Analysis of Backward Congestion Notification with Delay for Enhanced Ethernet Networks

Wanchun Jiang, Fengyuan Ren, *Member*, *IEEE*, Yongwei Wu, *Member*, *IEEE*, Chuang Lin, *Senior Member*, *IEEE*, and Ivan Stojmenovic, *Fellow*, *IEEE*

Abstract—At present, companies and standards organizations are enhancing Ethernet as the unified switch fabric for all of the TCP/IP traffic, the storage traffic and the high performance computing traffic in data centers. Backward congestion notification (BCN) is the basic mechanism for the end-to-end congestion management enhancement of Ethernet. To fulfill the special requirements of the unified switch fabric, i.e., losslessness and low transmission delay, BCN should hold the buffer occupancy 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 speed of Ethernet increases to 40 Gbps or 100 Gbps in the near future, the number of on-the-fly packets becomes the same order with the buffer size of switch. Accordingly, the impacts of delay will become significant. In this paper, we analyze BCN, paying special attention on the delay. We model the BCN system with a set of segmented delayed differential equations, and then deduce sufficient condition for the uniformly asymptotic stability of BCN. Subsequently, the bounds of buffer occupancy are estimated, which provides direct guidelines on setting buffer size. Finally, numerical analysis and experiments on the NetFPGA platform verify our theoretical analysis.

Index Terms—Backward congestion notification, data center ethernet, stability and delay

1 Introduction

Networking thousands of commercial computers, data centers provide elastic services to a large number of users. The efficiency of network is crucial to the performance of a data center. However, the condition that multiple application-specific networks run over different link layer technologies in data center, such as TCP/IP over Ethernet, Storage Area Networks (SANs) over Fibre Channel and High Performance Computing (HPC) networks over InfiniBand, not only increases the cost of redundant devices, but also complicates the design and management of data center networks (DCNs). To solve this problem, companies and standards organizations are developing a unified switch fabric for DCNs. Currently, Ethernet is enhanced by IEEE 802.1 Data Center Bridging (DCB) work group [1] to satisfy the special requirements of the unified switch fabric, such as losslessness and low transmission delay. Moreover, techniques such as Fiber Channel over Ethernet [2] and iSCSI [3] make it possible to accommodate storage traffic on the lossless Ethernet, while RoCEE [4], MXoE [5] and iWARP try to carry HPC traffic over Ethernet with low transmission delay. The enhanced Ethernet, named Data Center Ethernet (DCE), will become the unified switch fabric of DCNs in the near future.

Congestion management is one of the indispensable enhancements for Ethernet to be the unified switch fabric. To eliminate the transient congestion, the IEEE 802.1Qbb work group introduces the priority-based Pause mechanism into Ethernet [6]. Although the priority-based Pause mechanism can guarantee DCE is lossless, it produces the saturation tree problem [7], which will severely degrade the performance of Ethernet under the long-lived congestion. To eliminate the long-lived congestion, the end-to-end congestion management enhancement of Ethernet is developed by the IEEE 802.1Qau work group [8]. Comparing with the traditional congestion management scheme, whose goal is high link utilization and high throughput, the end-to-end congestion management enhancement of Ethernet set its goal to hold the buffer occupancy at the target point. In this way, the priority-based Pause mechanism would hardly be triggered since buffer is reserved for the burst traffic. Furthermore, the queuing delay also becomes controllable low. Another reason for deploying a uniform congestion management scheme in DCE is that, it's more economic comparing with deploying one for each type of traffic in the transport layer.

Nowadays, the speed of Ethernet is increasing rapidly, and this fact has great influence on the congestion management enhancement in DCE. Currently, the 10*Gbps* Ethernet switch is appearing in commercial applications. The standards of 40*Gbps* Ethernet and 100*Gbps* Ethernet have been ratified in Jun. 2010 [9]. Moreover, devices such as Juniper T1600 [10] and Brocade MLXE32 [11] begin to support ports of 100*Gbps*. In fact, the rapid increasing speed is one of the important reasons for Ethernet to be enhanced as the unified switch fabric. To adapt to the high speed, the congestion management scheme for DCE should be simple enough for the hardware implementation.

- *W. Jiang, F. Ren, Y.Wu, and C. Lin are with the Department of Computer Science and Technology,Tsinghua National Laboratory for Information Science and Technology, Tsinghua University, Beijing 100084, China. E-mail: {jiangwc, renfy, chlin}@csnet1.cs.tsinghua.edu.cn, wuyw@tsinghua.edu.cn.*
- *I. Stojmenovic is with Tsinghua National Laboratory for Information Science and Technology, Tsinghua University, Beijing 100084, China; and also with School of Information Technology and Engineering, University of Ottawa, Ottawa, Ontario K1N 6N5, Canada.*

Manuscript received 11 Sept. 2012; revised 27 June 2013; accepted 04 July 2013. Date of publication 04 Aug. 2013; date of current version 14 Oct. 2014. Recommended for acceptance by M. Guo.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TC.2013.157

More importantly, the bandwidth delay product, which is a key element in the design of the congestion management schemes, drastically changes with the speed of Ethernet. The bandwidth delay product is small in current data centers, since the link capacity is mainly $1Gbps$ or $10Gbps$ and the propagation delay is in the order of microseconds. For example, when the speed of Ethernet 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 speed of Ethernet becomes $100Gbps$, the number of on-the-fly packet is about 25, which is comparable to the buffer size of switch. With the increase of the speed of Ethernet, the delay will take a central role in the performance of the congestion management enhancement in DCE.

Up to now, four proposals for the end-to-end congestion management enhancement in DCE have been published and BCN [12] is the basic one, in which the framework is established. Currently, most of the investigations on BCN are based on simulations. Although simulations can provide proper parameters settings for BCN working on certain environment, these parameters settings can't adapt to the change of environment, especially when the link capacity increases to $40Gbps$ or $100Gbps$ in the near future. In contrast, theoretical analysis can provide direct choices of suitable parameters, as we have shown the sufficient condition for the stability of BCN in [16]. However, there are only a few theoretical works on BCN, acquiescing in the $1Gbps$ Ethernet or the $10Gbps$ Ethernet. Specifically, the impacts of delay on the performance of BCN are unidentified. In this paper, we firstly built a fluid-flow model for the BCN system, accounting for the delay. Then by analyzing the segmented delayed differential equations describing the BCN system, we draw the conclusion that, the BCN system is uniformly asymptotic stable when the delay is bounded. Subsequently, bounds of buffer occupancy are estimated to endow the guidelines towards setting buffer size, as well as show BCN's ability in disturbance attenuation. Finally, the numerical analysis and the experiments on the NetFPGA [13] platform are conducted to verify these theoretical results and demonstrate the impacts of parameters.

The remainder of this paper is organized as follows. The background and related work are introduced in Section 2. Next, the basic mechanism of BCN is described and then the fluid-flow model for BCN is built. The following section is the stability analysis on BCN. In Section 5, the bounds of buffer occupancy of BCN are explored. Subsequently, the numerical analysis and experiments are conducted. Finally, the conclusion is drawn in Section 7.

2 Related Works

Generally, the congestion management system can be described by delayed differential equations. However, the solution of the delayed differential equations is usually hard to be found. Because the transmission delay is small in DCNs, the delay is always approximately neglected to translate the delayed differential equations into the tractable differential equations.

For example, J. Jiang et al. [14] modeled BCN with difference equations, and then approximated them with ordinary differential equations, following the method of stochastic approximation. Ignoring the time lag, they concluded that BCN is stable, and its convergence speed generally decreases with the change of certain system parameters. When the speed of Ethernet is of $1Gbps$ or $10Gbps$, the delay can be ignored, since it is in the order of microseconds and there are only several on-the-fly packets. But when the speed of Ethernet increases to $40Gbps$ or $100Gbps$, the number of on-the-fly packets becomes the same order with the buffer size of switch. Hence, the delay will take an important role in BCN. Simulations also show that, with the increase of the delay, BCN's ability to control the buffer occupancy degrade [17].

On the other hand, Y. Lu et al. [15] constructed a model for BCN with a set of delayed differential equations and did frequency analysis on its linearized version. Making approximation on delay, they deduced sufficient condition for the stability of BCN. But the sufficient condition is composed of four inequalities, and these four inequalities are associated with each other such that impacts of parameters on the performance of BCN can't be shown explicitly. Moreover, their approximation on delay introduces inexactitude into the final result, as we will show in the numerical analysis. In this work, it is shown that all the characteristic roots the delayed differential equations have negative real part, when the delay is bounded. Namely the BCN system is asymptotically stable when the delay is bounded.

Moreover, it's important to set buffer large enough to reduce the chance of triggering the priority-based Pause mechanism in DCE. The suggestion on setting buffer size is absent in the aforementioned two works. In previous investigation, assuming that the delay is negligible, we have used the phase plane method to show the dynamics of BCN and deduced the sufficient condition, which is directly related with the buffer size, for the stability of BCN [16]. In this work, assuming the delay is smaller than the boundary we provided, we estimate the bounds of buffer occupancy to give suggestion on setting buffer size. Our theoretical results are consistent with that of [14] and [16] when the delay approaches to zero.

3 Modeling

As the basic proposal for the end-to-end congestion management enhancement of DCE, BCN is designed to hold the queue length around the target point at the bottleneck link. In fact, the congestion is measured by the queue length at the bottleneck link in the BCN system. Therefore, the dynamic of the queue length is focused in the process of modeling. Before building fluid flow model for BCN, we would firstly describe the core mechanism of BCN. More details can be found in [12].

3.1 Basic Mechanism of BCN

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 messages 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 core switch monitors the instantaneous buffer occupancy $q(t)$ and "samples" incoming packets with probability p. When packets are sampled, the feedback packet

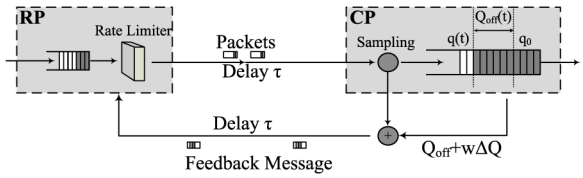

Fig. 1. BCN framework.

involving congestion information is constructed and sent back to the source of the sampled packet, i.e., the corresponding RP. The congestion information is represented by F_b, which consists of the current offset of the buffer occupancy ($Q_{off} = q(t) - q_0$) and the variance of the buffer occupancy in a sampling interval ($\Delta Q = q(t) - q_{old}$), where q_0 is the target buffer occupancy and q_{old} is the buffer occupancy at the sampling last time. F_b is given by

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

where w is a weight. Obviously, the sampling probability p should be large enough to satisfy the Nyquist-Shannon sampling theorem, such that there are enough feedback packets to deliver the congestion information. While the overhead will be increased as the increase of the value of the sampling probability p.

At RP, the AIMD-like algorithm is adopted for rate adjustment. The feedback information is included into the AIMD-like algorithm to adjust the degree 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.

In BCN, feedback messages follow the format of the 802.1Q tag. In fact, there are three types of feedback messages: normal messages involving F_b, PAUSE frames and STOP messages. When the buffer occupancy exceeds an upper bound q_{sc}, the PAUSE frame is sent as the feedback message to ask all its uplink neighbors to stop injecting packets. The STOP messages are generated to ask sources to stop for a random time and then restart, when severe congestion happens. Both the PAUSE message and the STOP message will result in large delay, and thus degrade the link utilization. But if the parameters of the BCN system and the buffer size is set properly, BCN will work on the normal state, namely hold the buffer occupancy around a target point. Thus, we concern on BCN's ability to hold the buffer occupancy at the target point, i.e., the core rate adjustment mechanism of BCN, similar to [14], [15], and [16]. 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 high throughput, low queuing delays, stability and so on.

3.2 Fluid-Flow Model of BCN

To model the BCN system, two assumptions are made. Firstly, since links are of high capacity in DCE networks, the number of bit streams in them is so large that it appears like continuous flow fluid, the fluid-flow approximation extensively used in network modeling work is assumed to be practicable, such as in [15] and [19] et al. This assumption means the buffer occupancy $q(t)$ and the sending rate $r(t)$ are continuous and differentiable. Secondly, considering the regular and symmetrical network topologies in data centers, such as Fat-Tree [20], Bcube [21] and the special traffic patterns driven by the parallel reads/writes in cluster file systems, such as Lustre [22] and Panasas [23], sources are assumed to be homogeneous, when only one bottleneck is concerned. That is, sources have the same characteristics, follow the same routes, and experience the same round-trip propagation delays. With this assumption, the sending rate of sources can be represented by $r(t)$ uniformly.

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 buffer occupancy equals to the input rate $Nr(t-\tau)$ minus the output rate C. Moreover, the difference of the buffer occupancy ΔQ in a sampling interval is

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

as $\frac{1}{pC}$ represents the length of the sampling interval. Combining (1), (3), and (4), the feedback variable F_b can be represented as

$$F_b = -\left\{[q(t) - q_0] + \frac{wN}{pC}\left[r(t-\tau) - \frac{C}{N}\right]\right\}. \tag{5}$$

Besides, the differential equation describing the AIMD like algorithm in RP is

$$\frac{dr(t)}{dt} = \begin{cases} G_d F_b(t-\tau) r(t) * pC, & \text{if } F_b(t-\tau) < 0, \\ G_i R_u F_b(t-\tau) * pC, & \text{if } F_b(t-\tau) > 0. \end{cases} \tag{6}$$

Up to now, the core mechanism of BCN has been described by (3), (5), and (6). Our model is similar to that of [15]. The main difference is that we use $r(t-\tau)$ instead of $r(t)$ when modeling the dynamics of CP. As a result, both τ and 2τ are involved in our model. This is more reasonable since the current injecting rate at the switch is the sending rate at sources before τ. Let $\tau = 0$, our model only adds a constant factor pC into equation (6), comparing to the model in [16]. In addition, all the differential equations are associated with only variable $q(t)$ and $r(t)$, while $r(t)$ can be expressed by $q(t)$ referring to equation (3). Hence, behaviors of the BCN system can be described by the behaviors of the queue. 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 a stable point of BCN.

Remark. Note that we focus on only the bottleneck link, when the fluid flow model of BCN is built. This is because the uncongested hop contributes nothing to our model except the value of the delay. Of cause, the assumption of only one single bottleneck link also poses

some limitations. However, recent measurement work [24] shows that the congestion point is actually the last-hop for both the Incast (or say many-to-one) transfer pattern and the shuffle (or say many-to-many) transfer pattern. In addition, the traffic between servers under the same ToR switch is actually the most significant since locality has been considered in job distribution in data centers. Thus, the last-hop single bottleneck will be dominated in data center. Namely, focusing on a single bottleneck link can provide insights about the stability of BCN.

4 Stability Analysis

Generally, the stability refers to the ability of converging to the stable state as the time goes on. Namely, when the solutions of the differential equations describing the BCN system can converge to the stable state $(q_0, \frac{C}{N})$, we say that BCN is stable. There are three obstacles to knowing the stability of BCN from the above differential. The first obstacle is that the BCN system is inherently divided into the rate increase subsystem and the rate decrease subsystem, and the stability of both subsystems is inadequate to deduce out the stability of BCN. To solve this problem, we focus on the uniformly asymptotic stability. When a congestion management system is uniformly asymptotic stable, it uninterruptedly moves close to the stable state $(q_0, \frac{C}{N})$. Obviously, if both subsystems of BCN are uniformly asymptotic stable, the BCN system would uninterruptedly move closer and closer to the stable state, namely the whole BCN system is uniformly asymptotic stable.

The second obstacle to analyzing the above fluid flow model of BCN is the delay. This is because there is no general method to find out the solutions of delayed differential equations, and accordingly the uniformly asymptotic stability of subsystems is also unknown. Luckily, there is a criterion, which declares that when all the roots of the characteristic equation of the differential equations describing the system have negative real part, the corresponding system is uniformly asymptotic stable. Although the characteristic equation of the differential equations is transcendental and accordingly its roots are hard to be found, we can estimate the boundary of the roots.

The third obstacle is the nonlinearity. The general method of solving this problem is to linearize the differential equations around the stable point approximately. Lyapunov has shown that when the Lipschitz condition is satisfied, this linear approximation is reasonable for the ordinary differential equations [25]. But referring to the delayed differential equations, another theorem is needed to promise the availability of the linear approximation. Fortunately, R. D. Driver has proven that when the global Lipschitz condition is satisfied, the delayed differential equations is uniformly asymptotic stable when its linearized version is uniformly asymptotic stable [26].

Subsequently, we will firstly show that both the rate increase subsystem and the rate decrease subsystem of BCN are uniformly asymptotic stability. 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 are made.

$$\begin{cases} x(t) = q(t) - q_0 \\ y(t) = N * r(t) - C. \end{cases} \tag{7}$$

With this substitution, (5) can be rewritten as

$$F_b(t) = -x(t) - ky(t-\tau). \tag{8}$$

Thus, when $F_b(t-\tau) < 0$, the rate decrease subsystem of BCN can be described by

$$\begin{cases} \frac{dx(t)}{dt} = y(t-\tau) \\ \frac{dy(t)}{dt} = -H_d[x(t-\tau) + ky(t-2\tau)][y(t)+C]. \end{cases} \tag{9}$$

And when $F_b(t-\tau) > 0$, the rate increase subsystem of BCN can be described by

$$\begin{cases} \frac{dx(t)}{dt} = y(t-\tau) \\ \frac{dy(t)}{dt} = -H_i[x(t-\tau) + ky(t-2\tau)]. \end{cases} \tag{10}$$

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

$$\frac{d\psi(t)}{dt} = A_i * \psi(t-\tau) + B_i * \psi(t-2\tau), \tag{11}$$

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 (11) is

$$\lambda^2 + H_i(k\lambda + 1)e^{-2\lambda\tau} = 0. \tag{12}$$

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

Proof. To show the uniformly asymptotic stability of the rate increase subsystem of BCN, we will prove that all the real parts of the roots of characteristic equation (12) are negative.

When the characteristic equation (12) has real root λ, we can assert that $\lambda < 0$. Or else if $\lambda \ge 0$, obviously there is

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

This is contrary to equation (12). Thus $\lambda < 0$.

When the characteristic equation (12) has complex root $\lambda = u + iv$, where i is the imaginary unit and $v \ne 0$, we can assert that $u < 0$. Or else, if $u \ge 0$, from equation (12) we have

$$\begin{aligned} 0 &= |\lambda^2 + H_i(k\lambda+1)e^{-2\lambda\tau}| \\ &\ge |\lambda|^2 - H_i k|\lambda e^{-2\lambda\tau}| - H_i|e^{-2\lambda\tau}| \\ &\ge |\lambda|^2 - H_i k|\lambda| - H_i. \end{aligned}$$

The second bound follows from $|e^{-2\lambda\tau}| < 1$ since $u \geq 0$. 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 (12) satisfies

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

where Im means taking the imaginary part of the complex. The first bound follows from $\frac{\sin 2v\tau}{2v\tau} < 1$. The second bound follows from $e^{-2u\tau} < 1$ and $cos(2v\tau) \geq cos\frac{\pi}{4} = \frac{\sqrt{2}}{2}$. The third bound follows from $H_i k\tau < \frac{\pi}{8} < 1$ and $\tau < \frac{\sqrt{2}}{4}k$. This is contrary to equation (12). Thus $u < 0$.

In sum, all roots of the characteristic equation (12) have negative real part. Hence, the rate increase subsystem of BCN is uniformly asymptotic stable. □

In the rate decrease area, i.e., when $F_b(t-\tau) < 0$, the delayed differential equations is nonlinear. It can be divided into the linear part and the nonlinear part.

$$\frac{d\psi(t)}{dt} = A_d{}^*\psi(t-\tau) + B_d{}^*\psi(t-2\tau) + F_d(t,\psi), \qquad (13)$$

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

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

Firstly of all, the linear version of the rate decrease subsystem is considered alone. The characteristic equation of the linear part of (13) is

$$\lambda^2 + H_d C(k\lambda+1)e^{-2\lambda\tau} = 0. \qquad (14)$$

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 asymptotic stable.*

Proof. Replacing H_i with $H_d C$ in the proof of Theorem 1, all the roots of characteristic equation (14) can be proven to have negative real part. Therefore, the linear version of rate decrease subsystem is uniformly asymptotic stable. □

Now, we consider the whole rate decrease subsystem of BCN. Note that the nonlinear part $F_d(t,\psi)$ of the rate decrease subsystem is polynomial of x and y.

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 asymptotic stable.*

Proof. According to Theorem 2, the linear version of the rate decrease subsystem of BCN is uniformly asymptotic 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, it is easy to prove that 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 [26], the rate decrease subsystem of BCN is also uniformly asymptotic stable. □

Now we have provided sufficient conditions for the asymptotic stability of both the rate increase subsystem and the rate decrease subsystem of BCN. 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 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 [18], the whole BCN system is uniformly asymptotic 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 asymptotic stable.*

Theorem 4 shows that the stability of the core mechanism of BCN depends on the delay τ directly. When the link capacity is 1*Gbps* or 10*Gbps*, the upper bounds of τ is large enough when the parameters setting follows the recommended values. Thus, the core mechanism of the BCN system is stable, just like the analysis in [14] and [16]. With the increase of the link capacity, all the three bounds of delay τ decrease. Delay τ can't be neglected as in [14] and [16]. Since delay is hard to be reduced, the parameters may need to be changed to enlarge the bounds of delay 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 100*Gbps* Ethernet directly.

Theorem 4 does reveal the relationship of τ and other parameters. Similar result is also obtained in [15], i.e., τ is bounded by $\frac{1}{aG_i R_u w}$ and $\frac{N}{aG_d Cw}$, along with other constraints $G_i R_u w^2 \geq \frac{pC}{b\sqrt{b^2+1}}$, $G_d w^2 \geq \frac{pN}{b\sqrt{b^2+1}}$, where a and b satisfy $\frac{b}{a} + \arctan b = \frac{\pi}{2}$. But, in [15], the parameters relationship is implicit. Moreover, the bounds are obtained with approximation on delay such that their results are inexactitude. We will show that the parameters settings used in the simulation of [15] fails to make BCN extremely stable by numerical analysis.

The upper bounds of τ is decided by both the rate increase subsystem and the rate decrease subsystem of BCN. Recalling the proof of theorem 1, we can find that the constant $\frac{\pi}{8}$ can be replaced by any real number $a < \frac{\pi}{4}$, while constant $\frac{\sqrt{2}}{4}$ is replaced by $\frac{cos2a}{2}$. In other words, Theorem 4 can be enhanced as "if $\tau \leq \min\left\{\frac{aN}{C(G_d+\sqrt{G_d pN})}, \frac{a}{(G_i R_u w+\sqrt{G_i R_u pC})}, \frac{w\cos 2a}{2Cp}\right\}$ where $a < \frac{\pi}{4}$, the core mechanism of BCN is uniformly asymptotic stable".

5 Bounds of Buffer Occupancy

Up to now, we have deduced the sufficient condition for uniformly asymptotic stability of the core mechanism of BCN.

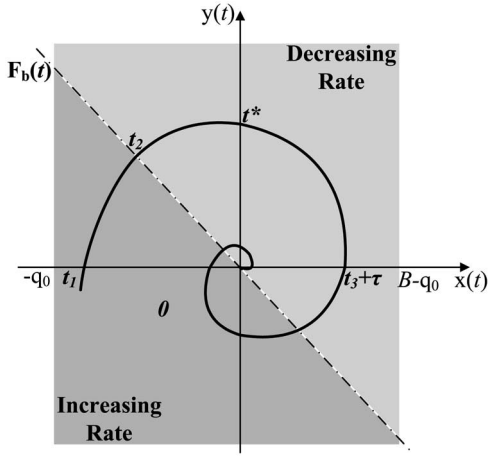

Fig. 2. Solution curve of BCN starting from initial state.

But the dynamics of the BCN system are also physically constrained by the buffer size. In the condition that the buffer is empty or overflow, the behaviors of BCN will deviate from our model. Fortunately, because the core mechanism of BCN is uniformly asymptotic stable, the empty buffer or full buffer must occur at the beginning. If the buffer is set large enough, BCN will never experience the empty buffer again after initializing with empty buffer, and the condition of severe congestion wouldn't occur, i.e., the occurrences of the PAUSE message and the STOP message are prevented. Therefore,it is necessary to estimate the bounds of buffer occupancy to give suggestion on setting buffer size. Two bounds are meaningful: the bound when BCN starts from initial state and the bound for the condition of an additional flow entering into the stable BCN system. The former bound indicates what size of the buffer should be set in BCN, while the latter bound measures BCN's ability of disturbance attenuation.

Firstly, we estimate the bound of buffer occupancy when BCN starts from the initial state. 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 (7), namely $x(t_3)$ represents the maximum of the buffer occupancy.

Next, we will find an upper bound of $y(t_2)$. Note that $y(t_1) = 0$, we have

$$\int_{t_1}^{t_2} y(t)\dot{y}(t)dt = \frac{y^2(t_2)}{2}. \tag{15}$$

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, $x(t) + ky(t - \tau) \le 0$ holds. Thus, when $t \in [t_1, t_2]$,

$$\dot{y}(t) = -H_i[x(t - \tau) + ky(t - 2\tau)] \ge 0. \tag{16}$$

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

$$\dot{x}(t) = y(t - \tau) \ge 0. \tag{17}$$

Since $x(t_1) \ge -q_0$ and $x(t_2) = -ky(t_2 - \tau) \le -ky(t_1) = 0$, $-q_0 \le x(t) \le 0$ when $t \in [t_1 + \tau, t_2]$. In addition, the solution curve of BCN moves from point $(-q_0, N\nu - C)$ to a point in the interval $([-q_0, 0], 0)$ before time t_1, there is $-q_0 \le x(t) \le 0$ 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 can deduce as follows

$$\begin{aligned}
\int_{t_1}^{t_2} y(t)\dot{y}(t)dt &= \int_{t_1}^{t_2} \dot{x}(t + \tau)\dot{y}(t)dt \\
&= -H_i \int_{t_1}^{t_2} [x(t - \tau) + k^* y(t - 2\tau)]\dot{x}(t + \tau)dt \\
&\le H_i \int_{t_1}^{t_2} [q_0 - k^* y(t - 2\tau)]\dot{x}(t + \tau)dt \\
&\le H_i q_0 [x(t_2 + \tau) - x(t_1 + \tau)] \\
&\le H_i q_0^2.
\end{aligned} \tag{18}$$

In the above expression, the first bound and last bound follow from $0 \ge x(t) \ge -q_0$. The second follows from $y(t) \ge 0$. From (15) and (18), we can obtain that

$$y(t_2) \le \sqrt{2H_i q_0}. \tag{19}$$

Subsequently, we find the upper bound of $x(t_3)$. Because t_3 denote the first time, subsequent to t_1, $y(t_3 + \tau) = 0$, there is

$$-\frac{y^2(t_2)}{2} = \int_{t_2}^{t_3+\tau} y(t)\dot{y}(t)dt = \int_{t_2}^{t_3+\tau} \dot{x}(t + \tau)\dot{y}(t)dt. \tag{20}$$

Recalling (17) and (16), $\dot{x}(t_2) = y(t_2 - \tau) > 0$ and $\dot{y}(t_2) = 0$. Go through the switching line Γ after time t_2, there must be $y(t) \ge 0$ before time $t_3 + \tau$. Thus, in time interval $[t_2, t_3 + \tau]$,

$$\dot{x}(t) = y(t - \tau) \ge 0. \tag{21}$$

Once the solution curve reaches Γ at time t', there is $\dot{y}(t') = -H_d[x(t' - \tau) + ky(t' - 2\tau)][y(t') + C] = 0$ and $\dot{x}(t') > 0$. After time t_2, the solution curve will move back into the rate decrease area again. Therefore, the solution curve keeps above the switching line Γ when $t \in [t_2, t_3 + \tau]$. Under this assertion, there is,

$$\dot{y}(t + \tau) = -H_d[x(t) + ky(t - \tau)][y(t) + C] \le 0. \tag{22}$$

Between time interval $[t_2, t_3 + \tau]$, there must be a time t^* such that $x(t^*) = 0$. The evolution trend of $x(t)$ and $y(t)$ in time interval $[t_2, t_3]$ is as shown in Fig. 2. Based on these analysis,

$$\begin{aligned}\frac{y^2(t_2)}{2} &= \int_{t_2}^{t_3+\tau} H_d[x(t-\tau)+ky(t-2\tau)][y(t)+C]\dot{x}(t+\tau)dt \\ &\geq \int_{t^*-\tau}^{t_3-\tau} H_d[x(t-\tau)+ky(t-2\tau)][y(t)+C]\dot{x}(t+\tau)dt \\ &\geq H_dC\int_{t^*-\tau}^{t_3-\tau}[x(t-\tau)+ky(t-2\tau)]\dot{x}(t+\tau)dt \\ &\geq H_dC\int_{t^*-\tau}^{t_3-\tau}[x(t+\tau)-2\tau y(t-2\tau) \\ &\quad + ky(t-2\tau)]\dot{x}(t+\tau)dt \\ &= H_dC[\frac{x^2(t_3)}{2}-\frac{x^2(t^*)}{2}] \\ &\quad + H_dC\int_{t^*-\tau}^{t_3-\tau}(k-2\tau)y(t-2\tau)\dot{x}(t+\tau)dt \\ &\geq \frac{1}{2}H_dCx^2(t_3). \end{aligned} \tag{23}$$

The first and second bounds follow from $y(t) \geq 0$, $\dot{x}(t+\tau) > 0$ and $x(t-\tau)+ky(t-2\tau) \geq 0$. The third bound follows from $\dot{x}(t+\tau) > 0$ and the differential mean value theorem, i.e., there exists $\xi \in [t-\tau, t+\tau]$ such that

$$x(t+\tau)-x(t-\tau) = 2\tau\dot{x}(\xi) = 2\tau y(\xi-\tau) \leq 2\tau y(t-2\tau).$$

The forth bound follows from $\tau < \sqrt{2}k/4 < k/2$, which is part of the sufficient condition for the stability of the core mechanism of BCN. From (20), (23), and (19), we can obtain that

$$x(t_3) \leq \frac{y(t_2)}{\sqrt{H_dC}} \leq \sqrt{\frac{2H_i}{H_dC}}q_0 = q_0\left\{1+\sqrt{\frac{2G_iR_uN}{G_dC}}\right\}. \tag{24}$$

Starting from the initial condition, when the solution curve reaches the positive part of x axis at the first time, $x(t)$ reaches its maximum, namely $x(t) \leq x(t_3)$, because of the uniformly asymptotic stability of the core mechanism of BCN.

Theorem 5. *If* $\tau \leq \min\{\frac{\pi N}{8C(G_d+\sqrt{G_dpN})}, \frac{\pi}{8(G_iR_uw+\sqrt{G_iR_upC})}, \frac{\sqrt{2}w}{4Cp}\}$, *and BCN starts from the initial state, the buffer occupancy satisfies* $q(t) \leq q_0\{1+\sqrt{\frac{2G_iR_uN}{G_dC}}\}$

The estimated bound of buffer occupancy in Theorem 5 has nothing to do with the delay τ, 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 [16], excepting for adding a constant 2 into the radical sign. Contradicting to the rule-of-thumb that buffer is set to be equal to the bandwidth delay product, the estimated bound of buffer occupancy decreases with the increases of the link capacity C in Theorem 5. Therefore, when the link capacity C increases to 10*Gbps* or 100*Gbps*, the buffer size of BCN can keep unchanged with the same parameters setting. However, the problem may occur in the BCN system with small link capacity.

Secondly, we consider the bound of the buffer occupancy for the condition of an additional flow entering into the stable BCN system. Assume that when BCN stays at the stable state, another flow of size C_0 comes into the system, i.e., the BCN system moves from stable state $(0,0)$ to state $(0,C_0)$ suddenly. So the analysis should start from point $(x(t^*),y(t^*)) = (0,C_0)$. Similar to above analysis, the solution curve of BCN will keep above of the switching line and move across the x axis at point $(x(t_3),0)$. Thus, in time interval $[t^*,t_3]$, there is $\dot{x}(t) \geq 0$ and $\dot{y}(t) \leq 0$. With the same method used in (23), there is

$$\begin{aligned}\frac{C_0^2}{2} &= -\frac{y^2(t^*)}{2} = -\int_{t^*}^{t_3} y(t)\dot{y}(t)dt \\ &\geq -\int_{t^*}^{t_3} y(t-\tau)\dot{y}(t)dt \\ &= \int_{t^*}^{t_3} H_d[x(t-\tau)+ky(t-2\tau)][y(t)+C]\dot{x}(t)dt \\ &\geq H_dC\int_{t^*}^{t_3}[x(t-\tau)+ky(t-2\tau)]\dot{x}(t)dt \\ &\geq H_dC\int_{t^*}^{t_3}[x(t)-\tau y(t-2\tau)+ky(t-2\tau)]\dot{x}(t)dt \\ &= H_dC[\frac{x^2(t_3)}{2}-\frac{x^2(t^*)}{2}+\int_{t^*}^{t_3}(k-\tau)y(t-2\tau)\dot{x}(t)dt] \\ &\geq \frac{1}{2}H_dCx^2(t_3). \end{aligned} \tag{25}$$

The first bound follows from $y(t) \leq y(t-\tau)$ since $\dot{y}(t) \leq 0$. The second bound follows from $y(t) \geq 0$, $\dot{x}(t) \geq 0$ and $x(t-\tau)+ky(t-2\tau) \geq 0$ when $t \in [t^*,t_3]$. The third bound follows from $\dot{x}(t) \geq 0$ and the differential mean value theorem, i.e., there exists $\xi \in [t-\tau, t]$ such that

$$x(t)-x(t-\tau) = \tau\dot{x}(\xi) = \tau y(\xi-tau) \leq \tau y(t-2\tau).$$

The forth bound follows from $x(t^*) = 0$ and $\tau < \sqrt{2}k/4 < k$, which is part of the sufficient condition for the stability of the core mechanism of BCN. From (25), we can obtain that

$$x(t_3) \leq \frac{C_0}{\sqrt{H_dC}} = \sqrt{\frac{N}{G_dp}}\frac{C_0}{C}. \tag{26}$$

Facing impulse flow of size C_0, when the solution curve reaches the positive part of x axis at the first time, $x(t)$ reaches its maximum, namely $x(t) \leq x(t_3)$, due to the uniformly asymptotic stability of the core mechanism of BCN.

Theorem 6. *If* $\tau \leq \min\{\frac{\pi N}{8C(G_d+\sqrt{G_dpN})}, \frac{\pi}{8(G_iR_uw+\sqrt{G_iR_upC})}, \frac{\sqrt{2}w}{4Cp}\}$, *and an addition flow of size* C_0 *enters into the stable BCN system, the buffer occupancy satisfies* $q(t) \leq \sqrt{\frac{N}{G_dp}}\frac{C_0}{C}$.

Theorem 6 shows that the impacts of impulse flow is not decided by its absolute size, but decided by the proportion between the size of the impulse flow and the link capacity. When the link capacity is larger, the BCN system can tolerate impulse of larger size. Moreover, the ability of disturbance attenuation of BCN is solely decided by its rate decrease

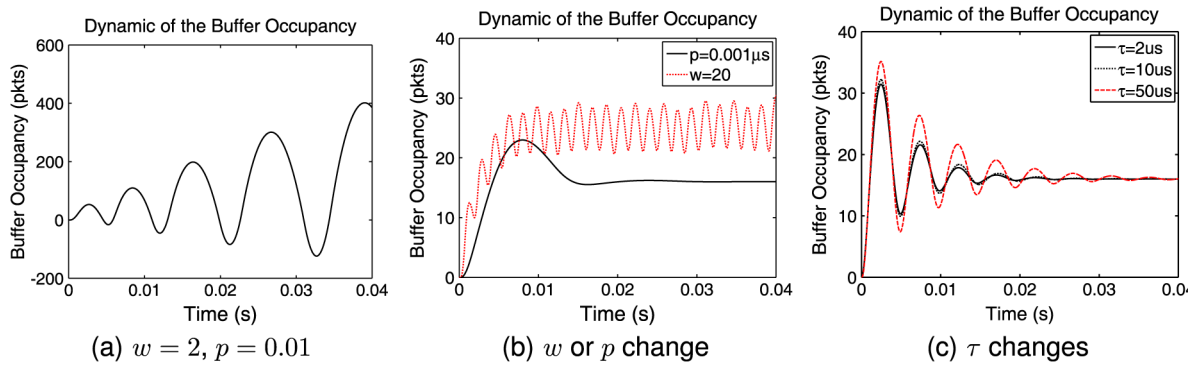

Fig. 3. Solution curves of the delayed differential equations describing the core mechanism of BCN.

subsystem. In addition, note that excepting C_0, all the other parameters of the boundary belongs to the previous stable BCN system. Namely, the maximum buffer occupancy is mainly decided by the size of the impulse flow. This is in accordance with intuition on the premise that the BCN system is asymptotically stable.

6 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, these parameters will stay unchanged and the unitary of link speed is packets per second in the calculation without declared explicitly.

6.1 Numerical Analysis

The same as in [15], we set $C = 10Gbps$, $q_0 = 16pkts$, $N = 50$, $\tau = 200\ \mu s$, all packets are of length $1.5KB$, and use MATLAB to compute the numerical solution of the delayed differential equations describing the core mechanism of BCN.

6.1.1 Impacts of Parameters on Stability

Under the default parameters setting, there is

$$\begin{cases} \frac{\pi N}{8C(G_d+\sqrt{G_d p N})} & = 3.35 * 10^{-4} \\ \frac{\pi}{8(G_i R_u w+\sqrt{G_i R_u p C})} & = 1.69 * 10^{-4} \\ \frac{\sqrt{2}w}{4Cp} & = 8.48 * 10^{-5}. \end{cases} \quad (27)$$

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. 3a. The solution curve differs from the simulation result of [15], 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 results of [15].

According to equations (27), the delay τ is mainly bounded by $\frac{\sqrt{2}w}{4Cp}$. Changing w from 2 to 20 can satisfy Theorem 4. When $w = 20$, the core mechanism of the BCN system is stable as shown in Fig. 3b. When $w = 20$, the variance of the buffer occupancy is taken more seriously when measuring the congestion. Thus, the rate adjustment algorithm becomes more sensitive to the variance of the buffer occupancy. That's why the buffer occupancy at the bottleneck link oscillates as shown in Fig. 3b Note that one of the bounds of delay in Theorem 4 decreases with the increase of w, w can't be enlarged too much.

In the same way, $p = 0.001$ can also move the core mechanism of BCN into the stable state, as shown in Fig. 3b. Note that when p decreases, all the bounds of delay in Theorem 4 increase. Moreover, with the decrease of p, the overhead of signaling feedback messages decreases. On the contrary, the Nyquist-Shannon sampling theorem constraints the smallest value of p. Therefore, p is expected to be set as small as possible, when it is large enough to satisfy the sampling theorem.

If the delay τ can be reduced, the core mechanism of BCN is also stable as shown in Fig. 3c. However, the delay is hard to be reduced. When the link capacity C becomes $100Gbps$, all the bounds of the delay in Theorem 4 decrease. If the delay keeps unchanged, the core mechanism of BCN won't keep stable. For example, in Fig. 4, when $\tau = 20\ \mu s$ and $C = 100Gbps$, the core mechanism of BCN becomes unstable. Consequently, BCN needs to be reconfigured according to Theorem 4.

6.1.2 Impacts of Parameters on Buffer Occupancy

Subsequently, we consider the impacts of parameters on the buffer occupancy in the condition that BCN starts from the initial state and its configuration satisfies Theorem 4. Firstly, τ is changed to be $20\ \mu s$ to move the core 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 1 with the bounds estimated by Theorem 5. In Table 1, all the maximum of buffer occupancy are smaller than the estimated bounds, and all the estimated bounds are less than three times of the

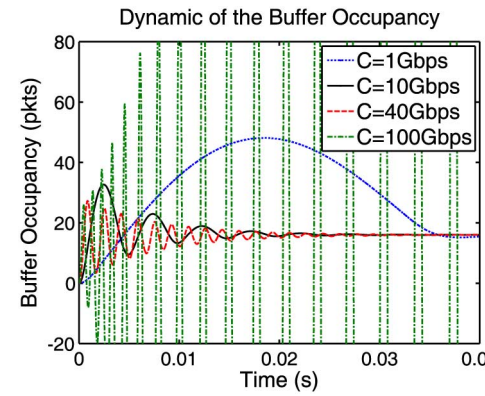

Fig. 4. Solution curves of the delayed differential equations describing the core mechanism of BCN when C changes.

TABLE 1
Maximum Buffer Occupancy

Parameters	Numerical Analysis	Estimation of Theorem 5
$N = 10, 25, 50$	21.6, 26.7, 32.7	32.2, 41.6, 52.1
$G_i = 1, 2, 4$	25.4, 28.6, 32.7	34.1, 41.6, 52.1
$R_u = 1, 4, 8(Mbps)$	32.7, 44.2, 52.6	52.1, 88.3, 118.2
$G_d = \frac{1}{128}, \frac{1}{64}, \frac{1}{32}$	32.7, 26.7, 22.7	52.1, 41.6, 34.1
$C = 1, 10, 40(Gbps)$	48.1, 32.7, 27.3	130.3, 52.1,34.1
$q_0 = 15, 20, 25(pkts)$	30.6, 40.8, 51.0	48.9, 65.2, 81.5
$p = 0.001, 0.002, 0.003$	21.4, 24.8, 28.3	52.1, 52.1, 52.1
$\tau = 2, 10, 50(\mu s)$	31.4, 32.2, 35.1	52.1, 52.1, 52.1
$w = 5, 10, 20$	23.9, 17.5, 16.0	52.1, 52.1, 52.1

corresponding maximum of buffer occupancy. Therefore, the bound estimated by Theorem 5 is reasonable. Similar numerical results, consisting with equation (26), are obtained, but omitted due to the limited space.

The buffer size suggested by Theorem 5 is larger than the bandwidth delay 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 17 *pkts*, the same as the bandwidth delay product according to the rule-of-thumb, and 53 *pkts* according to Theorem 5. Sizing buffer according to Theorem 5, the buffer will be neither full nor empty in the BCN system working normally.

6.2 Experiments

Although the numerical analysis shows that our estimation results are tight, the physical constraints of buffer are not included in the delayed differential equations. Therefore, we also implement the core mechanism of BCN and a delay module on the NetFPGA platform to verify our theoretical results. However, experiments on complex topology are left as the further work due to the limited number of ports of the NetFPGA board.

Firstly, the 2-sources dumbbell topology is used. In experiments, $C = 1Gbps$, $q_0 = 64pkts$, $B = 256pkts$ 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 change the delay τ by reconfiguring the delay modules. The dynamics of the buffer occupancy at the bottleneck link are shown in Fig. 5. When $\tau = 500\ \mu s > 330\ \mu s$, the buffer becomes empty frequently and accordingly the link utilization will degrade, the core mechanism of BCN becomes unstable. On the contrary, when Theorem 4 holds, the core mechanism of BCN is stable, as shown in Fig. 5a. These experimental results are consistent with Theorem 4. In addition, the maximum buffer occupancy is about 115, when the delay $\tau < 330\ \mu s$. This is a little larger than the theoretical bound 93, provided by Theorem 5. Taking the random factors in hardware experiment into account, we judge that these two results are consistent.

Next, after the BCN system reaches the stable state, a background flow of fixed size 500*Mbps* grabs the bandwidth of the bottleneck link. The corresponding dynamics of the buffer occupancy are shown Fig. 6. In this condition, according to Theorem 4, the delay τ should be smaller than $379\ \mu s$ for the stability of the core mechanism of BCN. In other word, the stable region of the BCN system is enlarged. This theoretical result can be inferred by comparing Fig. 6 with Fig. 5. Besides, the same as the prediction of Theorem 5, the buffer occupancies increase when the available bandwidth decreases.

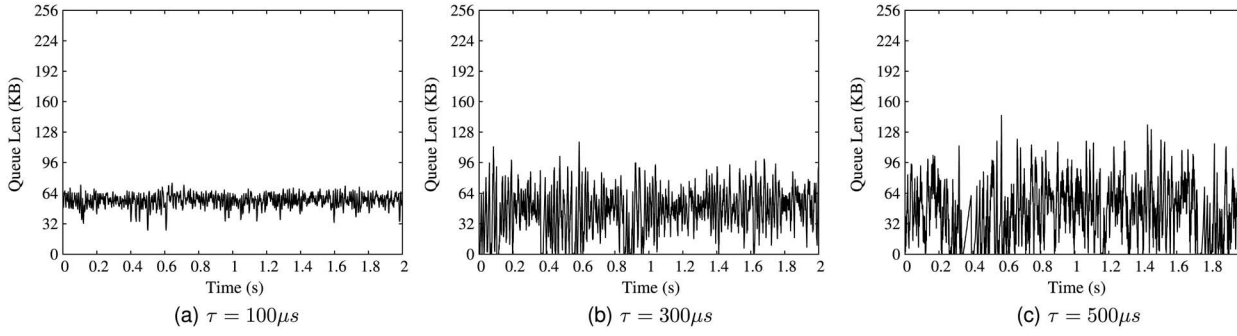

Fig. 5. Dynamics of the buffer occupancy at the bottleneck link when the delay τ changes.

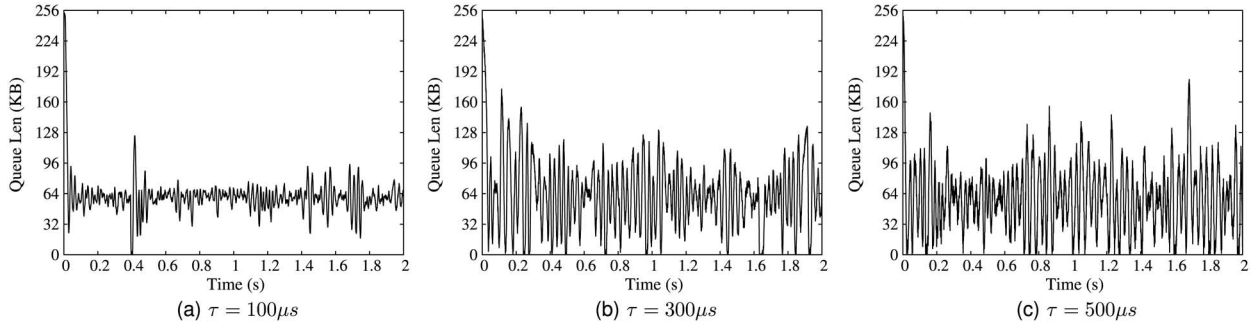

Fig. 6. Dynamics of the buffer occupancy at the bottleneck link with background flow of size 500*Mbps*.

7 Conclusion

BCN is the basic mechanism, which radicates the framework of the end-to-end congestion management scheme in DCE. In this paper, we theoretically analyze the BCN system, paying special attention on the delay. We reveal that the core mechanism of BCN is stable when the delay is bounded. When the speed of Ethernet increases to 40*Gbps* or 100*Gbps* 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, and also measure BCN's ability of disturbance attenuation. The numerical analysis and the experiments on the NetFPGA platform verify our theoretical analysis. Similar to BCN, when the speed of Ethernet becomes 40*Gbps* or 100*Gbps*, the impacts of delay on the performance of other congestion management schemes should be taken seriously too.

Acknowledgment

The authors gratefully acknowledge the anonymous reviewers for their constructive comments. This work is supported in part by National Basic Research Program of China (973 Program) under Grant No. 2014CB347800 and 2012CB315803, and National Natural Science Foundation of China (NSFC) under Grant No. 61225011.

References

[1] "IEEE 802.1: Data Center Bridging Task Group," http://www.ieee802.org/1/pages/dcbridges.html.
[2] C. DeSanti and J. Jiang, "FCoE in Perspective," *Proc. Int'l Conf. Advanced Infocomm Technology*, pp. 138:1-138:8, 2008.
[3] K.Z. Meth and J. Satran, "Design of the iSCSI Protocol," *Proc. Mass Storage Systems and Technologies*, pp. 116-122, 2003.
[4] D. Cohen, T. Talpey, A. Kanevsky, U. Cummings, M. Krause, R. Recio, D. Crupnicoff, L. Dickman, and P. Grun, "Remote Direct Memory Access over the Converged Enhanced Ethernet Fabric: Evaluating the Options," *Proc. High Performance Interconnects*, pp. 123–130, 2009.
[5] B. Goglin, "Design and Implementation of Open-MX: High-Performance Message Passing over Generic Ethernet Hardware," *Proc. Parallel and Distributed Processing*, 2008, pp. 1-7.
[6] "IEEE 802.1Qbb: Priority-Based Flow Control, Working Draft," http://www.ieee802.org/1/pages/802.1bb.html.
[7] G.F. Pfister and V.A. Norton, "Hotspot Contention and Combining in Multistage Interconnection Networks," *IEEE Trans. Computers*, vol. 34, no. 10, pp. 933-938, Oct. 1985.
[8] "IEEE 802.1Qau: End-to-end Congestion Management, Working Draft," http://www.ieee802.org/1/pages/802.1au.html.
[9] "IEEE P802.3ba: 40Gb/s and 100Gb/s Ethernet Task Force," http://www.ieee802.org/3/ba/index.html.
[10] "Juniper Networks T-series Routing Platforms: T320, T640, T1600, and TX Matrix, data sheet," http://egctechnologies.com/uploaded/T-Series%20Routers.pdf, 2007.
[11] "Brocade MLX Series Routers, Data Sheet," http://www.brocade.com/downloads/documents/data_sheets/product_data_sheets/brocade-mlx-series-ds.pdf, 2014.
[12] D. Bergamasco and R. Pan, "Backward Congestion Notification Version 2.0," http://www.ieee802.org/1/files/public/docs2005/new-bergamasco-bcn-september-interim-rev-final-0905.ppt, Sept. 2005.
[13] "NetFPGA Project," http://netfpga.org/.
[14] J. Jiang and R. Jain, "Analysis of Backward Congestion Notification (BCN) for Ethernet in Datacenter Applications," *Proc. IEEE Int'l Conf. Computer Comm. (INFOCOM) Minisymp.*, pp. 2456-2460, 2007.
[15] Y. Lu, R. Pan, B. Prabhakar, D. Bergamasco, V. Alaria, and A. Baldini, "Congestion Control in Networks with no Congestion Drops," *Proc. 44th Allerton Ann. Conf. Comm. Control and Computing*, pp. 891-898, Sept. 2006.
[16] F. Ren and W. Jiang, "Phase Plane Analysis of Congestion Control in Data Center Ethernet Networks," *Proc. Int'l Conf. Distributed Computing Systems (ICDCS)*, pp. 20-29, June 2010.
[17] B. Kwan and J. Ding, "Effects of Delay: Output Generated Multistage and Symmetric Topology w/ Single Hot Spot Scenarios," http://www.ieee802.org/1/files/public/docs2007/au-kwan-ding-bcn-effects-of-delay-02152007.pdf, 2007.
[18] R.M. Brooks and K. Schmit, "The Contraction Mapping Principle and Some Applications," *Electronic J. Differential Equations*, Monograph 09, pp. 15-17, 2009.
[19] V. Misra, W.B. Gong, and D. Towsley, "Fluid-Based Analysis of a Network of AQM Routers Supporting TCP Flows with an Application to RED," *Proc. ACM Special Interest Group on Data Comm. (ACM SIGCOMM)*, pp. 151-160, Aug. 2000.
[20] M. AI-Fares, A. Loukissas and A. Vahdat, "A Scalable, Commodity Data Center Network Architecture," *Proc. ACM Special Interest Group on Data Comm. (ACM SIGCOMM)*, pp. 63–74, Aug. 2008.
[21] C. Guo, G. Lu, D. Li, H. Wu, X. Zhang, Y. Shi, C. Tian, Y. Zhang and S. Lu, "BCube: A High Performance, Server-Centric Network Architecture for Modular Data Centers," *Proc. ACM Special Interest Group on Data Comm. (ACM SIGCOMM)*, Aug. 2009.
[22] P.J. Braam, "File Systems for Clusters from a Protocol Perspective," *Proc. Second Extreme Linux Topics Workshop*, http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.39.2657&rep=rep1&type=pdf, June. 1999, 5 sheets.
[23] B. Welch, M. Unangst, Z. Abbasi, G. Gibson, B. Mueller, J. Zelenka and B. Zhou, "Scalable Performance of the Panasas Parallel File System," http://www.lustre.org, Feb. 2008.
[24] S. Kandula, S. Sengupta, A. Greenberg, P. Patel, and R. Chaiken, "The Nature of Datacenter Traffic: Measurements & Analysis," *Proc. Integrated Marketing Comm. (IMC)*, 2009.
[25] E.A. Coddington and N. Levionson, *Theory of Ordinary Differential Equations*. McGraw Hill, 1975.
[26] R.D. Driver, *Ordinary and Delay Differential Equations*. Springer-Verlag, 1977, pp. 384-398.

Wanchun Jiang received the BE degree in computer science and technology from Tsinghua University, Beijing, China, in 2009 where he is pursuing the PhD degree. He is supervised by Prof. Fengyuan Ren and Prof. Chuang Lin now. His research interests include congestion control, data center networks, and the application of control theory in computer networks.

Fengyuan Ren received the BA and MSc degrees in automatic control and the PhD degree in computer science from Northwestern Polytechnic University, Xian, China, in 1993, 1996, and December 1999, respectively. He is a professor in the Department of Computer Science and Technology at Tsinghua University, Beijing, China. From 2000 to 2001, he worked in the Electronic Engineering Department of Tsinghua University as a postdoctoral researcher. In January 2002, he moved to the Computer Science and Technology Department of Tsinghua University. His research interests include network traffic management, control in/over computer networks, wireless networks, and wireless sensor networks. He (co)-authored more than 80 international journal and conference papers. He is a member of the IEEE, and has served as a technical program committee member and local arrangement chair for various IEEE and ACM international conferences.

Yongwei Wu received the PhD degree in applied mathematics from the Chinese Academy of Sciences, China, in 2002. He is presently a Professor with Computer Science and Technology, Tsinghua University, Beijing, China. His research interests include data center, cloud computing, distributed processing, and parallel computing.

Chuang Lin received the PhD degree in computer science from Tsinghua University, Beijing, China in 1994. He is a professor in the Department of Computer Science and Technology at Tsinghua University, Beijing, China. He is an honorary visiting professor at the University of Bradford, U.K. His current research interests include computer networks, performance evaluation, network security analysis, and Petri net theory and its applications. He has published more than 300 papers in research journals and IEEE conference proceedings in these areas and has published four books. He is a senior member of the IEEE and the Chinese Delegate in TC6 of IFIP. He served as the Technical Program Vice Chair, the 10th IEEE Workshop on Future Trends of Distributed Computing Systems (FTDCS 2004); the General Chair, ACM SIGCOMM Asia workshop 2005, and the 2010 IEEE International Workshop on Quality of Service (IWQoS 2010). He is an Associate Editor of the *IEEE Transactions on Vehicular Technology* and an area editor of *Computer Networks*, and the *Journal of Parallel and Distributed Computing*.

Ivan Stojmenovic received the PhD degree in mathematics. He has held regular and visiting positions in Serbia, Japan, USA, Canada, France, Mexico, Spain, U.K. (as chair in Applied Computing at the University of Birmingham), Hong Kong, Brazil, Taiwan, and China, and is a full professor at the University of Ottawa, Canada, and adjunct professor at the University of Novi Sad, Serbia. He has published over 300 different papers, and edited 7 books on wireless, ad hoc, sensor and actuator networks, and applied algorithms with Wiley. He is an editor of over dozen journals, editor-in-chief of the *IEEE Transactions on Parallel and Distributed Systems* (from January 2010 through December 2013), and founder and editor-in-chief of three journals (*Multiple-Valued Logic and Soft Computing*, *International Journal of Power Electronics and Drive Systems*, and *Ad Hoc & Sensor Wireless Networks*). He is one of about 250 computer science researchers with h-index at least 50, has top h-index in Canada for mathematics, and has $>11{,}000$ citations. He received four best paper awards and the Fast Breaking Paper for October 2003, by Thomson ISI ESI. He is a recipient of the Royal Society Research Merit Award, U.K., a Tsinghua 1000 Plan distinguished professor (2012-2015), and a fellow of the IEEE (Communications Society, class 2008), and Canadian Academy of Engineering (since 2012). He was an IEEE CS distinguished visitor 2010-2011. He received Excellence in Research Award of the University of Ottawa, 2009. He chaired and/or organized >60 workshops and conferences and served in >200 program committees. He was a program co-chair at IEEE PIMRC 2008, IEEE AINA-07, IEEE MASS-04&07, EUC-05&08-10, AdHocNow08, IFIP WSAN08, WONS-05, MSN-05&06, ISPA-05&07, founded workshop series at IEEE MASS, ICDCS, DCOSS, WoWMoM, ACM Mobihoc, IEEE/ACM CPSCom, FCST, MSN, and is/was workshop chair at IEEE INFOCOM 2011, IEEE MASS-09, and ACM Mobihoc-07&08.