical analysis can provide direct choices of suitable parameters, as we have shown the sufficient condition for the stability of BCN in [10]. However, there are only a few theoretical works on BCN, focusing on the $1Gbps$ or $10Gbps$ Ethernet. Specially, the impacts of delay on the performance of BCN are unidentified. In this paper, we firstly build a fluid-flow model for the BCN system, accounting for the delay. Then by analyzing the corresponding segmented delayed differential equations, we conclude that the BCN system is uniformly asymptotically stable when the delay is bounded. Subsequently, bound of buffer occupancy under the sufficient condition are explored, which endow the guidelines towards setting buffer size. Finally, the numerical analysis and the experiments on the NetFPGA [7] platform are conducted to demonstrate the impacts of parameters and verify the theoretical results, respectively.

978-1-4673-0775-8/12/$31.00 ©2012 IEEE

II. Modeling

A. Basic Mechanism of BCN

The core mechanism of BCN is introduced here. More details can be found in [6]. As shown in $Fig.1$, BCN is composed of two parts:

- Congestion Point(CP) refers to the switch, where congestion happens. The task of CP is to detect congestion, generate feedback massages and send them to the reaction point.
- Reaction Point(RP) refers to the rate regulator associated with the source or the edge switch. The goal of RP is to adjust the sending rate according to feedback messages.

At CP, the switch monitors the instantaneous queue length $q(t)$ and "samples" incoming packets with probability p. The congestion is measured by F_b, which consists of the current offset of the queue ($Q_{off} = q(t) - q_0$) and the variance of the queue length in a sampling interval ($\Delta Q = q(t) - q_{old}$), where q_0 is the target queue length and q_{old} is the queue length at the sampling last time. F_b is given by

$$F_b = -(Q_{off} + w * \Delta Q) \tag{1}$$

where w is a weight. The congestion message involving F_b is generated and carried to the source of the sampled packet, i.e., the RP.

At RP, the AIMD-like algorithm is adopted for rate adjustment. The feedback information is included into the AIMD-like algorithm to adjust the degrees of rate increase and rate decrease. The sending rate r is adjusted as follows:

$$r \leftarrow \begin{cases} r(1 + G_d F_b) & \text{if } F_b < 0 \\ r + G_i R_u F_b & \text{if } F_b > 0 \end{cases} \tag{2}$$

where G_d is a constant chosen such that $G_d|F_{bmax}| = \frac{1}{2}$, i.e., the sending rate decreases no more than 50% each time, G_i is the factor of rate increase and R_u is the unit of rate increase.

With the collaboration of CP and RP, BCN aims to adapt the injecting rate to the capacity of the network such that buffer occupancy stays at the target point q_0. The stable queue is also beneficial for achieving lossless and controllable low delay.

B. Fluid-Flow Model of BCN

Considering the queue associated with the bottleneck link, the dynamics of CP can be modeled by

$$\frac{dq(t)}{dt} = N\left[r(t-\tau) - \frac{C}{N}\right] \tag{3}$$

where N is the number of active flows, C denotes the capacity of the bottleneck link and τ is the propagation delay as shown in $Fig.1$. Equation (3) means the differential of the queue length equals to the input rate $Nr(t-\tau)$ minus the output rate C. Moreover, the difference of the queue length ΔQ in a sampling interval is

$$\Delta Q = \Delta t \frac{dq(t)}{dt} = \frac{1}{pC}\frac{dq(t)}{dt} \tag{4}$$

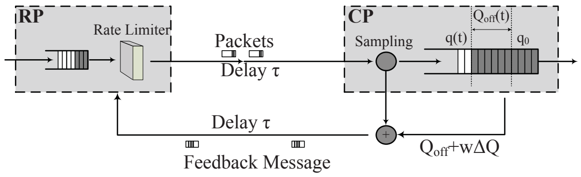


Fig. 1. BCN Framework

where p is the sampling probability. Besides, the differential equation describing the AIMD like algorithm in RP is

$$\frac{dr(t)}{dt} = \begin{cases} G_d F_b(t-\tau) r(t) * pC & \text{if } F_b(t-\tau) < 0 \\ G_i R_u F_b(t-\tau) * pC & \text{if } F_b(t-\tau) > 0 \end{cases} \tag{5}$$

Obviously, $q(t) = q_0$ and $r(t) = \frac{C}{N}$ is a solution of the delayed differential equations, namely $(q_0, \frac{C}{N})$ is the stable point of BCN.

III. Stability Analysis

The stability of above model will be analyzed in this section. For the sake of simplicity, variables $k = \frac{w}{pC}$, $H_i = G_i R_u pC$, $H_d = \frac{G_d pC}{N}$ are defined, and linear substitutions (6) are made.

$$\begin{cases} x(t) = q(t) - q_0 \\ y(t) = N * r(t) - C \end{cases} \tag{6}$$

To facilitate the expression, we define

$$\psi(t) \triangleq \begin{bmatrix} x(t) \\ y(t) \end{bmatrix} \quad and \quad G(t,\psi) \triangleq \frac{d\psi(t)}{dt}.$$

In the rate increase area, i.e., when $F_b(t-\tau) > 0$, the BCN system model can be written into the form of matrix

$$\frac{d\psi(t)}{dt} = A_i * \psi(t-\tau) + B_i * \psi(t-2\tau) \tag{7}$$

where

$$A_i = \begin{bmatrix} 0 & 1 \\ -H_i & 0 \end{bmatrix} \quad and \quad B_i = \begin{bmatrix} 0 & 0 \\ 0 & -H_i k \end{bmatrix}$$

The characteristic equation of equation (7) is

$$\lambda^2 + H_i(k\lambda + 1)e^{-2\lambda\tau} = 0 \tag{8}$$

Theorem 1: If $\tau \leq \min\{\frac{\pi}{8(G_i R_u w + \sqrt{G_i R_u pC})}, \frac{\sqrt{2}w}{4Cp}\}$, i.e., $\tau \leq \min\{\frac{\pi}{8(H_i k + \sqrt{H_i})}, \frac{\sqrt{2}}{4}k\}$, the rate increase subsystem of BCN is uniformly asymptotically stable.

Proof:

When the characteristic equation (8) has real root λ, we can assert that $\lambda < 0$. Or else if $\lambda \geq 0$, obviously there is

$$\lambda^2 + H_i(k\lambda + 1)e^{-2\lambda\tau} > 0$$

This is contrary to equation (8). Thus $\lambda < 0$.

When the characteristic equation (8) has complex root $\lambda = u + iv$, where i is the imaginary unit and $v \neq 0$, we can assert that $u < 0$. Or else, if $u \geq 0$, from equation (8) we have

$$0 = |\lambda^2 + H_i(k\lambda + 1)e^{-2\lambda\tau}| \geq |\lambda|^2 - H_i k|\lambda| - H_i$$

Hence, there is

$$|\lambda| \leq \frac{H_i k \pm \sqrt{(H_i k)^2 + 4H_i}}{2} \leq (H_i k + \sqrt{H_i})$$

According to the condition of this theorem,

$$|\lambda|\tau = \sqrt{u^2 + v^2}\tau \leq (H_i k + \sqrt{H_i})\tau \leq \frac{\pi}{8}$$

Hence, $u\tau < \frac{\pi}{8}$ and $v\tau < \frac{\pi}{8}$. However, when $u \geq 0$ and $v \neq 0$, the imaginary part of equation (8) satisfies

$$\begin{aligned} & Im[\lambda^2 + H_i(k\lambda + 1)e^{-2\lambda\tau}]/v \\ & \quad \geq 2u - 2uH_i k\tau e^{-2u\tau} + (k\cos 2v\tau - 2\tau)H_i e^{-2u\tau} \\ & \quad \geq 2u(1 - H_i k\tau) + (\frac{\sqrt{2}}{2}k - 2\tau)H_i e^{-2u\tau} \\ & \quad > 0 \end{aligned}$$

This equation contradicts to equation (8). Thus $u < 0$.

In sum, all roots of the characteristic equation (8) have negative real part. Hence, the rate increase subsystem of BCN is uniformly asymptotically stable. ■

In the rate decrease area, i.e., when $F_b(t-\tau) < 0$, the BCN system model can be written as

$$\frac{d\psi(t)}{dt} = A_d * \psi(t - \tau) + B_d * \psi(t - 2\tau) + F_d(t, \psi) \quad (9)$$

where $F_d(t, \psi)$ is the nonlinear polynomial part and

$$A_d = \begin{bmatrix} 0 & 1 \\ -H_d C & 0 \end{bmatrix} \quad and \quad B_d = \begin{bmatrix} 0 & 0 \\ 0 & -H_d kC \end{bmatrix}$$

Firstly of all, the linear version of the rate decrease subsystem is considered alone. The characteristic equation of the linear part of (9) is

$$\lambda^2 + H_d C(k\lambda + 1)e^{-2\lambda\tau} = 0 \quad (10)$$

Theorem 2: If $\tau \leq \min\{\frac{\pi N}{8C(G_d + \sqrt{G_d pN})}, \frac{\sqrt{2}w}{4Cp}\}$, namely, $\tau \leq \min\{\frac{\pi}{8(H_d Ck + \sqrt{H_d C})}, \frac{\sqrt{2}}{4}k\}$, the linear version of the rate decrease subsystem of BCN is uniformly asymptotically stable.

The proof of Theorem 2 is the same as that of Theorem 1. Note that the nonlinear part $F_d(t, \psi)$ of the rate decrease subsystem is polynomial of x and y. Considering the whole rate decrease subsystem of BCN, we have

Theorem 3: If $\tau \leq \min\{\frac{\pi N}{8C(G_d + \sqrt{G_d pN})}, \frac{\sqrt{2}w}{4Cp}\}$, the rate decrease subsystem of BCN is uniformly asymptotically stable.

Proof: According to Theorem 2, the linear version of the rate decrease subsystem of BCN is uniformly asymptotically stable when $\tau \leq \min\{\frac{\pi N}{8C(G_d + \sqrt{G_d pN})}, \frac{\sqrt{2}w}{4Cp}\}$. Since both $G(t, \psi)$ and $F_d(t, \psi)$ are polynomials of x and y, the global Lipschitz condition is satisfied, i.e., there exist K and N, for any (t, ψ) and $(t, \tilde{\psi})$, there are $||G(t, \psi) - G(t, \tilde{\psi})|| \leq K||\psi - \tilde{\psi}||$ and $F_d(t, \psi) \leq N \max\{||\psi||, ||\frac{d\psi}{dt}||/K\}$. According to the theorem in [12], the rate decrease subsystem of BCN is also uniformly asymptotically stable. ■

Now we have proven that both the rate increase subsystem and the rate decrease subsystem are uniformly asymptotically stable. It means that, with the increase of time t, $(x(t), y(t))$ will approach to the origin, no matter it is in the rate increase area or the rate decrease area. That is, each time the phase trajectory of the BCN system reaches the switching line $F_b = 0$ at point d, d becomes closer to the origin, no matter from the rate increase area or the rate decrease area. According to the contraction mapping principle[11], the whole BCN system is uniformly asymptotically stable.

Theorem 4:

If $\tau \leq \min\{\frac{\pi N}{8C(G_d + \sqrt{G_d pN})}, \frac{\pi}{8(G_i R_u w + \sqrt{G_i R_u pC})}, \frac{\sqrt{2}w}{4Cp}\}$, the core mechanism of BCN is uniformly asymptotically stable.

Theorem 4 shows that the stability of the core mechanism of BCN depends on the delay τ directly. When the link capacity is $1Gbps$ or $10Gbps$, and the parameters set to be the recommended values, the upper bound of τ is large enough. The core mechanism of the BCN system is stable, just like the analysis in [8] and [10]. With the increase of the link capacity, all the bounds of τ decrease. The delay τ can't be neglected as in [8] and [10]. Since the delay is hard to be reduced, the parameters may need to be changed to enlarge the bounds of τ for the stability of the core mechanism of BCN. Besides, the stability of the core mechanism of BCN has nothing to do with the target point q_0. But, q_0 is directly associated to the bound of buffer occupancy, as we will show in the next section. In sum, Theorem 4 provides guidelines for BCN working on the $100Gbps$ Ethernet directly.

IV. Bound of Buffer Occupancy

In this section, the bounds of buffer occupancy will be estimated to provide guidelines on sizing buffer size. The initial state of BCN is $q(0) = 0$ and $r(0) = \nu$, where ν is the initial sending rate of sources. Assume that $N\nu < C$, i.e., the initial state $(x(t), y(t)) = (-q_0, N\nu - C)$ is in the rate increase area. At the beginning of the accumulation of packets in the buffer, there must exist a time point such that $r(t) = C$, that is the solution curve of BCN will pass through $([-q_0, 0], 0)$. Let t_1 denote the first time the solution curve of BCN crosses the interval $([-q_0, 0], 0)$ after the initial state. Let t_2 denote the first time, subsequent to t_1, $F_b(t_2) = 0$. And let t_3 denote the first time, subsequent to t_1, $y(t_3 + \tau) = 0$. Thus, $\frac{dq(t)}{dt} = 0$ at time t_3 referring to equation (3) and (6), namely $x(t_3)$ represents the maximum of the buffer occupancy.

At first, we estimate the upper bound of $y(t_2)$. Note that $y(t_1) = 0$, we have

$$\int_{t_1}^{t_2} y(t)\dot{y}(t)dt = \frac{y^2(t_2)}{2} \quad (11)$$

Starting at time t_1, the solution curve is below the switching line Γ $(F_b(t) = 0)$. So before the solution curve reaches Γ at time t_2, $x(t) + ky(t - \tau) \leq 0$ holds. Thus, when $t \in [t_1, t_2]$,

$$\dot{y}(t) = -H_i[x(t - \tau) + ky(t - 2\tau)] \geq 0 \quad (12)$$

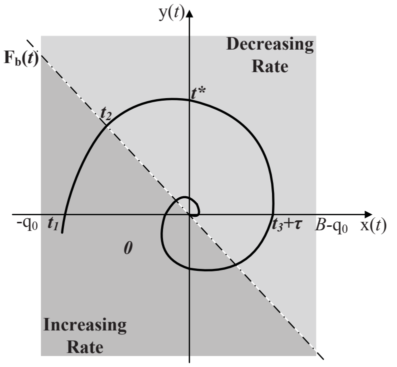


Fig. 2. Solution Curve of BCN Starting from Initial State

Since $y(t_1) = 0$, $y(t_2) \geq y(t) \geq y(t_1) = 0$. A further result is that, when $t \in [t_1 + \tau, t_2]$

$$\dot{x}(t) = y(t - \tau) \geq 0 \tag{13}$$

Since $x(t_1) \geq -q_0$ and $x(t_2) = -ky(t_2 - \tau) \leq -ky(t_1) = 0$, $-q_0 \leq x(t) \leq 0$ when $t \in [t_1 + \tau, t_2]$. Obviously, $-q_0 \leq x(t) \leq 0$ when $t \in [t_1, t_1 + \tau]$. Hence, when $t \in [t_1, t_2]$, the solution curve of BCN is in the second quadrant as shown in $Fig.2$. Known the evolution trend of $x(t)$ and $y(t)$ in time interval $[t_1, t_2]$, we have

$$\begin{aligned}\int\limits_{t_1}^{t_2} y(t)\dot{y}(t)dt &= \int\limits_{t_1}^{t_2} \dot{x}(t+\tau)\dot{y}(t)dt \\ &\leq H_i \int\limits_{t_1}^{t_2} [q_0 - k * y(t - 2\tau)]\dot{x}(t+\tau)dt \\ &\leq H_i q_0 [x(t_2 + \tau) - x(t_1 + \tau)] \\ &\leq H_i q_0^2\end{aligned} \tag{14}$$

In the above expression, the first bound and last bound follow from $0 \geq x(t) \geq -q_0$. The second follows from $y(t) \geq 0$. From (11) and (14), we can obtain that

$$y(t_2) \leq \sqrt{2H_i} q_0 \tag{15}$$

With the same method, we have

$$\frac{y^2(t_2)}{2} \geq \frac{1}{2} H_d C x^2(t_3) \tag{16}$$

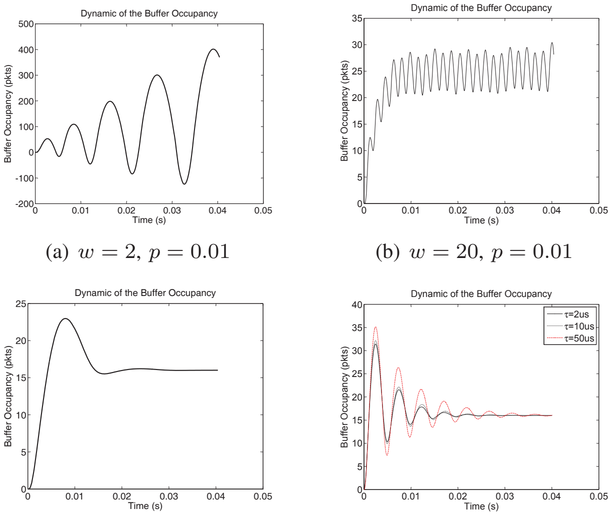
Therefore,

$$x(t_3) \leq \frac{y(t_2)}{\sqrt{H_d C}} \leq \sqrt{\frac{2H_i}{H_d C}} q_0 = q_0\{1 + \sqrt{\frac{2G_i R_u N}{G_d C}}\} \tag{17}$$

When the core mechanism of BCN is uniformly asymptotically stable, $x(t_3)$ represents the max buffer occupancy. Hence,

Theorem 5: If $\tau \leq \min\{\frac{\pi N}{8C(G_d + \sqrt{G_d p N})}, \frac{\pi}{8(G_i R_u w + \sqrt{G_i R_u p C})}, \frac{\sqrt{2}w}{4Cp}\}$, and BCN starts from the initial state, the queue length satisfies $q(t) \leq q_0\{1 + \sqrt{\frac{2G_i R_u N}{G_d C}}\}$

The delay τ doesn't occurs in Theorem 5, namely the delay τ has little influence on the maximum of required buffer occupancy. The estimated bound in Theorem 5 is almost the same as the result in [10], excepting for adding a constant 2 into the radical sign.



(a) $w = 2$, $p = 0.01$

(b) $w = 20$, $p = 0.01$

(c) $w = 2$, $p = 0.001$

(d) τ changes

Fig. 3. Solution curves of the delayed differential equations describing the core mechanism of BCN

V. Numerical Analysis and Experiments

In numerical analysis and experiments, the default configuration of BCN is $G_i = 4$, $R_u = 1Mbps$, $w = 2$, $G_d = \frac{1}{128}$ and $p = 0.01$. Subsequently, without declared explicitly, these parameters will stay unchanged.

A. Numerical Analysis

The same as in [9], we set $C = 10Gbps$, $q_0 = 16pkts$, $N = 50$, $\tau = 200\mu s$,and use Matlab to compute the numerical solution of the delayed differential equations describing the core mechanism of BCN. All packets are of length $1.5KB$.

1) Impacts of Parameters on System Stability: Under the default parameters setting, there is

$$\begin{cases} \frac{\pi N}{8C(G_d + \sqrt{G_d p N})} & = 3.35 * 10^{-4} \\ \frac{\pi}{8(G_i R_u w + \sqrt{G_i R_u p C})} & = 1.69 * 10^{-4} \\ \frac{\sqrt{2}w}{4Cp} & = 8.48 * 10^{-5} \end{cases} \tag{18}$$

Since $\tau = 2 * 10^{-4} > 8.48 * 10^{-5}$, Theorem 4 is not satisfied. The BCN system is unstable as the solution curve shown in $Fig.3$(a). The solution curve differs from the simulation result of [9], because the physical constraints of buffer are not included in the delayed differential equations. In reality, the buffer may be emptied or overflowed temporarily as shown in the simulation result of [9]. The BCN system can be stabilized by either changing parameters or reducing the delay, as shown in $Fig.3$(b), (c) and (d). However, the delay is hard to be reduced. When the link capacity C becomes $100Gbps$, all the bounds of the delay decrease, the core mechanism of BCN will be unstable.

2) Impacts of Parameters on Buffer Occupancy: We consider the impacts of parameters on the buffer occupancy in the condition that BCN starts from the initial state and satisfies Theorem 4. Firstly, τ is changed to be $20\mu s$ to move the core

TABLE I
MAX BUFFER OCCUPANCY AND THE CORRESPONDING ESTIMATION

Parameters	Max Queue Length	Estimation of Th.5
$N = 10, 25, 50$	21.59, 26.74, 32.67	32.16, 41.55, 52.13
$G_i = 1, 2, 4$	25.40, 28.55, 32.67	34.07, 41.55, 52.13
$R_u = 1, 4, 8(Mbps)$	32.67, 44.15, 52.64	52.13, 88.26, 118.20
$G_d = \frac{1}{128}, \frac{1}{64}, \frac{1}{32}$	32.67 , 26.74, 22.65	52.13, 41.55, 34.07
$C = 1, 10, 40(Gbps)$	48.11, 32.67, 27.26	130.26, 52.13, 34.07
$q_0 = 15, 20, 25(pkts)$	30.62, 40.84, 51.04	48.87, 65.16, 81.46
$p = 0.001, 0.002, 0.003$	21.36, 24.84, 28.29	52.13, 52.13, 52.13
$\tau = 2, 10, 50(\mu s)$	31.42, 32.18, 35.14	52.13, 52.13, 52.13
$w = 5, 10, 20$	23.90, 17.53, 16.00	52.13, 52.13, 52.13

mechanism of BCN into the stable state. Then, parameters are varied respectively such that their impacts on buffer occupancy are exhibited. The maximum of buffer occupancy are obtained through numerical analysis and then listed in TABLE I with the bounds estimated by Theorem 5. In TABLE I, all the maximum of buffer occupancy are smaller than the estimated bounds, and all the estimated bounds are less than three times of the corresponding maximum of buffer occupancy. Therefore, the bound estimated by Theorem 5 is reasonable.

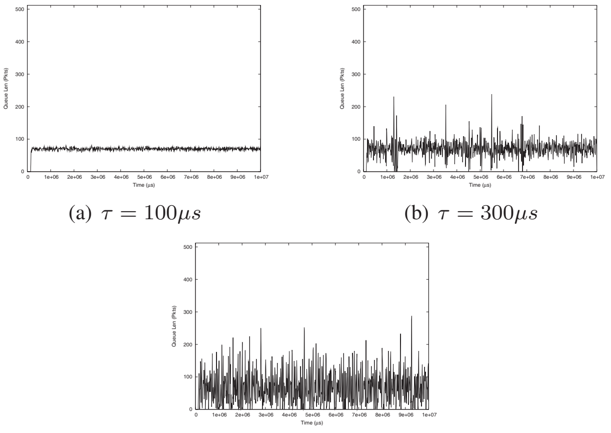
The buffer size suggested by Theorem 5 is larger than the delay bandwidth product, which is the rule-of-thumb for sizing the buffer. For example, under the default parameters setting, where τ is changed to be $20\mu s$, the buffer size is suggested to be $17pkts$, the same as the delay bandwidth product according to the rule-of-thumb, and $53pkts$ according to Theorem 5. Sizing buffer according to Theorem 5, the buffer will be neither full nor empty.

B. Experiments

We also implement the core mechanism of BCN and a delay module on the NetFPGA platform. Experiments use the 2-sources dumbbell topology and the dynamics of the queue length at the bottleneck link are shown. In experiments, $C = 1Gbps$, $N = 2$, $q_0 = 64pkts$, $B = 512pkts$ and all packets are of length $1KB$. According to Theorem 4, the delay τ should be smaller than $330\mu s$ for the stability of the core mechanism of BCN. We changes the delay τ by reconfiguring the delay module in experiments,. The results are shown in $Fig.4$. When $\tau = 500\mu s > 300\mu s$, the core mechanism of BCN becomes unstable and the link utilization degrades. On the contrary, when Theorem 4 holds, the core mechanism of BCN is stable. Therefore, this experiment results is consistent with Theorem 4. Furthermore, the buffer occupancy shown in $Fig.4$ is consistent with Theorem 5.

VI. CONCLUSION

BCN is the basic mechanism, which radicates the framework of the end-to-end congestion management scheme in DCE. This paper analyzes the BCN system theoretically, paying special attention on the impacts of the delay. It also reveals that the core mechanism of BCN is stable when the delay is bounded. When the Ethernet speed increases to $40Gbps$ or $100Gbps$ in the near future, either the delay should be decreased or BCN needs to be reconfigured. We estimate the maximum of buffer occupancy, providing guidelines towards setting buffer size. The numerical analysis and the experiments on the NetFPGA platform verified the theoretical analysis.



(a) $\tau = 100\mu s$

(b) $\tau = 300\mu s$

(c) $\tau = 500\mu s$

Fig. 4. Dynamics of the queue length at the bottleneck link in the NetFPGA based experiments

VII. ACKNOWLEDGEMENT

The authors gratefully acknowledge the anonymous reviewers for their constructive comments. This work is supported in part by the National Natural Science Foundation of China (NSFC) under Grant No. 60971102 and 60932003, National Basic Research Program of China (973 Program) under Grant No.2009CB320504,2010CB328105 and 2012CB315803, and National Science and Technology Major Project of China (NSTMP) under Grant No.2011ZX03002-002-02.

REFERENCES

[1] *IEEE 802.1: Data Center Bridging Task Group* http://www.ieee802.org/1/pages/dcbridges.html

[2] *IEEE 802.1Qbb: Priority-based Flow Control, Working Draft*, http://www.ieee802.org/1/pages/802.1bb.html

[3] G. F. Pfister and V. A. Norton, *Hotspot contention and combining in multistage interconnection networks*, IEEE Trans. on Computers, Vol. 34, No. 10, pages 933-938, Oct. 1985.

[4] *IEEE 802.1Qau: End-to-end congestion management, Working Draft*,http://www.ieee802.org/1/pages/802.1au.html

[5] *IEEE P802.3ba: 40Gb/s and 100Gb/s Ethernet Task Force*, http://www.ieee802.org/3/ba/index.html

[6] D. Bergamasco and R. Pan, *Backward Congestion Notification Version 2.0*, Sep. 2005, http://www.ieee802.org/1/files/public/docs2005/new-bergamasco-bcn-september-interim-rev-final-0905.ppt

[7] NetFPGA Project, http://netfpga.org/

[8] J. Jiang and R. Jain, *Analysis of Backward Congestion Notification (BCN) for Ethernet In Datacenter Applications*, IEEE INFOCOM Minisymposium, 2007

[9] Y. Lu, R. Pan, B. Prabhakar, D. Bergamasco, V. Alaria and A. Baldini, *Congestion control in networks with no congestion drops*, 44nd Allerton Annual Conference on Communication, Control and Computing, Sep. 2006.

[10] F. Ren and W. Jiang, *Phase Plane Analysis of Congestion Control in Data Center Ethernet Networks*, ICDCS, Jun. 2010.

[11] R. M. Brooks and K Schmit, *The Contraction Mapping Principle and Some Applications*, Electronic Journal of Differential Equations, Monograph, 2009

[12] R. D. Driver, *Ordinary and Delay Differential Equations*, Springer-Verlag, New York, Heigelberg and Berlin, pages 384-398, 1977.