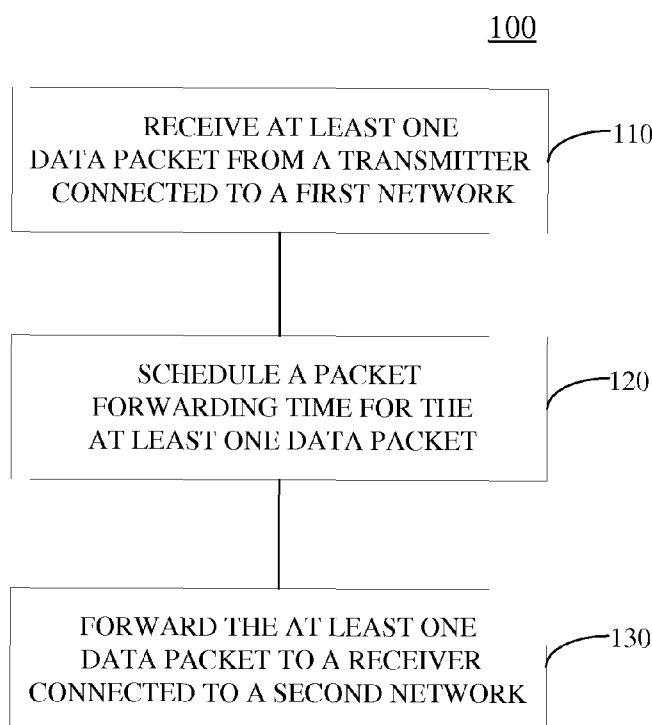




- (51) International Patent Classification:
H04L 12/66 (2006.01) *H04L 29/02* (2006.01)
- (21) International Application Number:
PCT/CN2011/084762
- (22) International Filing Date:
27 December 2011 (27.12.2011)
- (25) Filing Language: English
- (26) Publication Language: English
- (30) Priority Data:
61/427,675 28 December 2010 (28.12.2010) US
- (71) Applicant (for all designated States except US): **THE CHINESE UNIVERSITY OF HONG KONG** [CN/CN]; Shatin, N.T., Hong Kong (CN).
- (72) Inventors; and
- (75) Inventors/Applicants (for US only): **LEE, Yiu Bun** [CN/CN]; 12B, Harbourview Garden, 337 Prince Edward Road West, Kowloon City, Hong Kong (CN). **WAN, Wing San** [CN/CN]; Rm. 2505, 25/F, Ming Chau House, Kin Ming Estate, Tseung Kwan O., NT, Hong Kong (CN).
- (74) Agent: **INSIGHT INTELLECTUAL PROPERTY LIMITED**; 19 A, 19B, Tower A, InDo Building, No. 48A Zhichun Road, Haidian District, Beijing 100098 (CN).
- (81) Designated States (unless otherwise indicated, for every kind of national protection available): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.
- (84) Designated States (unless otherwise indicated, for every kind of regional protection available): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

[Continued on next page]

- (54) Title: SYSTEMS AND METHODS TO IMPROVE PERFORMANCE OF TCP OVER LARGE BANDWIDTH-DELAY-PRODUCT NETWORKS



- (57) Abstract: The apparatus, systems, and methods described herein may operate to receive at least one data packet from a transmitter connected to a first network. A packet forwarding time may be scheduled for the at least one data packet. The at least one data packet may be forwarded to a receiver connected to a second network based on the packet forwarding time such that transmission of the at least one data packet prevents a buffer overflow at the receiver.

FIG 4



Published:

— *with international search report (Art. 21(3))*

SYSTEMS AND METHODS TO IMPROVE PERFORMANCE OF TCP OVER LARGE BANDWIDTH-DELAY-PRODUCT NETWORKS

Cross Reference to Related Applications

[0001] The present application claims the priority benefit of U.S. Provisional Patent Application Serial No. 61/427,675 filed December 28, 2010 and entitled "SYSTEMS AND METHODS TO IMPROVE PERFORMANCE OF TCP OVER LARGE BANDWIDTH-DELAY-PRODUCT NETWORKS," of which application is incorporated herein by reference in its entirety.

Background

[0002] The emergence of mobile data networks, such as 3G, 3.5G, and 4G is progressively reshaping the Internet landscape from almost entirely wired broadband Internet users, to more and more users connected via wireless links. Although widely used as a core Internet transport protocol for the Internet, Transmission Control Protocol (TCP) has shown performance limitations, especially when operated in mobile data network environments. For example, three fundamental problems of running TCP in modern mobile data networks have limited achieving a higher speed of data packet transmission and, thus, enhanced user experience: (a) flow-control-limited throughput due to larger bandwidth-delay product (BDP); (b) wireless link capacity estimation; and (c) false congestion avoidance due to random loss.

Brief Description of the Drawings

[0003] FIG. 1 shows the dynamics of reported window size according to various embodiments.

[0004] FIG. 2 shows stop-and-go behavior due to large BDP compared with $Awnd$ according to various embodiments.

[0005] FIG. 3 shows an emulation platform according to various embodiments.

[0006] FIG. 4 shows a flow chart illustrating various methods to transmit and receive data packets according to various embodiments.

[0007] FIG. 5 shows a flow chart illustrating various methods to transmit and receive data packets according to various embodiments.

[0008] FIG. 6A shows CWnd and Throughput dynamics without accelerator according to various embodiments; FIG. 6B shows CWnd and Throughput dynamics of the first 15 seconds with sender's flow control mechanism disabled according to various embodiments; and FIG. 6C shows CWnd and Throughput dynamics with sender's flow control mechanism disabled according to various embodiments.

[0009] FIG. 7 shows a transmission timing diagram between receiver and accelerator according to various embodiments.

[0010] FIG. 8 shows the evolution of the AWnd, CWnd, Virtual AWnd, buffer occupancy, sending rate and throughput achieved with transmission scheduling, according to various embodiments.

[0011] FIG. 9 shows the dynamic of CWnd and throughput in the presence of packet loss according to various embodiments of the invention.

[0012] FIG. 10 illustrates the slow recovery of discarded packets caused by disabling flow control, according to various embodiments of the invention.

[0013] FIG. 11 shows how proactive retransmission works according to various embodiments.

[0014] FIG. 12 shows pseudo code for data transmission with the local retransmission capability according to various embodiments.

[0015] FIG. 13 shows the evolution of CWnd, throughput and sending rate with proactive retransmission according to various embodiments.

[0016] FIG. 14 shows the RTT incurred by local retransmission according to various embodiments.

[0017] FIG. 15 shows the evolution of modulated RTO (RTOM), RTO and RTT versus time according to various embodiments.

[0018] FIG. 16 shows the evolution of modulated RTO (RTOM), RTO and RTT versus time (enlarged around loss) according to various embodiments.

[0019] FIG. 17 shows the logic of packet scheduling according to various embodiments.

[0020] FIG. 18 shows throughput variation of two competing flows without accelerator according to various embodiments.

[0021] FIG. 19 shows throughput variation of two competing flows with accelerator according to various embodiments.

[0022] FIG. 20 shows throughput variation of two competing flows without accelerator according to various embodiments.

[0023] FIG. 21 shows throughput variation of two competing flows with accelerator according to various embodiments.

[0024] FIG. 22 shows throughput variation of flows with different weight according to various embodiments.

[0025] FIG. 23 shows a practical network setup according to various embodiments.

[0026] FIG. 24 shows loss ratio with 95% confidence interval for data rates of 3-6.5 Mbps according to various embodiments.

[0027] FIG. 25 shows achievable throughput versus various receive buffer size according to various embodiments.

[0028] FIG. 26 shows a block diagram of various systems and apparatuses for transmitting and receiving data packets according to various embodiments.

[0029] FIG. 27 shows a block diagram of an article of manufacture, including a specific machine, according to various embodiments.

Detailed Description

[0030] The emergence of mobile data networks such as 3G, 3.5G, and 4G is progressively reshaping the Internet landscape from almost entirely wired broadband Internet users, to more and more users connected via wireless links. Investigation is made with respect to the performance limitations of the core Internet transport protocol – Transmission Control Protocol (TCP), when operated in mobile data network environments; and various embodiments are proposed to develop a novel network-centric mobile accelerator to overcome such limitations without changing the core TCP transport module at either end of the connection (i.e., server and client). For example, three fundamental problems of running TCP in modern mobile data networks are investigated:

(a) flow-control-limited throughput due to larger bandwidth-delay product (BDP); (b) wireless link capacity estimation; and (c) false congestion avoidance due to random loss, and novel solutions are proposed for implementation in the network-centric accelerator. Experimental results conducted in production 3G HSPA networks show that the accelerator according to various embodiments can increase the throughput performance of TCP by up to about 2.5 times of the unaccelerated TCP. In addition, the proposed accelerator according to various embodiments does not require modification to software (S/W) and/or Internet applications (hereinafter “application”), TCP implementation at the hosts, or operating system; and thus can be readily deployed in current and future mobile data networks.

[0031] Chapter 1: Introduction

[0032] The emergence of mobile data networks, such as 3G, 3.5G, and 4G is progressively reshaping the Internet landscape from almost entirely wired broadband Internet users, to more and more users connected via wireless links. As demonstrated below in detail, this change could have profound impact on the performance of many Internet applications and services given the many fundamental differences between wired and wireless networks. Various embodiments of the invention investigate the performance limitations of the core Internet transport protocol – Transmission Control Protocol (TCP), when operated in mobile data network environments; and propose a novel network-centric mobile accelerator to overcome such limitations without changing the core TCP transport module at either end of the connection (i.e., server and client).

[0033] Specifically, three fundamental problems of running TCP in modern mobile data networks are investigated: (a) flow-control-limited throughput due to larger bandwidth-delay product (BDP); (b) wireless link capacity estimation; and (c) false congestion avoidance due to random loss. For (a), various embodiments are proposed to develop a novel virtual advertised window mechanism to decouple flow control between the client and the server so that the server’s transmission throughput is not limited by the receiver advertised window size. For (b), various embodiments are proposed to develop a rate-based algorithm (as oppose to credit-based approach in TCP) to continuously estimate and adapt the data transmission rate over the mobile data network. For (c),

various embodiments are proposed to develop a novel RTT modulation technique to effectively suppress packet loss events from the TCP sender so that high throughput can be sustained despite the existence of random packet loss.

[0034] Unlike previous works which rely on modifications to the TCP protocol stack at the sender, at the receiver, at both ends, or at the network, various embodiments are proposed to develop a network-centric approach where the above-mentioned protocol processing algorithms may be implemented in a network device - mobile accelerator, which in turn may be deployed at the edge of the mobile data network linking it to the rest of the Internet.

[0035] This network-centric approach has, for example, four significant advantages over existing technologies. First, it may eliminate the need to modify the TCP module in either the TCP sender or the TCP receiver. Considering that many modern operating systems are proprietary and slow to adopting any protocol modifications, this network-centric approach may allow a far more rapid adoption and deployment of the proposed protocol optimizations. Second, even if the TCP module can be modified, some of the optimizations such as local retransmission (c.f., Chapter 5) simply cannot be effectively implemented in the end hosts. Third, the proposed protocol optimization algorithms may be designed specifically for mobile data networks. Thus, if it is implemented in the end host such as the server, it may need to determine the type of network a client is accessing before the appropriate protocol optimizations can be applied. Finally, in practice, many mobile operators may deploy their own web proxies to reduce Internet bandwidth consumption and to implement proprietary value added services. Consequently, the Internet server (e.g., web server) may not even be communicating with the client device directly at all. In this case the TCP connection between the server and the proxy may be independent from the TCP connection between the proxy and the client. Hence protocol optimization performed at the server end may never reach the mobile data network, rendering it useless.

[0036] In addition to TCP throughput improvement, the proposed mobile accelerator according to various embodiments may also open up a new platform for implementing sophisticated network traffic controls. For example, various embodiments

are proposed to develop a packet scheduling mechanism which can perform precise bandwidth allocation among competing TCP flows.

[0037] The rest of the instant patent application is organized as follows: [0038] reviews previous related works. [0053] investigates the performance problem of TCP's flow control mechanism in mobile data networks. Chapter 4 presents the virtual advertized window mechanism to overcome TCP's flow control bottleneck. Chapter 5 investigates packet loss in mobile data networks and presents a local retransmission algorithm. Chapter 6 presents the RTT modulation mechanism for loss event suppression. Chapter 7 presents a packet scheduling mechanism for precise bandwidth allocation. Chapter 8 reports performance of the proposed mobile accelerator in production 3G networks. Chapter 9 presents a machine-readable medium and Chapter 10 concludes the study.

[0038] Chapter 2: Background and Related Work

[0039] Much research had been done to improving the performance of TCP in large BDP networks. The existing works may be classified into four categories: modifying both sender and receiver; modifying the sender only; modifying the receiver only; and network-centric approach.

[0040] 2.1 Sender-receiver-based approaches

[0041] Approaches where both the sender and the receiver are modified are considered first. Jacobson et al. proposed in RFC 1323 the LWS extension to TCP which is currently the most widely supported solution. It works by scaling the Advertised Window (AWnd) by a constant factor throughout the connection. With the maximum LWS factor 14, the maximum AWnd can be increased up to about 1 GB $((2^{16}-1)*2^{14} \approx 2^{30})$. The strengths and weaknesses of LWS will be discussed in more detail in [0053].

[0042] Alternatively, an application(s) can be modified to initiate multiple TCP connections in parallel to increase throughput by aggregating multiple TCP connections. This approach effectively multiplies the AWnd and Congestion Window (CWnd) by the number of TCP flows and so can mitigate the AWnd limitation. However, aggregating multiple TCP connections will also allow the application(s) to gain unfair amount of

bandwidth from competing TCP flows and the aggregate CWnd may increase too rapidly. Hacker et al. solved this problem by deferring CWnd increase until multiple ACKs are received so as to compensate for the rapid CWnd growth.

[0043] 2.2 Sender-based approaches

[0044] Apart from AWnd limit, the CWnd maintained by the sender may also limit TCP's throughput in large BDP networks. Specifically, the growth of the CWnd is triggered by the reception of ACKs. Thus, in a long delay path, it may take longer time for the CWnd to grow to a sufficiently large value so that the link bandwidth can be fully utilized.

[0045] To tackle this problem, Allman et al. proposed in RFC 3390 to initialize the CWnd to a larger value (as opposed to one TCP segment) so that it can grow more quickly in large delay networks to ramp up TCP's throughput. Since then, much effort had been put into developing more sophisticated congestion control algorithms, such as CUBIC, BIC, FAST, and H-TCP to further shorten TCP's throughput ramp up time and provide better fairness among different TCP connections. These solutions tackled the limitation of CWnd growth.

[0046] Another problem affecting the CWnd and throughput is the existence of non-congestion-induced packet loss in mobile data networks. As TCP would reduce CWnd upon detection of packet loss, its throughput performance will be severely degraded in the presence of non-congestion-induced packet losses. Lai et al. proposed a TCP-NCL protocol for wireless networks such as WiFi to differentiate between congestion loss from random packet loss; and to handle packet reordering in case the link layer also retransmit packets. Nevertheless, the performance of these sender-based approaches in mobile data networks is still subject for further investigation, and in cases where the mobile network implemented proxy servers, then they will have no effect at all as the sender will only be communicating with the proxy server.

[0047] 2.3 Receiver-based approaches

[0048] At the receiving end, Fisk and Feng proposed dynamic right-sizing of the AWnd by estimating the CWnd at the receiver and then dynamically adapt the receiver buffer size, i.e., the AWnd, to twice the size of the estimated CWnd. This ensures that

when the sender's CWnd doubles (behavior of TCP New Reno after receiving an ACK), the AWnd will not become the bottleneck.

[0049] More recent Operating Systems such as Linux[®] 2.4 and Microsoft[®] Windows[®] Vista[®] also implemented receiver buffer size auto-tuning by estimating the BDP of the connection and the data consumption rate of the application. In comparison, the accelerator according to various embodiments of the invention does not require any modification to the receiver application or require support from the receiver operating system, and so can be more readily deployed by an ISP or a satellite operator to accelerate all bandwidth demanding TCP traffics.

[0050] 2.4 Network-centric approach

[0051] The fourth approach is to implement protocol optimizations within the network. The Snoop protocol adopts this approach by keeping copies of transmitted packets in the base station's cache, and then retransmitting them on behalf of the sender when packet loss occurs. However, the extra delay incurred by the retransmission process could also trigger the sender to timeout which in turn will lead to severe throughput degradation. Various embodiments of the invention tackled this problem through a novel RTT modulation mechanism (c.f. Chapter 6) which can actively prevent sender timeout in case of local retransmission.

[0052] Another study by Hu and Yeung proposed a new active queue management protocol whereby the network device such as a router or a base station, generates duplicate ACKs to trigger CWnd reduction at the sender whenever the device's packet queue length exceeds a certain threshold, thus preventing congestion from taking place.

[0053] Chapter 3: TCP flow control revisited

[0054] TCP's built-in flow control mechanism is designed to prevent a fast sender from overflowing a slow receiver. It works by reporting the receiver's buffer availability, i.e., the AWnd, back to the sender via a 16-bit field inside the TCP header so that the sender would not send more data than the receiver's buffer can store.

[0055] Over the years, computer processing power has grown tremendously such that even today's modest computers can easily keep up with the arriving stream of data at

relatively high data rates (e.g., tens of Mbps). Thus, an arrived packet will quickly be retrieved by an application from the receiver buffer, and in most cases this can be completed even before the next packet arrives. As a result, the reported AWnd simply stays at the maximum receiver buffer size as illustrated in FIG. 1 which shows the actual AWnd of a receiver at a throughput of about 1.5 Mbps according to various embodiments. In this case, TCP's flow control mechanism is clearly not necessary as the sender never transmits data faster than the receiver's processing rate.

[0056] Due to the delayed AWnd, the sender cannot send more than the reported AWnd and thus cannot make use of the new freed buffer space at the receiver. In cases where the BDP is larger than the maximum AWnd, the sender will operate in a stop-and-go manner, as illustrated in FIG. 2, resulting in severe underutilization of the network channel.

[0057] For example, a 3G HSPA data network has bandwidth ranging from 3.6 Mbps to 14.4 Mbps. Take the commonly offered bandwidth 7.2 Mbps with round-trip-delay (RTT) of 100 ms as an example. This set of network parameters will lead to a BDP of 90 KB which already exceeds TCP's maximum advertised window size of 64 KB. In this case TCP's flow control mechanism will limit the throughput to no more than 5.1 Mbps which leave about 29% of bandwidth unused even if there are no competing traffics in the network.

[0058] The conventional solution – TCP's Large Window Scale extension, was designed specifically to address this limitation by allowing the two hosts of a TCP connection to negotiate a constant multiplying factor to be applied to the AWnd value during connection setup so that AWnd larger than 64 KB can be used. Obviously, this requires both communicating hosts to support and activate the use of LWS prior to connection setup is carried out. Unfortunately, although most modern operating systems support LWS, there is no standard way for the application to activate the use of LWS. Some operating systems activate LWS by default (e.g., Windows Vista), some requires manual configuration (e.g., Windows XP), and yet others require an application to explicitly configure a large (e.g., >64 KB) socket buffer to activate it.

[0059] For example, the common operating system such as Microsoft® Windows® XP® defaults its advertised window size to 17 KB for connections with link capacity at or

below 10 Mbps. Running over a 3G HSPA data network, this advertised window size will limit the achievable throughput of TCP to about 1.4 Mbps at a RTT of 100 ms. While applications could configure a larger window size, e.g., by explicitly setting a larger socket buffer size, the inventors found that this is rarely done in practice. In fact, almost all common Internet applications evaluated on the Microsoft® Windows® platform, including web browsers (Internet Explorer, Firefox®, Google® Chrome®), FTP clients (Windows' built-in FTP client), and email clients (Outlook® Express), simply employ the operating system's advertised window size (e.g., 17 KB) and do not make use of TCP's LWS extension, therefore limiting their throughput performance in networks with large bandwidth-delay product.

[0060] Another limitation of LWS is that the AWnd value is still interpreted as the amount of buffer physically available at the receiver (and sender as well). Thus, for networks with very large BDP, the resultant buffer requirement can be very large as well (e.g., with LTE's 172.8 Mbps over 100 ms RTT a receiver buffer size of 21.6 MB will be needed). This could become a problem for applications that make use of large number of sockets (e.g., server applications) or for mobile devices with limited physical memory.

[0061] If FIG. 1 is reconsidered, then it is expected that such a large buffer at the receiver will end up mostly unused anyway as the received data will be retrieved out of the TCP receive buffer quickly by the application. Thus, instead of strictly following the reported receiver buffer size, various embodiments of the invention propose to reinterpret it as an indicator of receiver processing capacity and employ it for use in a rate-based flow control mechanism to be discussed in detail in the next Chapters.

[0062] Chapter 4: Opportunistic Transmission

[0063] Results from Chapter 3 reveal that the AWnd may turn out to become the bottleneck to achieving high throughput in modern mobile data networks. A trivial solution may be to eliminate the AWnd altogether as it is no longer serving its intended function.

[0064] 4.1 General Description

[0065] To experiment with this idea, according to an embodiment of the present application, a computer-implemented method 100 is provided. As shown in FIG. 4, the

method 100 includes receiving at least one data packet from a transmitter connected to a first network in step 110, scheduling a packet forwarding time for the at least one data packet in step 120, and forwarding the at least one data packet to a receiver connected to a second network based on the packet forwarding time such that transmission of the at least one data packet prevents a buffer overflow at the receiver in step 130.

[0066] According to an embodiment of the present application, the number of the at least one data packets to be forwarded to the receiver can be determined and informed to the transmitter. In an example, the number of the at least one data packets to be forwarded to the receiver is calculated based on the amount of buffers free-up at the receiver and the amount of data packets received at the receiver during a round-trip-delay during a round-trip-delay. In this way, the buffers free-up at the receiver can be utilized and the transmission efficiency is increased. The number of the data packets can be informed to the transmitter by dynamically adjusting a size of the AWnd based on the number of the at least one data packet to be forwarded to the receiver determined by the transmitter.

[0067] According to an embodiment of the present application, as shown in FIG. 5, in order to schedule the packet forwarding time, a packet forwarding delay for the at least one data packet can be calculated in step 121. The packet forwarding delay can be calculated based on a packet reception rate by the receiver in step 122. The packet reception rate can be calculated based on a difference in acknowledged sequence numbers associated with the at least one data packet in step 123.

[0068] According to an embodiment of the present application, when detecting a lost data packet in step 140, the out-of-window packets to be discarded by the receiver can be determined in step 150. Then, the out-of-window packets are retransmitted to the receiver before the transmitter times out and retransmits one of the out-of-window packets in step 160. In an example, the out-of-window packets are retransmitted based on suppressing forwarding a duplicate acknowledgement (ACK) received from the receiver to the transmitter. In another example, the out-of-window packets are retransmitted based on delaying forwarding an acknowledgement (ACK) received from the receiver to the transmitter. The forwarding of the ACK can be delayed by increasing an amount of delayed time according to a number of retransmission attempts.

[0069] According to an embodiment of the present application, in order to schedule the packet forwarding time, a weight is assigned to each data flow, the weight indicating a number of packets to be forwarded for each data flow.

[0070] 4.2 Detailed Description of Opportunistic Transmission

[0071] To experiment with this idea, in various embodiments, an accelerator may be implemented between the sender and the receiver as illustrated in FIG.3. The accelerator's function may be, for example, to modify the AWnd field inside the TCP header of ACK packets from the receiver, to a value much larger than the BDP of the network to prevent it from becoming the bottleneck. In this experiment, for example, the Linux Netem module may be used to emulate a typical 3G HSPA data network. The emulation parameters including network delay and bandwidth may be obtained from measurement of a production 3G HSPA data network.

[0072] The sender, the accelerator, and the emulator may all run on Linux[®] with kernel 2.6 (with CUBIC as the default TCP congestion control module) and the receiver may run on Windows[®] XP[®] SP2. This network configuration may have a BDP of 90 KB which is larger than Windows XP's default receiver buffer size of 17 KB. The receiver may initiate the connection and then the sender may keep sending data to the receiver as fast as TCP allows.

[0073] As a benchmark for comparison, an experiment may be conducted without any modification to the AWnd reported by the receiver FIG. 6A. shows the evolution of AWnd, CWnd, and throughput (as measured by the receiver) of the first 20 seconds of the experiment according to various embodiments. As expected, the AWnd may stay at the maximum value at all times. The CWnd, once increased passing the AWnd value, may stop increasing further even though there is no packet loss during the experiment. As a result, the achievable throughput may be limited by the window size and, as in this case, may be significantly lower than the network's capacity (i.e., about 1.41 Mbps versus about 7.2 Mbps).

[0074] In the second experiment, the accelerator may modify the AWnd reported by the receiver to 10 MB, regardless of the actual value of the AWnd reported. This may effectively disable TCP's flow control mechanism as the modified AWnd size may be larger than the largest value reached by the CWnd in the experiment.

[0075] FIG. 6B shows the evolution of CWnd and throughput for the first 15 seconds of the experiment according to various embodiments. Compared to FIG. 6A, it is observed that the CWnd may no longer be constrained by the AWnd and thus can continue to grow all the way up to about 3 MB. Similarly, the achievable throughput also continued to increase, only to be limited by the network's link capacity at 7.2 Mbps. However, the high throughput did not last very long and it abruptly dropped to zero after about 10 s.

[0076] Not surprisingly, this abrupt drop may be due to network congestion as the large CWnd allowed the sender to transmit at a rate even higher than the network link capacity. In fact, TCP was designed to induce congestion events in its probe for additional bandwidth in the network and react to congestion by lowering its transmission rate (via reduction in CWnd). What is unexpected is the drop to nearly zero throughput after the loss event. Moreover, this condition may continue for an extensive duration (over 600 s), as illustrated in FIG. 6C, before the TCP flow can recover from it. This suggests that TCP congestion control algorithm may not function well without its flow control mechanism.

[0077] To understand the cause of the blackout period after a loss event, in various embodiments, the behavior of the receiver may be considered. Normally, when there is no packet loss, TCP segments arriving at the receiver may be quickly processed and passed to the application, thus releasing the occupied buffer space. However, when a packet loss occurs, all the subsequent data may need to be buffered until the lost packet is successfully retransmitted as TCP guarantees in-sequence data delivery. In normal TCP, the sender may stop transmitting data once the AWnd is used up and thus buffer overflow may never occur. Now, in case the AWnd is set to a value larger than the actual one, the sender may continue to send data, even beyond the buffer space available at the receiver. Consequently, these out-of-bound packets may all be discarded by the receiver due to buffer overflow, resulting in a very large burst of packet loss.

[0078] This may explain the evolution of the CWnd in FIG. 6B. At time 8.49 s, the sender (running the default CUBIC congestion module in Linux) began receiving duplicate ACKs and started to reduce its CWnd linearly at a rate of one per two DUP-ACKs, until it reached a lower limit. This continued until the lost packet timed out, at

which point the CWnd reset to one. Due to the large burst of packet loss induced, numerous timeout events followed and this was the reason for the extensive blackout period in FIG. 6C. During this period, the sender repeatedly timed out and then retransmitted the lost packet, one at a time. Only after all the lost packets were successfully retransmitted, can the CWnd and hence, throughput resume growth.

[0079] The previous experiment clearly shows that the trivial solution of disabling TCP's flow control may end up degrading throughput performance rather than improving it. To prevent such blackout condition from occurring, in various embodiments, a way may be devised to control the packet transmission to prevent congesting the receiver and the network, and to develop a new retransmission algorithm to incorporate the effect of modifying the AWnd. Various embodiments to tackle the former problem will be discussed in the following sections and the retransmission algorithm will be further investigated in Chapter 5.

[0080] 4.2.1 Link bandwidth estimation

[0081] To prevent congesting the network, the accelerator according to various embodiments may regulate the rate at which packets are forwarded to the receiver over the mobile data network. In various embodiments, a system model is developed, for example, to estimate the link bandwidth by assuming that (a) the accelerator always has data to forward; (b) network delays and receiver's processing capability remain constant; (c) network delays of the channel between the accelerator and the receiver are symmetric in the forward and the reverse direction; (d) the receiver generates an ACK immediately upon the arrival of a TCP segment, i.e., zero processing delay; and (e) the uplink, i.e., from the receiver to the accelerator, is not the bottleneck.

[0082] Let rtt be the RTT between the accelerator and the receiver. Obviously, the RTT is not known a priori and thus may need to be estimated from passive measurements of packets traversing the accelerator. Let f_i be the time packet i was forwarded by the accelerator to the receiver, and let t_i be the time at which the corresponding ACK arrived at the accelerator. Then, the RTT as measured by packet i , denoted by rtt_i , may be computed from:

$$rtt_i = t_i - f_i \quad (1)$$

[0083] To smooth out random fluctuations in the rtt_i , the accelerator may apply exponentially weighted moving average to the measured values to obtain the smoothed RTT rtt :

$$rtt = (1 - \alpha) \times rtt + \alpha \times rtt_i \quad (2)$$

where the weight $\alpha = 0.125$ may follow the one used in TCP's internal RTT estimator.

[0084] Next, the link bandwidth may be estimated from the acknowledged sequence number and the arrival time of a pair of ACK packets using a formula similar to, assuming the receiver has infinite processing capacity. Specifically, let t_i be the arrival time of ACK packet i with acknowledged sequence number n_i . Then, for some positive integer k , the link bandwidth, denoted by r_i , may be estimated from

$$r_i = \frac{(n_{i+k} - n_i)}{t_{i+k} - t_i} \quad (3)$$

where the numerator may be the amount of data acknowledged during the time interval (t_i, t_{i+k}) . The intuition behind (3) may be that the receiver, having infinite processing capacity, may generate an ACK packet immediately upon receiving a packet from the network. Thus, the transmission rate may be faster than the link bandwidth. Then, the rate at which packets depart from the network may be determined by the link bandwidth available.

[0085] The parameter k may control the duration of the estimation interval (in number of ACK packets) and may be adjusted to tradeoff between accuracy and timeliness of rate estimation. Also, exponentially weighted moving averaging similar to (2) may be applied to smooth out random fluctuations and obtain the smoothed estimated link bandwidth denoted by r .

[0086] In various embodiments, since the assumption of infinite receiver processing capacity may not be valid in practice, this assumption may be relaxed as discussed in the next section.

[0087] 4.2.2 Reception rate estimation

[0088] When packets arrive at a rate higher than the receiver's processing capacity, the received data may be buffered at the receiver buffer awaiting processing. The lower buffer availability then may result in smaller reported $AWnd$. Thus, by

monitoring the AWnd reported from the receiver, the accelerator may incorporate the receiver's processing capability into the system model described according to (3).

[0089] Specifically, let a_i be the value of AWnd reported by ACK packet i . Then, the amount of data processed by the receiver between time t_i and t_{i+k} may be given by:

$$(a_{i+k} + n_{i+k}) - (a_i + n_i) \quad (4)$$

which is the difference in the acknowledged sequence number plus the difference in the reported AWnd. In case the receiver is slower than the incoming data rate, then a_{i+k} may decrease, thus reducing the amount.

[0090] The reception rate, denoted by R_i , may then be computed from

$$R_i = \frac{(a_{i+k} - a_i) + (n_{i+k} - n_i)}{t_{i+k} - t_i} \quad (5)$$

[0091] Similar to (3), the parameter k may control the width of the estimation interval. In various embodiments, exponentially weighted moving averaging may also be applied to (5), for example, to smooth out the random fluctuations and obtain the smoothed reception rate denoted by R . Note that if the receiver processing capacity is infinite, then $a_{i+k} = a_i$ for all i and k , and (5) may reduce to the special case in (3). The estimated reception rate may incorporate the effect of both link bandwidth availability and receiver processing capacity. This may be used to schedule the forwarding of packets in the accelerator.

[0092] 4.2.3 Transmission scheduling

[0093] Armed with an estimate of the reception rate, transmission of packets from the accelerator to the receiver may be scheduled such that it may not cause buffer overflow at the receiver. The challenge is that the AWnd reported in a ACK packet may be delayed information – for example, it may be the receiver's buffer availability about $0.5rtt$ s ago. During the time the ACK takes to travel to the accelerator, additional TCP segments may arrive at the receiver, and the receiver application may also take out more data from the receive buffer.

[0094] Considering the timing diagram in FIG. 7, in various embodiments, the minimum f_k for the k -th packet may be found such that the expected buffer availability

bk, when the k-th packet arrives at the receiver, is non-negative (i.e., no buffer overflow).
fk may further be expressed as

$$f_k = t_{now} + \Delta t_k \quad (6)$$

where tnow is the current time and Δt_k is the delay in forwarding of the k-th packet.

[0095] Next, in various embodiments, the data arrives and departures may be considered between the time the last ACK was transmitted by the receiver, i.e., ($t_i - 0.5rtt$), and the time packet k arrives at the receiver, i.e., ($f_k + 0.5rtt$). First, during this time, the receiver may continue to process packets and free-up buffers at a rate of R. Thus, the expected amount of buffer space being freed up in this period, denoted by Dproc, may be given by

$$\begin{aligned} D_{proc} &= R(f_k + 0.5rtt - t_i + 0.5rtt) \\ &= R(t_{now} + \Delta t_k - t_i + rtt) \end{aligned} \quad (7)$$

[0096] Similarly, the expected amount of data arriving at the receiver during this period, denoted by Drecv, may be given by

$$D_{recv} = \sum_{\forall \{f_j > (t_i - rtt) \cap f_j \leq f_k\}} q_j \quad (8)$$

[0097] Hence the buffer availability bk at time ($f_k + 0.5rtt$) may be computed from the buffer availability at the beginning of the period, for example, ai, plus the amount of buffer freed up, for example, Dproc, minus the amount of data arrivals, for example, Drecv:

$$\begin{aligned} b_k &= \min \{ a_i + D_{proc} - D_{recv}, a_{max} \} \\ &= \min \left\{ a_i + R(t_{now} + \Delta t_k + 0.5rtt - t_i + 0.5rtt) - \sum_{\forall \{f_j > (t_i - rtt) \cap f_j \leq f_k\}} q_j, a_{max} \right\} \quad (9) \\ &= \min \left\{ a_i + R(t_{now} + \Delta t_k + rtt - t_i) - \sum_{\forall \{f_j > (t_i - rtt) \cap f_j \leq f_k\}} q_j, a_{max} \right\} \end{aligned}$$

where amax is the maximum AWnd of the receiver. Assuming $b_k < a_{max}$, then the packet forwarding delay may be computed from

[0098]

$$\begin{aligned}
0 &\leq b_i \\
0 &\leq a_i + R(t_{\text{now}} + \Delta t_i + rtt - t_i) - \sum_{j \in \{j \mid t_i - rtt \leq t_j \leq t_i\}} q_j \\
R \times \Delta t_i &\geq \sum_{j \in \{j \mid t_i - rtt \leq t_j \leq t_i\}} q_j - a_i - R(t_{\text{now}} + rtt - t_i) \\
\Delta t_i &\geq \max \left[\frac{1}{R} \left(\sum_{j \in \{j \mid t_i - rtt \leq t_j \leq t_i\}} q_j - a_i \right) - (t_{\text{now}} + rtt - t_i), 0 \right] (\because \Delta t_i \geq 0)
\end{aligned} \tag{10}$$

[0099] The accelerator may compute the packet forwarding time for each head-of-line packet in a TCP flow, and transmit them according to a round-robin scheduler in Chapter 7.

[00100] 4.2.4 Performance

[00101] For example, to evaluate the performance of the developed estimation and scheduling algorithms, the experiment illustrated in FIG. 3 may be repeated. In various embodiments, the accelerator may be equipped with the reception rate estimation algorithm and schedule packet forwarding according to the transmission scheduler, for example, described in Section 4.2.3.

[00102] FIG. 8 shows the evolution of the AWnd, CWnd, Virtual AWnd, buffer occupancy at accelerator, sending rate and the throughput achieved by the TCP flow, according to various embodiments. Unlike the previous experimental results in FIG. 5, the bottleneck link may not be congested even though the sender's CWnd grows to very large value (up to about 10 MB). As a result, the TCP throughput may reach and sustain at the link's capacity limit. Note that the initial sending rate may be higher than the TCP throughput as the accelerator reports its own buffer availability in the AWnd sent back to the sender. Once the accelerator buffer becomes full, for example, around 25 seconds, the AWnd may realize flow control between the sender and the accelerator, of which the sender quickly reduces its sending rate to the TCP throughput. Compared to the case without the accelerator, the achievable throughput (payload only) may be increased from 1.34 Mbps to 6.71 Mbps.

[00103] Nevertheless, unlike wired networks where packet losses are primarily due to congestion, our measurements show that mobile data networks do exhibit random packet losses which may not be related to congestion. Thus, although the transmission scheduler may prevent congestion-induced packet losses, the occurrences of random packet loss may still degrade TCP's throughput performance. Loss recovery is discussed in detail in the next chapter.

[00104] Chapter 5: Local Retransmission

[00105] As shown in Chapter 4, the introduction of a virtual AWnd, which may be purposely made larger than the actual receiver AWnd, may lead to large packet loss bursts which may in turn blackout the throughput performance for hundreds of seconds (c.f. FIG. 9). The transmission scheduler in Chapter 4 may prevent loss induced by congestion but experiments according to various embodiments reveal that even a very low level of residual packet loss may still cause the blackout phenomena. Various embodiments are proposed in this Chapter to develop a new local retransmission algorithm, for example, to tackle this burst-loss problem so that the throughput of the TCP flow may be maintained in the event of packet loss.

[00106] 5.1 The blackout period

[00107] To illustrate the problem, an experiment is conducted using the same network topology as in FIG. 3. The accelerator may implement the transmission scheduler, for example, to prevent network congestion from occurring. The network emulator is configured, for example, to introduce random loss at about 0.01%. For ordinary TCP packet, loss at this low loss rate may not cause significant throughput degradation as such losses may be readily recovered via TCP's fast retransmit algorithm.

[00108] By contrast, as shown in FIG. 9 which shows the evolution of the CWnd versus time, the accelerated TCP flow may suffer from the same blackout period as reported in Chapter 4, even if there are only two packet losses before the blackout period begins. The blackout period may continue for over about 400 seconds while the sender repeatedly timeout and retransmit lost packets, for example, one at a time, to the receiver.

[00109] In various embodiments, measurements of production 3G HSPA networks reveal that the loss rate may not be zero even at low data rate. Table 1 summarizes the

loss rate for sending UDP data in about 1 KB datagram size to a mobile receiver at data rates ranging from about 1 to 7 Mbps. The 3G HSPA network and the 3G modem may have a maximum bandwidth of about 7.2 Mbps. It is clear from this result that even at the lowest data rate of about 1 Mbps, some packet losses may remain. Analysis of the traffic trace data suggests that these packet losses may be unlikely to be caused due to congestion.

[00110] To understand the cause for the blackout period, for example, the packet exchanges between the sender and the receiver in FIG. 1 may be traced. The accelerator in-between may simply forward packets between the sender and receiver, except for rewriting the AWnd field in ACKs sent by the receiver to a large value (e.g., about 10 MB). At the beginning of Phase I, the receiver buffer may be empty as it is assumed that it takes zero time to process the incoming packet and passes the data to the receiver application. Now suppose packet i is lost. Then, while packets $i+1$, $i+2$, etc., arrive at the receiver successfully, they may not be passed to the application as packet i has not been received. Instead, in one embodiment, the subsequent packets may be buffered up until the buffer is full when packet j ($j>i$) arrives as illustrated in Phase II of FIG. 1. In this case, the receiver may discard packet j and all subsequent packets as well.

[00111] Eventually, the sender may use up the CWnd, say, after packet $k-1$ (assuming $k>j+1$) is transmitted and then it may halt transmission. Meanwhile, when three duplicate ACKs for packet i are received, the sender may trigger fast retransmit by retransmitting packet i and decreasing the CWnd (e.g., to half in Reno). Suppose the retransmission is successful, then the ACK for packet j (i.e., the last buffered packet) may return to the sender at the end of Phase II. At the receiver side, all the buffered data may now be passed to the application, thus freeing up the receiver buffer.

[00112] In various embodiments, the blackout period may begin in Phase III. Here, the sender may not be able to transmit any packet as the CWnd has already been used up. Note that receiving the ACK for packet j may not relieve this condition as long as:

$$(k-1) - j + 1 \geq (k-i+1)/2$$

for the case of Reno. Similar conditions may be derived for other TCP variants such as CUBIC. In any case, this condition may be due to two factors: (a) the reduction of CWnd

in fast retransmit; and (b) the transmission of packets beyond the actual AWnd (i.e., packet j to $k-1$).

[00113] The sender in this case may wait until the RTO for packet j expires, at which time it may retransmit packet j and reset the CWnd to 1. By the time the ACK returns, the RTO for packet $j+1$ may be likely to have expired as well and so the sender may retransmit packet $j+1$, and the process may repeat. During Phase III, the sender may effectively transmit just one packet per RTT, resulting in the long stretch of very low throughput observed in FIG. 9. Eventually, all the lost packets up to packet $k-1$ may be successfully retransmitted and then the sender may begin slow-start again in Phase IV of FIG. 10. Note that this black-out behavior may occur because the accelerator sends packets beyond the receiver's AWnD. It is verified via experiments that the same behavior does not occur in ordinary TCP, even when a large burst of consecutive packets are dropped.

[00114] Therefore, while rewriting the AWnd eliminates the throughput limit constrained by AWnd, it may introduce a new problem in packet loss recovery. Various embodiments are proposed to tackle this problem, for example, by developing a new predictive retransmission algorithm in the next section.

Sending rate (Mbps)	1	2	3	4	5	6	7
Loss ratio	0.01	0.01	0.01	0.04	0.21	0.33	0.47

Table 1. Average loss ratio at various sending rate

[00115] 5.2 Proactive retransmission

[00116] Local retransmission has previously been proposed to handle random loss in wireless networks. The principle is to buffer unacknowledged packets in the intermediate gateway so that it can retransmit lost packets on behalf of the sender to suppress duplicate ACKs from triggering congestion control at the sender.

[00117] In various embodiments, local retransmission may be extended to tackle the blackout problem. Referring to FIG. 10, it is noted that when packet i is lost, all the out-of-window packets from packet j to $k-1$ may be discarded by the receiver. As the accelerator knows the actual AWnd reported by the receiver, it may compute j from

$$j = i + A W n d$$

[00118] Thus, instead of waiting for the sender to timeout and retransmit packet j , the accelerator may simply retransmit packets j to $k-1$ after retransmitting packet i . FIG. 11 illustrates this proactive retransmission algorithm according to various embodiments using the same scenario in FIG. 10. The pseudo-code for the algorithm according to various embodiments is listed in FIG. 12. Note that the retransmitted packets may also be subject to the transmission scheduling described in Section 4.2.3, for example, to prevent a large burst of retransmitted packets from congesting the network. In case the receiver supports TCP SACK, in various embodiments, the accelerator may also scan through the SACK list and retransmit the missing packets within the $AWnd$.

[00119] In various embodiments, the accelerator may suppress the duplicate ACK (i.e., not forwarding it to the sender) if it can retransmit the packet. This may prevent the sender from triggering its congestion control algorithm. On the other hand, if the packet is lost before it reaches the accelerator, then the duplicate ACK may be forwarded to the sender as the accelerator cannot retransmit a packet never received. This may also preserve TCP's congestion control of the path between the sender and the accelerator.

[00120] 5.3 Performance

[00121] The experiment may be repeated using the setup in FIG. 3 with local retransmission implemented at the accelerator. FIG. 13 shows the evolution of $CWnd$, throughput, and the sender's sending rate versus time according to various embodiments. Firstly, it is observed that the two packet losses within the first 5 seconds do not cause the sender to reduce its $CWnd$. This is because the loss events may be successfully suppressed by the accelerator.

[00122] However, the third packet loss may result in sender timeout, despite that the lost packet may be successfully retransmitted by the accelerator. This is because the extra delay incurred by local retransmission may delay the ACK so much that it may exceed the RTO of the sender. This shows that local retransmission alone may not be able to suppress all loss events. Various embodiments are proposed to develop a novel RTT modulation technique in the next section, for example, to tackle this problem.

[00123] Chapter 6: Loss Event Suppression

[00124] A fundamental limitation of local retransmission is that the lost packet may not be retransmitted and the corresponding ACK received in time to prevent the sender from timeout. Once timeout occurs, the TCP flow's throughput will decrease significantly as the CWnd is reset to one. One solution to this problem is to split the TCP flow into two, one between the sender and the accelerator and the other between the accelerator and the receiver. This split-TCP approach decouples loss recovery between the sender and the receiver by generating ACKs independently of the receiver. Specifically, a split-TCP gateway will return ACKs to the sender as if it is the receiver. The received packets are then forwarded to the receiver in a separate TCP connection. Thus, loss events between the gateway and the receiver are handled completely by the gateway, without involving the sender at all. This prevents the sender from triggering its congestion control algorithm.

[00125] However the split-TCP approach suffers from a tradeoff – it breaks the reliable service guarantee provided by TCP. In particular, it is possible for a packet to be acknowledged by the gateway but ultimately fail to be delivered to the receiver (e.g., the network link goes down before the gateway can forward the packet to the receiver). In this case, the sender and the receiver may be in an inconsistent state, i.e., the sender assumes successful delivery of a packet which is in fact not received by the receiver.

[00126] To avoid this problem, a novel RTT modulation mechanism according to various embodiments are proposed to enable the suppression of loss events without the need for splitting the TCP flow.

[00127] 6.1 RTT modulation

[00128] The challenge in retransmitting lost packet locally by the accelerator is the time needed for the retransmissions. In particular, if the retransmission cannot be done so that the ACK can reach the sender before it times out, congestion control may kick in to reduce the sender's transmission rate significantly (for example, by reducing CWnd to 1 and then restart from slow-start again). Specifically, TCP may set its timeout threshold, denoted by RTO, according to:

$$RTO = srtt \div 4 \times RTT_{dev} \quad (9)$$

where $srtt$ is the smoothed RTT and $rttdev$ is the smoothed mean deviation of RTT. Thus, to avoid timeout, the ACK may arrive at the sender no later than the average RTT plus four times the mean RTT deviation.

[00129] Considering the scenario in FIG. 14, let RTT_{SR} , RTT_{SA} , RTT_{AR} be the RTT between sender and receiver, sender and accelerator and accelerator and receiver, respectively. Normally RTT_{SR} is the sum of RTT_{SA} and RTT_{AR} . However, once local retransmission takes place, measured RTT of the lost packet may increase to $RTT_{SA}+2RTT_{AR}$. Depending on the extent of RTT variation of the mobile data network, the extra delay RTT_{AR} may be larger than four times the mean deviation of the RTT as measured by the sender.

[00130] In various embodiments, measurement of a production 3G HSPA network recorded a mean RTT deviation at around 7 ms while the mean RTT_{AR} is around 100 ms. Thus, in this case even the packet can be successfully retransmitted in one attempt, the extra delay incurred (i.e., 100 ms) may still trigger timeout at the sender, leading to a significant drop in the sender's CW_{nd} .

[00131] Now, the time needed for retransmission may depend on a number of factors, including the RTT between the accelerator and the receiver, transmission time, and accelerator processing time. The first two of these factors may be network dependent and thus may not be reduced further by the transport layer. On the other hand, it is noted that the timeout threshold at the sender may be continuously updated through measurement of the average RTT and mean RTT deviation between the sender and the receiver whenever an ACK comes back.

[00132] In (9), it is noted that $srtt$ may be computed from measured RTT, which may be in turn computed from the difference between the ACK packet arrival time and the transmission time of the original data packet. While the way the sender computes the $srtt$ may not be changed, the accelerator may change the RTT as measured by the sender, simply by delaying the forwarding of ACK packets received from the receiver – RTT modulation.

[00133] The next question may be the amount of delay to add. Intuitively, a longer delay may allow more time to carry out local retransmission. However, if the delay is too long, then it may exceed the RTO threshold at the beginning of the connection, resulting

in false triggering of timeout. Moreover, a long RTT may also lower the rate at which the sender ramps up the CWnd, thus degrading throughput at the beginning.

[00134] To tackle this dilemma, various embodiments are proposed to increase the forwarding delay in an incrementally manner. Specifically, the accelerator may continuously measure the RTT deviation between the accelerator and the receiver by

$$rtt_{dev_i} = |rtt_i - srtt| \quad (10)$$

and then compute the smoothed mean RTT deviation from

$$rtt_{dev} = \frac{3}{4} rtt_{dev} + \frac{1}{4} rtt_{dev_i} \quad (11)$$

[00135] Although the measurement does not include the delay and delay variations between the sender and the accelerator, the impact is generally insignificant as (a) the mobile link typically exhibits larger RTT and RTT variations than the wired link; and (b) RTT variations of the wired link is generally independent of the RTT variations of the mobile link. Thus, in general, the mean RTT deviation measured by the sender may not be smaller than the mean RTT deviation of the mobile link.

[00136] Define L to be the number of retransmission attempts to accommodate. Then, an ACK-forwarding delay, denoted by D, may be added such that

$$D \geq L \times rtt \quad (12)$$

where rtt is the smoothed RTT between the accelerator and the receiver.

[00137] Denote the initial ACK-forwarding delay as D₀=0. Then whenever an ACK arrives at the accelerator, it will increase the ACK-forwarding delay by

$$D_{i+1} = D_i + \beta \times rtt_{dev} \quad (13)$$

where rttdev is the smoothed mean RTT deviation between the accelerator and the receiver. The parameter β controls how fast the ACK-forwarding delay may be increased. An obvious limit is $\beta < 4$ as the sender RTO may expire if the ACK arrives 4 times the smoothed mean RTT deviation later than the mean RTT.

[00138] 6.2 Performance

[00139] To evaluate the performance of RTT modulation, in various embodiments, the experiment may be repeated using the setup in FIG. 3 by implementing the transmission scheduler, proactive retransmission, and RTT modulation algorithms in the accelerator. In one embodiment, it may be set that $\beta = 2$ in (13). FIG. 15 shows the

evolution of modulated RTO (RTOM), RTO and RTT versus time according to various embodiments. Random packet loss occurred at 8.21 s. FIG. 16 shows an enlarged view of FIG. 15 around the time the loss event occurred according to various embodiments. As the lost packet was retransmitted by the accelerator, the RTT of the locally retransmitted packet is increased beyond the original RTO. Thus, without RTT modulation, this may trigger sender timeout. By contrast, modulated RTT may allow the extra margin for the local retransmission to complete, and in this case the increased RTT may stay within the modulated RTT and thus the loss event may be successfully suppressed.

[00140] Chapter 7: Fairness

[00141] Congestion control between the accelerator and the receiver may no longer follow the conventional TCP AIMD algorithm. In a mobile data network, although network channel resources are allocated in a per-device basis, TCP flows destined to the same receiver may still compete against each other for bandwidth. In various embodiments, the mobile data network may also be shared by multiple hosts using an NAT. In this case, TCP flows destined to all hosts connected via the same mobile data link may compete against each other for bandwidth. In this chapter, the bandwidth sharing behavior of conventional TCP and that of the accelerator according to various embodiments are discussed in detail.

[00142] 7.1 Packet forwarding

[00143] In some embodiment, the accelerator may allocate channel resources in a round robin manner as illustrated in FIG. 17 for TCP flows destined towards the same receiver. Specifically, packets may be organized in per-flow queues awaiting forwarding to the receiver. Each packet may have a scheduled forwarding time as determined by the transmission scheduler (c.f. Section 4.2.3). The accelerator may iterate through all queues in a round-robin manner. Empty queues may be skipped. Otherwise, the scheduled forwarding time of the queue's head-of-line packet may be compared against the current time, and the packet may be forwarded if the scheduled forwarding time is up. Otherwise, the queue is skipped.

[00144] To evaluate this component, in various embodiments, two connections transferring about 16 MByte of data may be used while one of the connections may start 5 seconds earlier than another one such that the earlier link can achieve its maximum throughput and let the latter one compete with it. In various embodiments, for the case of not using the accelerator, the receiver buffer size may be manually set to be about 128 KB (which is larger than the BDP of the emulated link at about 108 KB), and random loss may be set to be zero to ensure that the two ordinary TCP connections (using Linux's default CUBIC congestion module) can fully utilize the link capacity.

[00145] FIG. 18 and FIG. 19 show the 1-s averaged throughput of the two TCP flows for the case without and with accelerator, respectively, according to various embodiment. In one embodiment, the throughput may be measured at the receiver. It is observed that normal TCP flows without the accelerator exhibit more throughput fluctuations than the case with the accelerator. The throughput fluctuations may be due to the bandwidth probing actions of the two TCP flows. By contrast, the accelerator's round-robin scheduler may allocate near equal bandwidth to the two competing TCP flows with negligible fluctuations.

[00146] The differences may widen even further if the throughput is averaged over a shorter timescale, e.g., about 100 ms in FIG. 20 and FIG. 21. In terms of fairness, in various embodiments, Jain's fairness index may be employed to quantify the comparisons. The fairness index may range from $1/n$ to 1 where n is the number of competing flows. Larger index value may represent, for example, better fairness among the competing TCP flows. Table 4 lists the fairness indices for three timescales, i.e., about 1000 ms, 100 ms, and 10 ms. It is noted that the accelerator may maintain a high fairness index even for timescale as short as about 10 ms, at which point normal TCP performs poorly. These results demonstrate that although the accelerator does not implement TCP's AIMD congestion control algorithm, it nevertheless may achieve even better fairness in bandwidth sharing.

Scales	1000 ms	100 ms	10 ms
Normal TCP	0.95	0.76	0.82
Using accelerator	0.95	0.98	0.99

Table 2 Jian's fairness index with different scale

[00147] 7.2 Non-uniform bandwidth allocation

[00148] In addition to round-robin scheduling, in various embodiments, the accelerator may open the door to implement more sophisticated bandwidth allocation and channel scheduling algorithms. As an example, in one embodiment, a generalized queuing discipline called weight fair queuing (WFQ) may be implemented. Each TCP flow may be assigned a weight, denoted by w_i for flow i . In each round of packet forwarding, up to w_i packets may be forwarded for flow i . This may result in unequal throughput with the flows throughput approximating the ratios of the weights. FIG. 22 shows the throughput of 3 TCP flows with weights $w_0 = 1$, $w_1 = 2$, and $w_2 = 3$, respectively, according to various embodiments. Considering the stationary period from time about 20 to 40 seconds, the resultant throughput ratios may be 1:2.0001:3.0003, which closely approximate the assigned weights.

[00149] WFQ may be applied in many ways, such as to provide different bandwidth to different applications (e.g., more bandwidth for web browsing versus P2P), or to adapt the bandwidth allocation based on application demands (e.g., video streaming versus file download), and so on.

[00150] Chapter 8: Experiments

[00151] In various embodiments, extensive measurements and experiments may be conducted with respect to a production 3G HSPA network, for example, to evaluate the characteristics of mobile data networks, to study the performance of TCP, and to evaluate the performance of the proposed accelerator.

[00152] 8.1 Experiment setup

[00153] FIG. 23 illustrates the experiment setup according to various embodiments. In one embodiment, the sender and the accelerator both may run on Linux with kernel 2.6 (with CUBIC as the TCP congestion control module). They may connect to the Internet via high-speed wired network links running at about 100 Mbps. The receiver host may run Windows XP SP2 and be connected to the mobile data network via a USB 3G HSPA modem. All experiments may be conducted with the receiving host in a stationary position.

[00154] 8.2 Packet loss

[00155] In various embodiments, the packet loss rate of the mobile link may be studied by sending UDP datagrams directly from the sender to the receiver, bypassing the accelerator, at various fixed rates ranging from about 3 to 6.5 Mbps. FIG. 24 shows the packet loss ratio with 95% confidence interval versus various UDP data rates according to various embodiments. The measurement shows that the mobile link may have an upper throughput limit at approximately 5.40 Mbps. Nevertheless, as the results in FIG. 24 show, the packet loss ratio may be significant even at data rates lower than the mobile link's throughput limit. Moreover, the results consistently show that the loss rates may be data rate dependent – higher data rate generally results in higher packet loss ratio.

[00156] A second observation is that packet loss may remain even if the data rate is low. For example, in one embodiment, at a data rate of about 3 Mbps, a loss rate of approximately 0.6% may be consistently measured. These losses may be due to radio transmission errors which may not be recovered by the link layer's retransmission mechanism such as HARQ.

[00157] It is also worth noting that the extent of packet losses may depend on many factors, such as the location of the mobile device, mobility of the mobile device, amount of radio interference, etc. In various embodiments, similar measurements may also be conducted with respect to other mobile data networks and different levels (higher) of packet losses may be found. Nevertheless, the previous two observations according to various embodiments are consistent even across different mobile operators.

[00158] 8.3 Unaccelerated TCP throughput

[00159] The throughput performance of normal TCP over the mobile data network may be evaluated. In addition to using the default receiver window size, which is about 17 KB for Microsoft® Windows® XP®, a custom application according to various embodiments may explicitly increase the receiver window size by increasing the socket buffer size via the sockets API.

[00160] FIG. 25 shows the achievable TCP throughput versus receiver buffer size ranging from about 17 KB to about 128 KB according to various embodiments. The BDP of the path may be approximately 90 KB and thus the 128 KB and the 192 KB settings already exceeded the BDP. Not surprisingly, the throughput may increase with

larger receiver buffer size. However, it is noted that even with a receiver buffer size of about 192 KB, the achievable throughput may be still lower than the throughput limit of about 5.40 Mbps as measured using UDP in Section 8.2.

[00161] 8.4 Accelerated TCP throughput

[00162] Next, in various embodiments, the throughput test may be repeated by introducing the accelerator between the sender and the receiver. The receiver in this case may employ the default receiver buffer size of about 17 KB. The test may be set to transfer a file of about 32 MB from the sender to the receiver.

[00163] Table 3 summarizes the overall throughput and steady state throughput for both accelerated and unaccelerated cases according to various embodiments. The former may include the ramp-up period of TCP while the latter may not. In either case, the accelerator may increase the achievable throughput by approximately 2.5 times. Moreover, the accelerated TCP may achieve a throughput (about 4.03 Mbps) higher than the case for normal TCP with about 192 KB receiver buffer size (3.80 Mbps). This is because the accelerator may not only eliminate the AWnd-induced bottleneck, but also suppress packet loss events, for example, to keep the sender CWnd at a high level.

	Overall Throughput (bps)	Steady State Throughput (bps)
Without Accelerator	1617918.03	1696274.89
With Accelerator	4033170.61	4255799.80
Ratio (With / Without)	2.49	2.51

Table 3 Improvement in both overall throughput and steady state throughput

[00164] 8.5 Fairness

[00165] In various embodiments, for example, to evaluate the fairness in bandwidth sharing across competing TCP flows, two concurrent TCP connections may be initiated between the sender and the receiver, both transferring data as fast as TCP allows. For the unaccelerated case, the receiver window size may be explicitly set to be about 128 KB while the default window size may be set to be about 17 KB for the accelerated case.

[00166] In various embodiments, the throughput of the two TCP flows at the receiver may be captured and measured, and then the Jain's fairness index may be computed using throughput data averaged over about 100 ms intervals. The resultant fairness indices may be about 0.74 and 0.99 for the unaccelerated and accelerated cases, respectively. Consistent with the results in Section 7.1, the accelerator may achieve much better fairness in sharing bandwidth across the competing TCP flows.

[00167] 8.6 Mobile handset performance

[00168] In various embodiments, the experiment setup may be changed to replace the PC-based receiver host by a mobile handset. For example, in one embodiment, three mobile handsets may be tested: iPhone 3G, iPhone 3GS, and Nexus One. All three mobile handsets may have built-in support for 3G HSPA. Throughput may be measured by downloading a 4 MB image file from a web server running in the sender host. Table 4 shows a summary of measured throughput according to various embodiments.

	Without accelerator	With accelerator
iPhone 3G throughput (Mbps)	2.50	2.6
iPhone 3GS throughput (Mbps)	3.56	3.69
Nexus One throughput (Mbps)	2.93	4.02

Table 4 Throughput obtained with handsets with and without the accelerator

[00169] From the packet trace data, it is observed that both iPhone[®] 3G and iPhone[®] 3GS may have a fixed receiver buffer size of about 128 KB while Nexus[®] One may implement adaptive buffer sizing, with a dynamically variable receiver buffer size up to about 81 KB.

[00170] As shown in Table 4, in various embodiments, for iPhone[®] 3G and iPhone[®] 3GS, the accelerator may not offer significant improvement in TCP throughput. Analysis of the traffic trace data reveals that the RTT measured by the sender may be about 400 ms. However, if the mobile handset is replaced by a PC-based receiver, the measured RTT may be only about 120 ms at the same data rate. As the same mobile data network is used in both cases, the differences in RTT may be due to processing delays incurred by the mobile handsets. In other words, in the cases of iPhone[®] 3G and iPhone[®]

3GS, the achievable throughput may be in fact limited by the processing capacity of the receiver. It is worth nothing that this capacity limit may be incorporated into the accelerator's transmission scheduler and thus even without TCP's AIMD congestion control, the accelerator may still correctly estimate the receiver's processing capacity and achieved its throughput limit.

[00171] For the Nexus One, in one embodiment, the accelerator may increase the achievable throughput from about 2.93 Mbps to about 4.02 Mbps. This suggests that the Nexus One may have a higher processing capacity which allows, for example, the accelerated TCP flow to reach a higher throughput.

[00172] Chapter 9: Machine-readable medium

[00173] The system 200 and apparatuses 202, 204, 206 in FIG. 26 may be implemented in a machine-accessible and readable medium that is operational over one or more networks 208. The networks 208 may be wired, wireless, or a combination of wired and wireless. Also, at least one of the networks 208 may be a satellite-based communication link, such as the WINDS (Wideband InterNetworking engineering test and Demonstration Satellite) communication link or any other commercial satellite communication links. The system 200 and apparatuses 202, 204, 206 may be used to implement, among other things, the processing associated with the methods 100 of FIG. 5. Modules may comprise hardware, software, and firmware, or any combination of these. Additional embodiments may be realized.

[00174] For example, FIG. 26 is a block diagram of an article 300 of manufacture, including a specific machine 302, according to various embodiments of the invention. Upon reading and comprehending the content of this disclosure, one of ordinary skill in the art will understand the manner in which a software program can be launched from a computer-readable medium in a computer-based system to execute the functions defined in the software program.

[00175] One of ordinary skill in the art will further understand the various programming languages that may be employed to create one or more software programs designed to implement and perform the methods disclosed herein. The programs may be structured in an object-oriented format using an object-oriented language such as Java or

C++. Alternatively, the programs can be structured in a procedure-oriented format using a procedural language, such as assembly or C. The software components may communicate using any of a number of mechanisms well known to those of ordinary skill in the art, such as application program interfaces or interprocess communication techniques, including remote procedure calls. The teachings of various embodiments are not limited to any particular programming language or environment. Thus, other embodiments may be realized.

[00176] For example, an article 300 of manufacture, such as a computer, a memory system, a magnetic or optical disk, some other storage device, and/or any type of electronic device or system may include one or more processors 304 coupled to a machine-readable medium 308 such as a memory (e.g., removable storage media, as well as any memory including an electrical, optical, or electromagnetic conductor) having instructions 312 stored thereon (e.g., computer program instructions), which when executed by the one or more processors 304 result in the machine 302 performing any of the actions described with respect to the methods above.

[00177] The machine 302 may take the form of a specific computer system having a processor 304 coupled to a number of components directly, and/or using a bus 316. Thus, the machine 302 may be similar to or identical to the apparatuses 202, 204, 206 or system 200 shown in FIG. 26.

[00178] Turning now to FIG. 26, it can be seen that the components of the machine 302 may include main memory 320, static or non-volatile memory 324, and mass storage 306. Other components coupled to the processor 304 may include an input device 332, such as a keyboard, or a cursor control device 336, such as a mouse. An output device 328, such as a video display, may be located apart from the machine 302 (as shown), or made as an integral part of the machine 302.

[00179] A network interface device 340 to couple the processor 304 and other components to a network 344 may also be coupled to the bus 316. The instructions 312 may be transmitted or received over the network 344 via the network interface device 340 utilizing any one of a number of well-known transfer protocols (e.g., HyperText Transfer Protocol and/or Transmission Control Protocol). Any of these elements coupled to the

bus 316 may be absent, present singly, or present in plural numbers, depending on the specific embodiment to be realized.

[00180] The processor 304, the memories 320, 324, and the storage device 306 may each include instructions 312 which, when executed, cause the machine 302 to perform any one or more of the methods described herein. In some embodiments, the machine 302 operates as a standalone device or may be connected (e.g., networked) to other machines. In a networked environment, the machine 302 may operate in the capacity of a server or a client machine in server-client network environment, or as a peer machine in a peer-to-peer (or distributed) network environment.

[00181] The machine 302 may comprise a personal computer (PC), a tablet PC, a set-top box (STB), a PDA, a cellular telephone, a web appliance, a network router, switch or bridge, server, client, or any specific machine capable of executing a set of instructions (sequential or otherwise) that direct actions to be taken by that machine to implement the methods and functions described herein. Further, while only a single machine 302 is illustrated, the term “machine” shall also be taken to include any collection of machines that individually or jointly execute a set (or multiple sets) of instructions to perform any one or more of the methodologies discussed herein.

[00182] While the machine-readable medium 308 is shown as a single medium, the term “machine-readable medium” should be taken to include a single medium or multiple media (e.g., a centralized or distributed database, and/or associated caches and servers, and or a variety of storage media, such as the registers of the processor 304, memories 320, 324, and the storage device 306 that store the one or more sets of instructions 312). The term “machine-readable medium” shall also be taken to include any medium that is capable of storing, encoding or carrying a set of instructions for execution by the machine and that cause the machine 302 to perform any one or more of the methodologies of the present invention, or that is capable of storing, encoding or carrying data structures utilized by or associated with such a set of instructions. The terms “machine-readable medium” or “computer-readable medium” shall accordingly be taken to include tangible media, such as solid-state memories and optical and magnetic media.

[00183] Various embodiments may be implemented as a stand-alone application (e.g., without any network capabilities), a client-server application or a peer-to-peer (or

distributed) application. Embodiments may also, for example, be deployed by Software-as-a-Service (SaaS), an Application Service Provider (ASP), or utility computing providers, in addition to being sold or licensed via traditional channels.

[00184] Chapter 10: Conclusion

[00185] Various embodiments of the invention address the performance problem of running TCP over mobile data networks by introducing an accelerator between the sender and the receiver. The accelerator may implement, for example, three acceleration algorithms: (a) opportunistic transmission to overcome the AWnd-induced throughput bottleneck; (b) local retransmission to prevent throughput blackout in the event of packet loss; and (c) loss event suppression to keep the sender CWnd at a high level.

[00186] Experimental results conducted in production 3G HSPA networks show that the accelerator can increase the throughput performance of TCP by up to about 2.5 times of the unaccelerated TCP. Moreover, the accelerator can achieve better fairness among competing TCP flows and can also be equipped with dynamic resource allocation algorithms to offer more sophisticated traffic control for the TCP flows.

[00187] Last but not least, the proposed accelerator according to various embodiments does not require modification to related applications, TCP implementation at the hosts, or operating system; and thus can be readily deployed in current and future mobile data networks.

[00188] The accompanying drawings that form a part hereof show, by way of illustration and not of limitation, specific embodiments in which the subject matter may be practiced. The embodiments illustrated are described in sufficient detail to enable those skilled in the art to practice the teachings disclosed herein. Other embodiments may be utilized and derived therefrom, such that structural and logical substitutions and changes may be made without departing from the scope of this disclosure. This Detailed Description, therefore, is not to be taken in a limiting sense, and the scope of various embodiments is defined only by the appended claims and the full range of equivalents to which such claims are entitled.

[00189] Such embodiments of the inventive subject matter may be referred to herein individually or collectively by the term “invention” merely for convenience and

without intending to voluntarily limit the scope of this application to any single invention or inventive concept, if more than one is in fact disclosed. Thus, although specific embodiments have been illustrated and described herein, any arrangement calculated to achieve the same purpose may be substituted for the specific embodiments shown. This disclosure is intended to cover any and all adaptations or variations of various embodiments. Combinations of the above embodiments and other embodiments not specifically described herein will be apparent to those of skill in the art upon reviewing the above description.

[00190] The Abstract of the Disclosure is provided to comply with 37 C.F.R. §1.72(b) requiring an abstract that will allow the reader to quickly ascertain the nature of the technical disclosure. It is submitted with the understanding that it will not be used to interpret or limit the scope or meaning of the claims. In the foregoing Detailed Description, various features are grouped together in a single embodiment for the purpose of streamlining the disclosure. This method of disclosure is not to be interpreted to require more features than are expressly recited in each claim. Rather, inventive subject matter may be found in less than all features of a single disclosed embodiment. Thus the following claims are hereby incorporated into the Detailed Description, with each claim standing on its own as a separate embodiment.

Claims

What is claimed is:

1. A computer-implemented method, comprising:
receiving at least one data packet from a transmitter connected to a first network;
scheduling a packet forwarding time for the at least one data packet; and
forwarding the at least one data packet to a receiver connected to a second network
based on the packet forwarding time such that transmission of the at least one data packet
prevents a buffer overflow at the receiver.
2. The method of claim 1, wherein the scheduling of the packet forwarding time
includes computing a packet forwarding delay for the at least one data packet.
3. The method of claim 2, wherein the computing of the packet forwarding delay
includes computing a packet reception rate by the receiver.
4. The method of claim 3, wherein the computing of the packet reception rate includes
calculating a difference in acknowledged sequence numbers associated with the at least
one data packet.
5. The method of claim 1, wherein the forwarding of the at least one data packet
includes:
based on detecting a lost packet, determining out-of-window packets to be
discarded by the receiver; and
retransmitting the out-of-window packets to the receiver before the transmitter
times out and retransmits one of the out-of-window packets.
6. The method of claim 5, wherein the retransmitting of the out-of-window packets
includes suppressing forwarding a duplicate acknowledgement (ACK) received from the
receiver to the transmitter.

7. The method of claim 5, wherein the retransmitting of the out-of-window packets includes delaying forwarding an acknowledgement (ACK) received from the receiver to the transmitter.
8. The method of claim 7, wherein the delaying of the forwarding of the ACK includes increasing an amount of delayed time according to a number of retransmission attempts.
9. The method of claim 1, wherein the scheduling of the packet forwarding time includes assigning a weight to each data flow, the weight indicating a number of packets to be forwarded for each data flow.
10. The method of claim 1, wherein the scheduling of the packet forwarding time includes allocating channel resources in a round robin manner to forward the at least one data packet.
11. The method of claim 1, wherein the at least one data packet comprises a Transmission Control Protocol (TCP) package.
12. The method of claim 1, further comprising:
 - determining the number of the at least one data packet to be forwarded to the receiver; and
 - informing the transmitter the number of the at least one data packet.
13. The method of claim 12, wherein the determining the number of the at least one data packet to be forwarded to the receiver includes: computing the amount of buffers free-up at the receiver during a round-trip-delay.
14. The method of claim 12, wherein the determining the number of the at least one data packet to be forwarded to the receiver includes: computing the amount of data packets received at the receiver during a round-trip-delay.

15. The method of claim 12, wherein the informing the transmitter the number of the at least one data packet includes: dynamically adjusting a size of a buffer advertized by the receiver based on the number of the at least one data packet to be forwarded to the receiver determined by the transmitter.

16. An apparatus, comprising:

a receiving module to receive at least one data packet from a transmitter connected to a first network;

a scheduling module to schedule a packet forwarding time for the at least one data packet; and

a forwarding module to forward the at least one data packet to a receiver connected to a second network based on the packet forwarding time such that transmission of the at least one data packet prevents a buffer overflow at the receiver.

17. The apparatus of claim 16, wherein the scheduling module is configured to compute a packet forwarding delay for the at least one data packet to schedule the packet forwarding time.

18. The apparatus of claim 17, wherein the scheduling module is configured to compute a packet reception rate by the receiver to compute the packet forwarding delay.

19. The apparatus of claim 18, wherein the scheduling module is configured to calculate a difference in acknowledged sequence numbers associated with the at least one data packet to compute the packet reception rate.

20. The apparatus of claim 16, wherein the forwarding module is configured to:

determine out-of-window packets to be discarded by the receiver based on detecting at least one lost packet; and

retransmit the out-of-window packets to the receiver before the transmitter times out and retransmits one of the out-of-window packets.

21. The apparatus of claim 20, wherein the forwarding module is configured to delay forwarding an acknowledgement (ACK) received from the receiver to the transmitter.
22. The apparatus of claim 21, wherein the forwarding module is configured to increase an amount of delayed time according to a number of retransmission attempts.
23. The apparatus of 16, wherein the forwarding module is configured to enable suppression of lost events without the need for splitting data flow.
24. The apparatus of claim 16, wherein the scheduling module is configured to assign a weight to each data flow, the weight indicating a number of packets to be forwarded for each data flow.
25. The apparatus of claim 16, wherein the receiver comprises a mobile handset.
26. The apparatus of claim 16, wherein the scheduling module is configured to allocate channel resources in a round robin manner to forward the at least one data packet.
27. The apparatus of claim 16, wherein the at least one data packet comprises a Transmission Control Protocol (TCP) package.
28. The apparatus of claim 16, wherein the scheduling module is configured to determine the number of the at least one data packet to be forwarded to the receiver; and the forwarding module is configured to inform the transmitter the number of the at least one data packet.
29. The apparatus of claim 28, wherein the scheduling module is configured to compute the amount of buffers free-up at the receiver during a round-trip-delay.
30. The apparatus of claim 28, wherein the scheduling module is configured to compute the amount of data packets received at the receiver during a round-trip-delay.

31. The apparatus of claim 28, wherein the scheduling module is configured to dynamically adjust a size of a buffer advertized by the receiver based on the number of at least one data packet to be forwarded to the receiver determined by the transmitter.

32. A computer-readable non-transitory storage medium storing instructions that, when executed by a processor, cause the processor to perform a method comprising:

- receiving at least one data packet from a transmitter connected to a first network;
- scheduling a packet forwarding time for the at least one data packet; and

- forwarding the at least one data packet to a receiver connected to a second network based on the packet forwarding time such that transmission of the at least one data packet prevents a buffer overflow at the receiver.

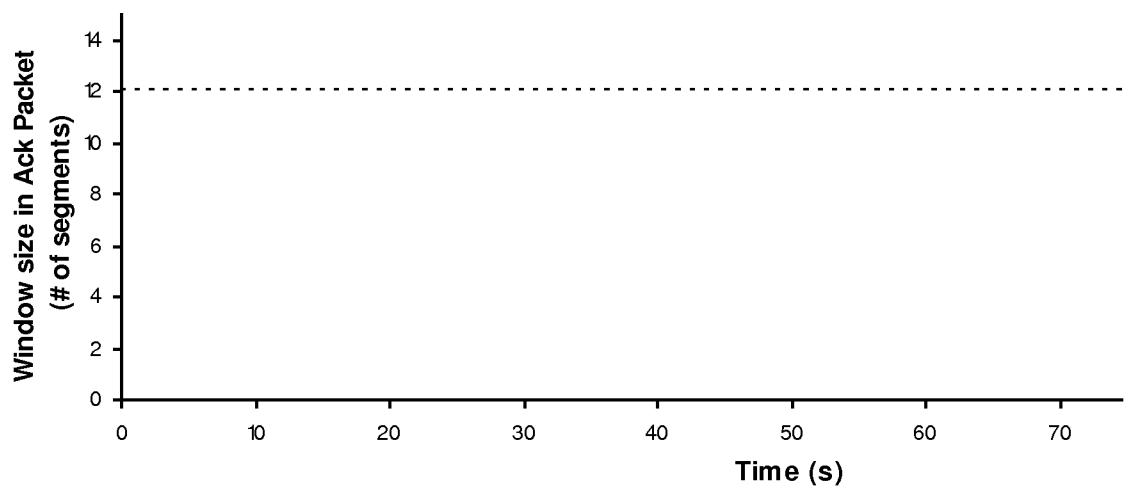


FIG 1

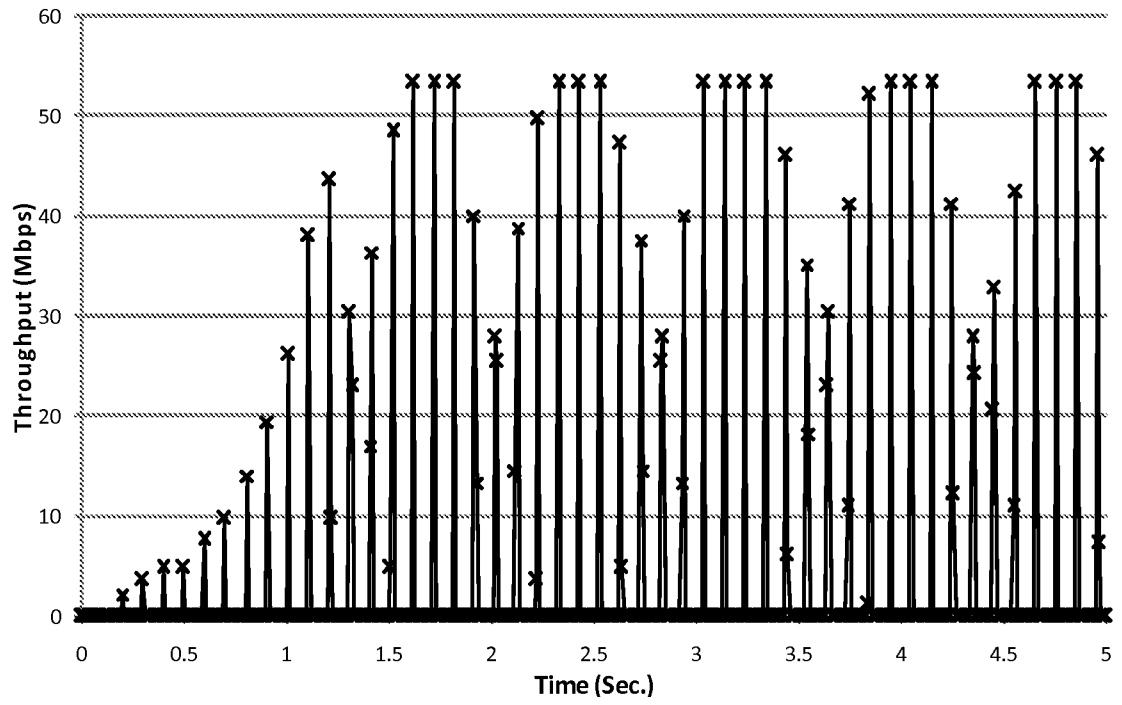


FIG 2

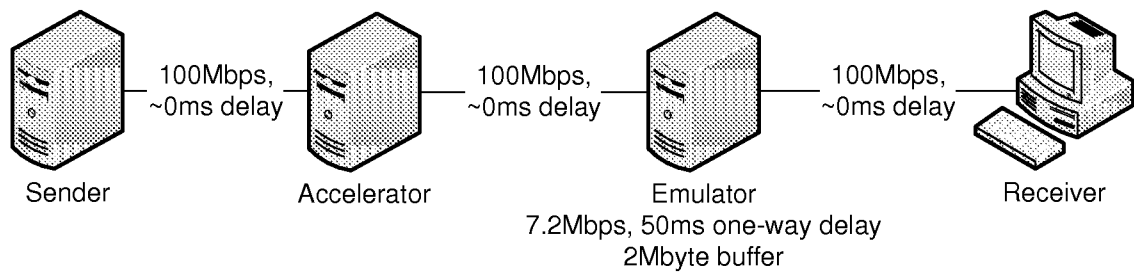


FIG 3

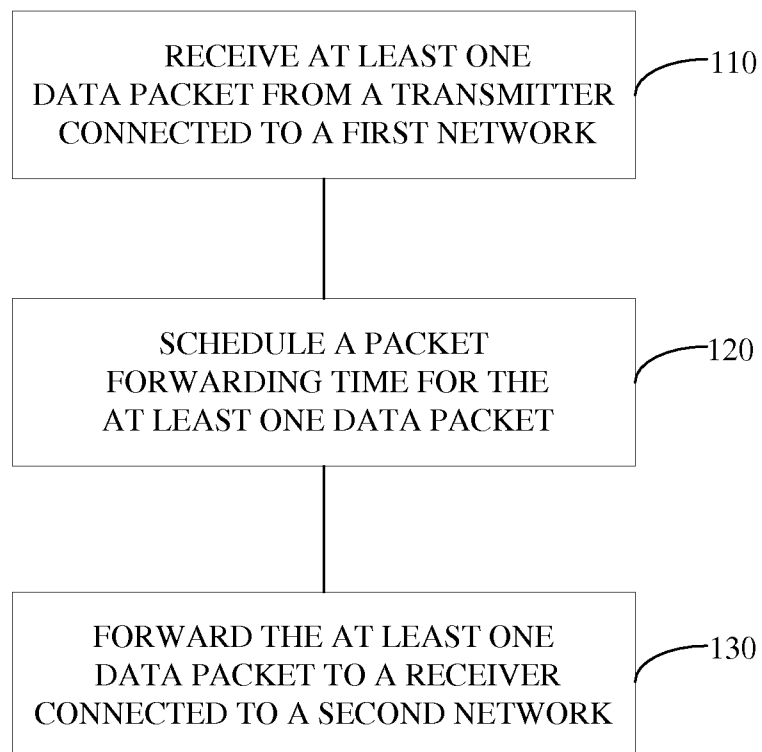
100

FIG 4

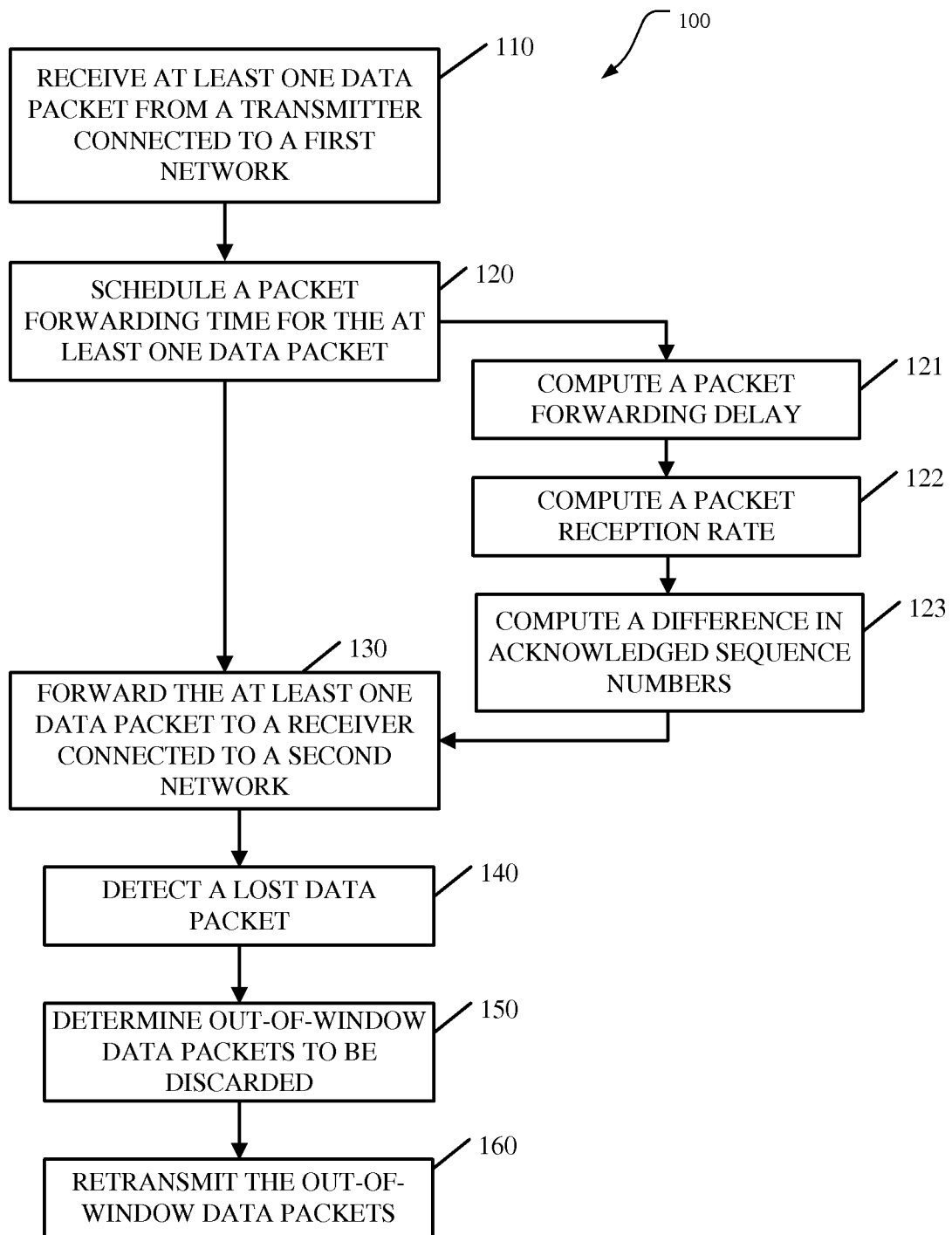


FIG 5

6/29

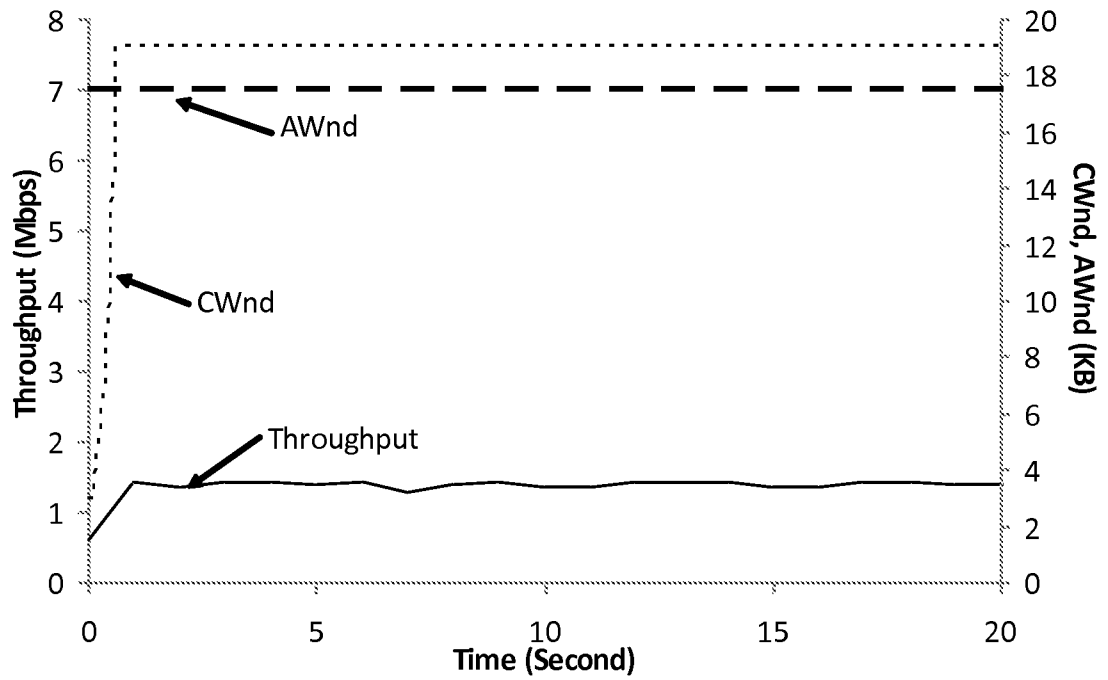


FIG 6A

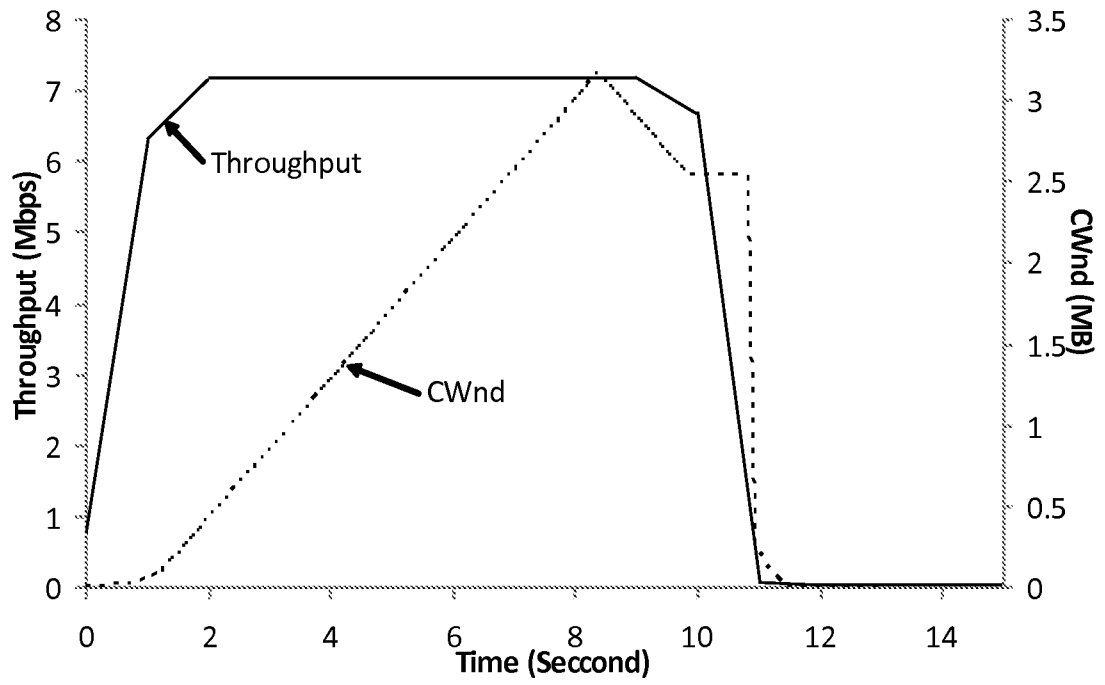


FIG 6B

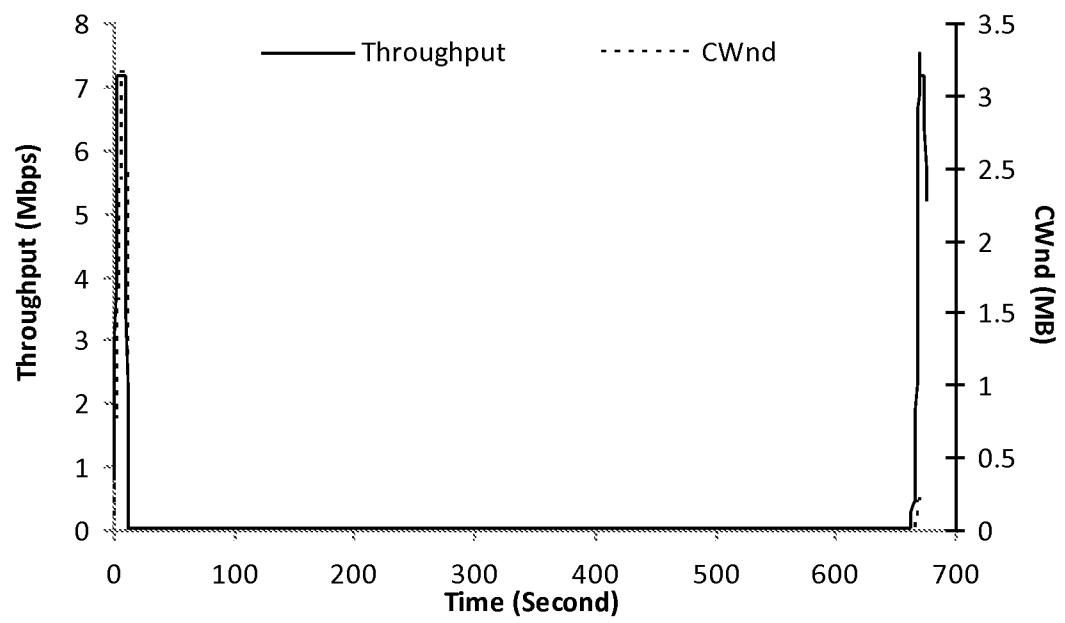


FIG 6C

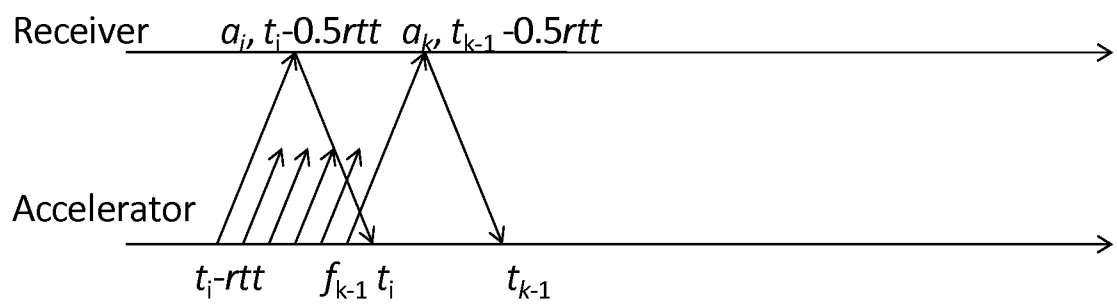


FIG 7

10/29

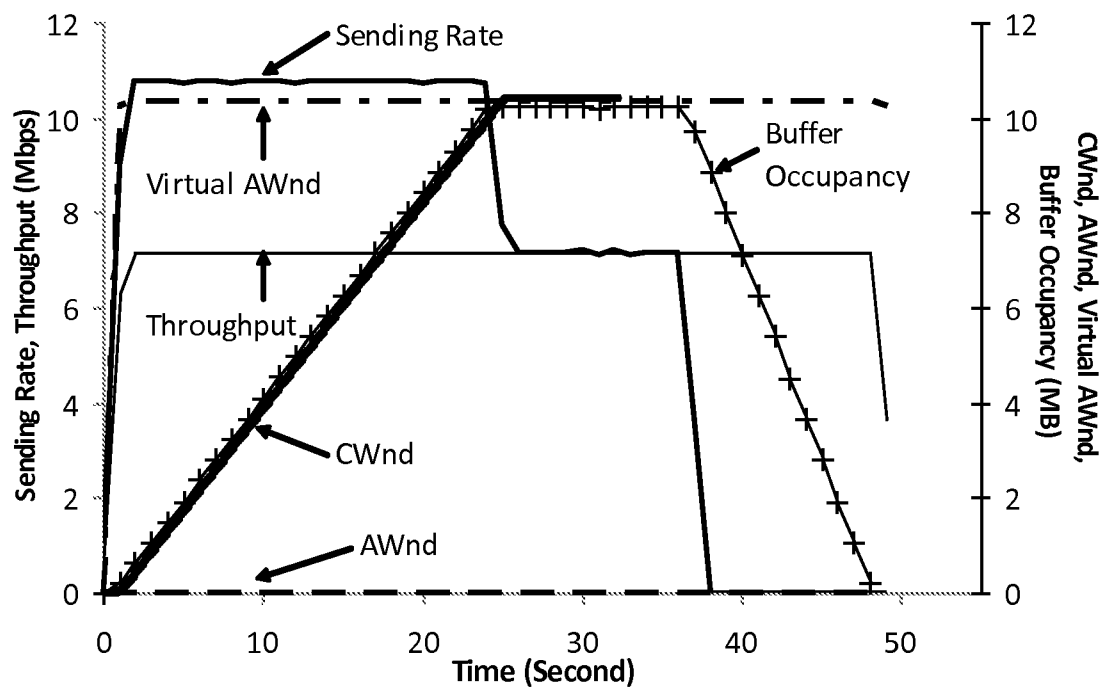


FIG 8

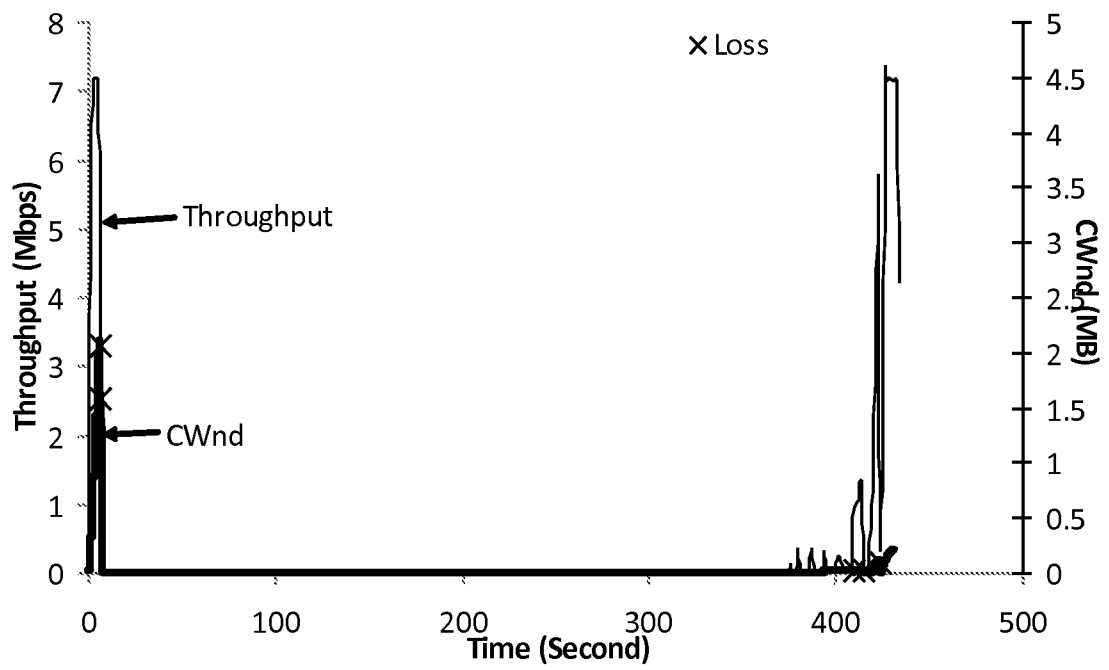


FIG 9

12/29

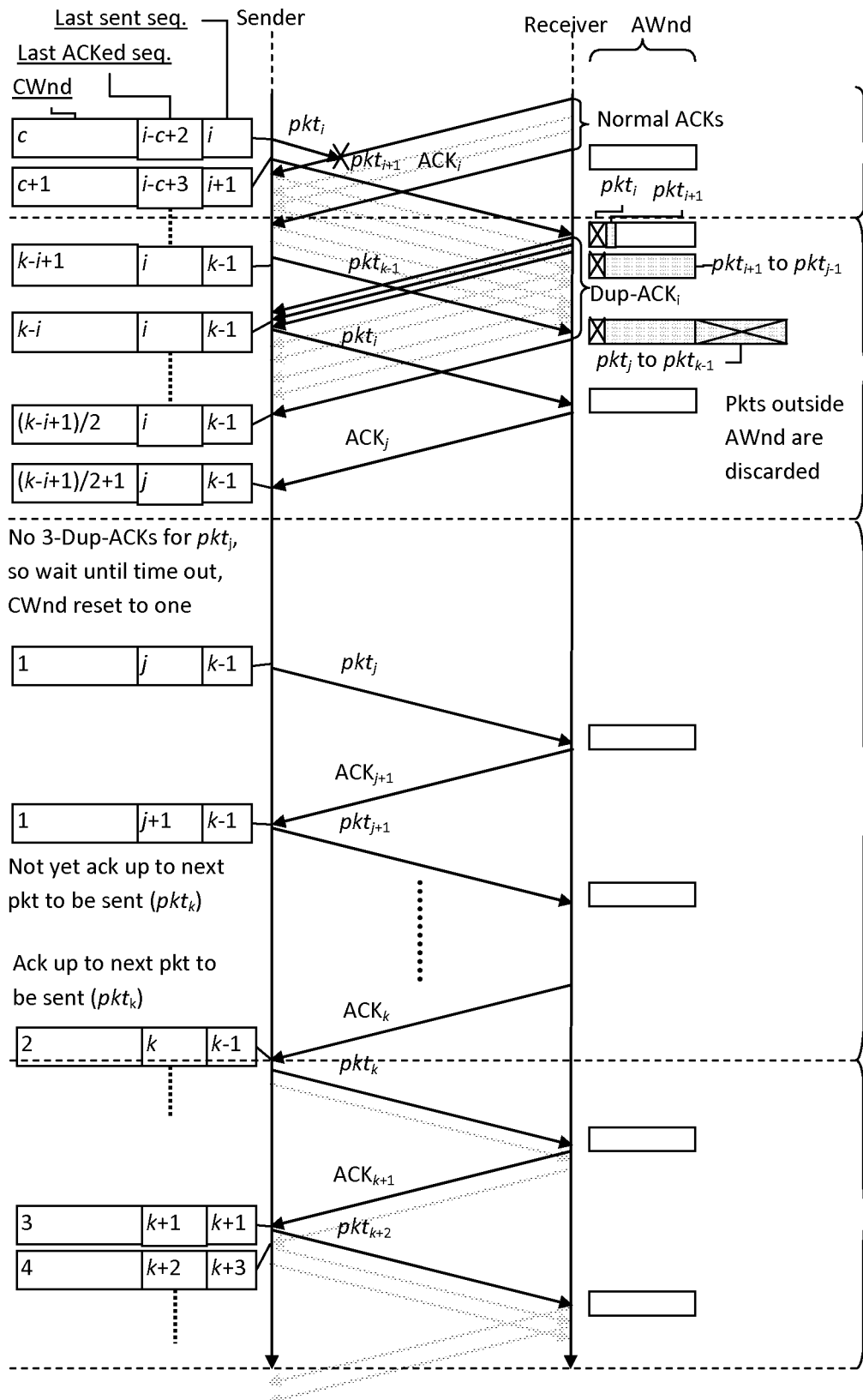


FIG 10

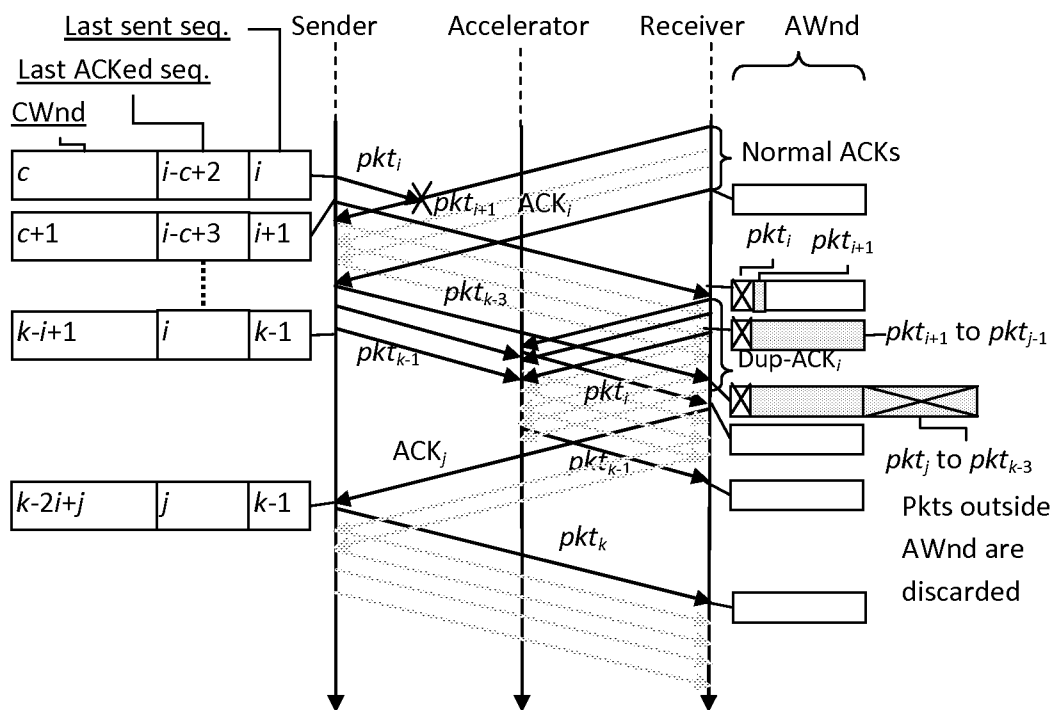


FIG 11

```
i ← 0 //start from the 0-th packet
loss ← NULL
WHILE i != tail+1 DO //tail marks the last packet received from sender
    IF loss == NULL THEN
        schedule_transfer(pkti) //request the scheduling algorithm to transfer
    ELSE
        IF  $i \geq loss + a_{\max}$  THEN //  $a_{\max}$  as the maximum AWnd
            schedule_transfer(pkti)
        END IF
    END IF

    IF DupACK == NULL THEN //DupACK is packet no. to be retransmitted
        i = i + 1
    ELSE
        i ← DupACK
        loss ← DupACK
        DupACK ← NULL
    END IF
END DO
```

FIG 12

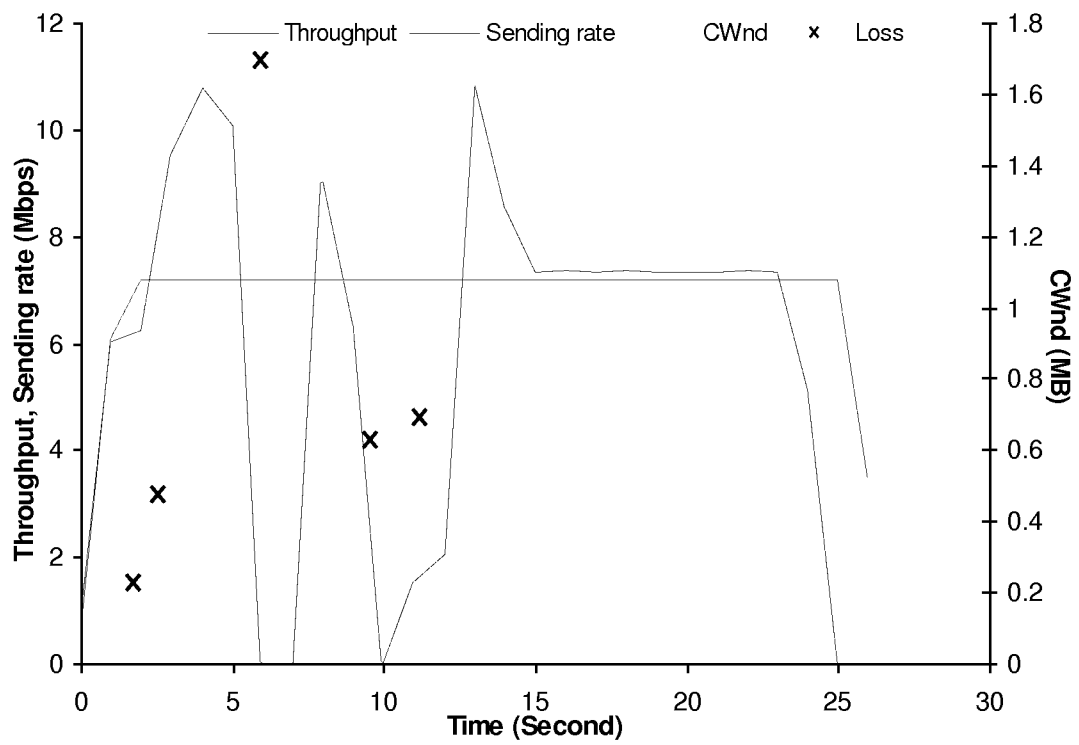


FIG 13

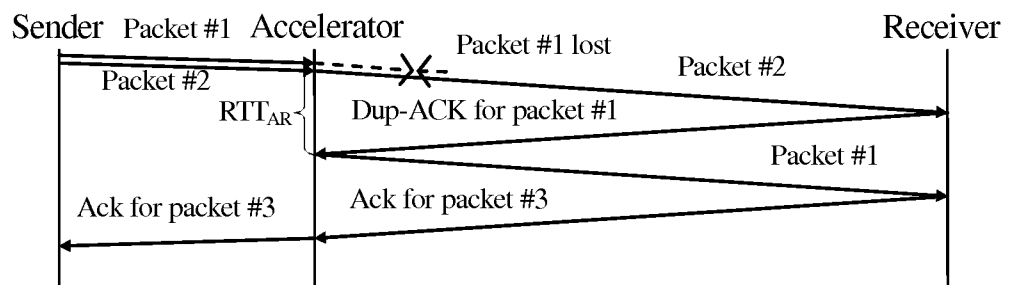


FIG 14

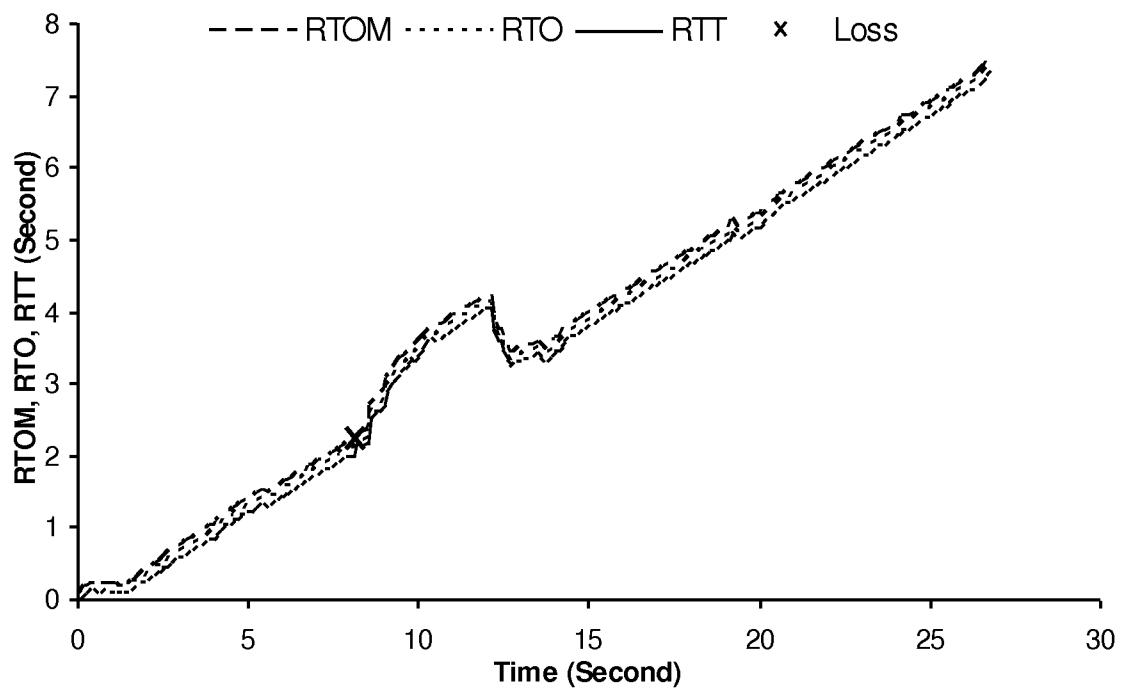


FIG 15

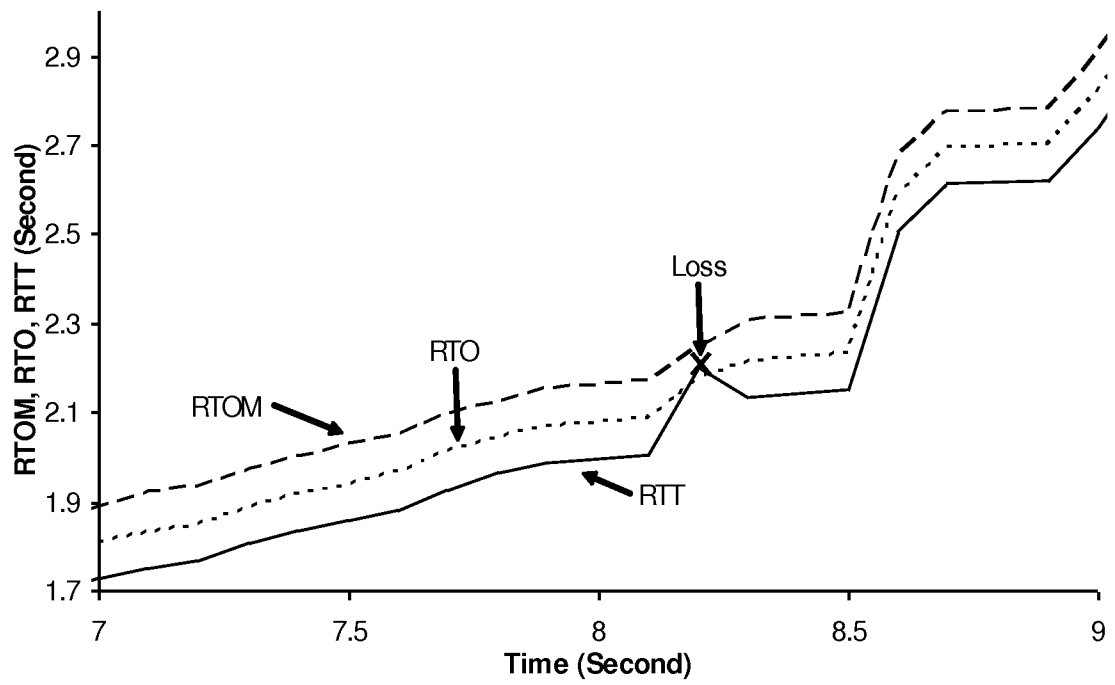


FIG 16

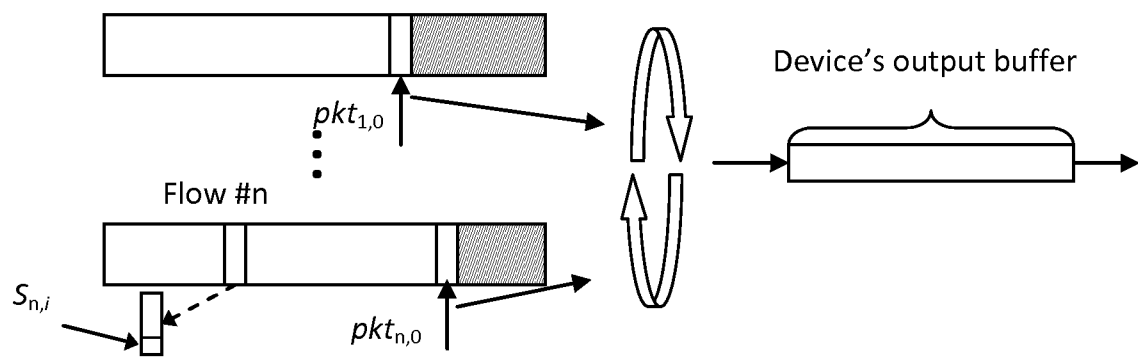


FIG 17

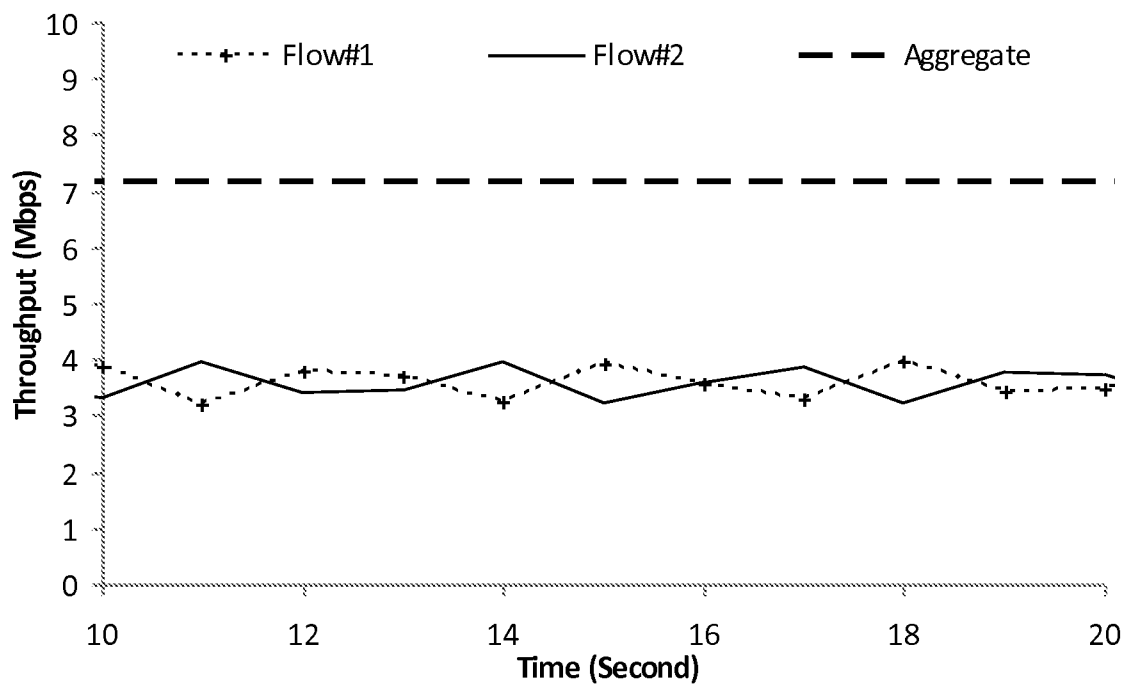


FIG 18

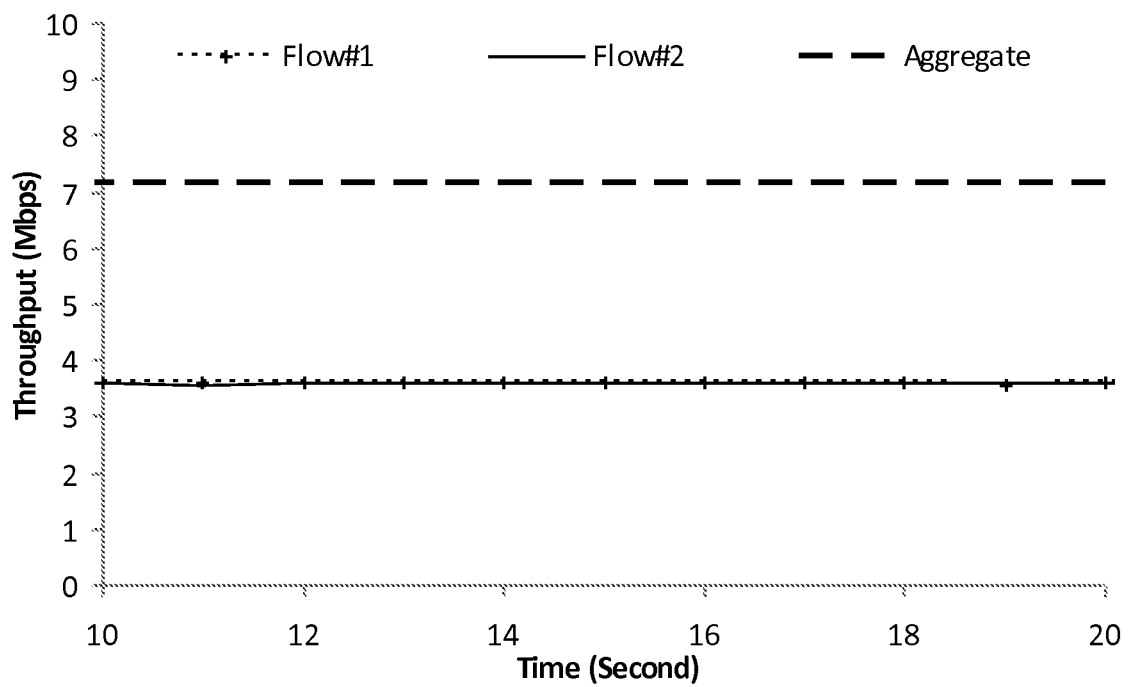


FIG 19

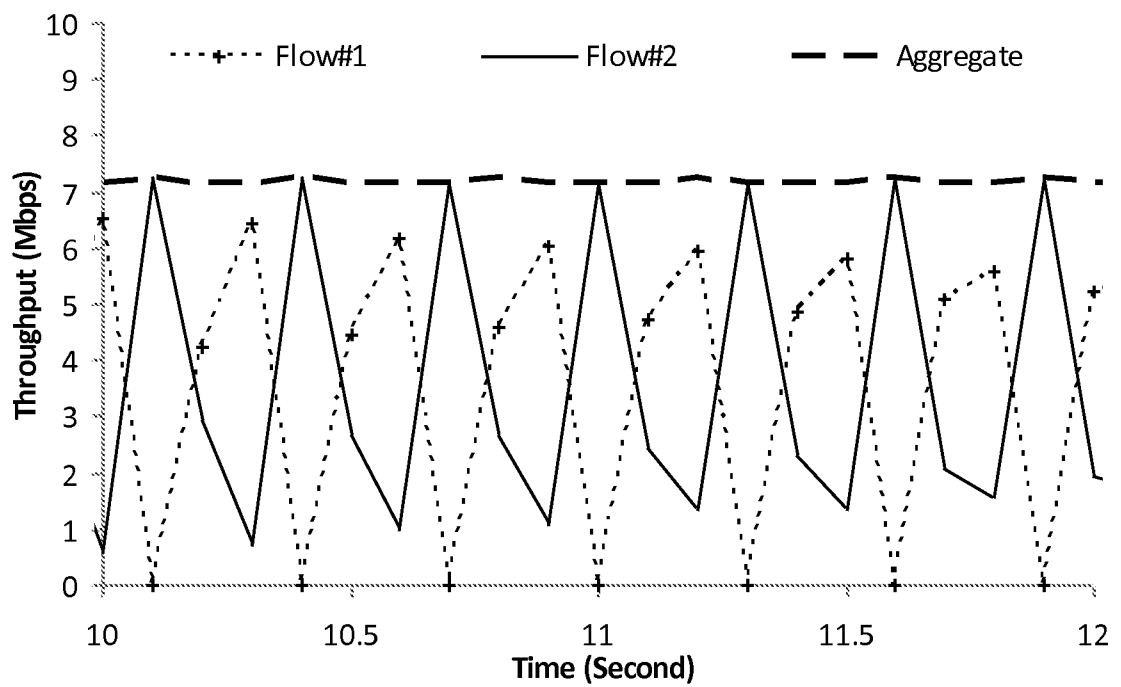


FIG 20

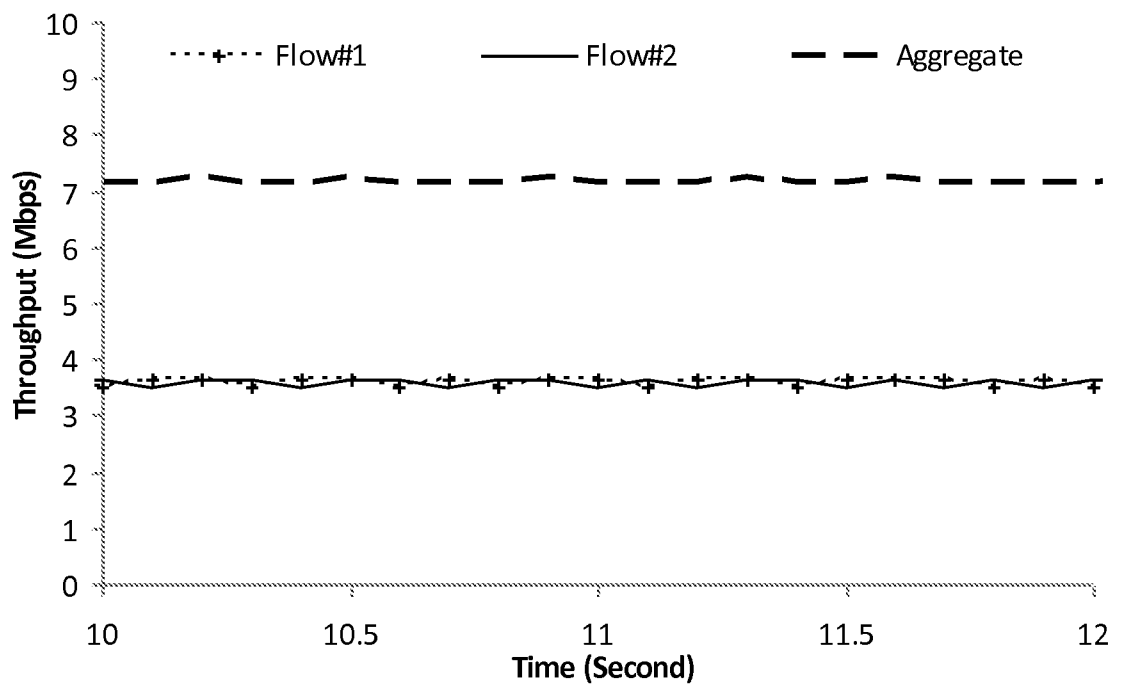


FIG 21

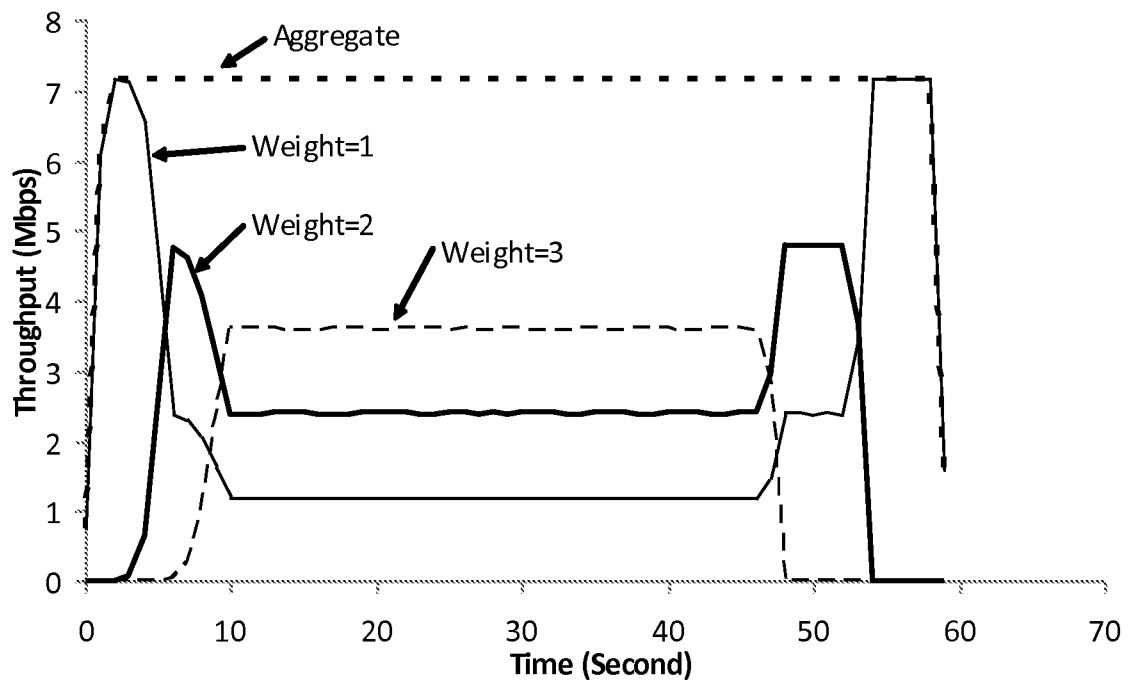


FIG 22

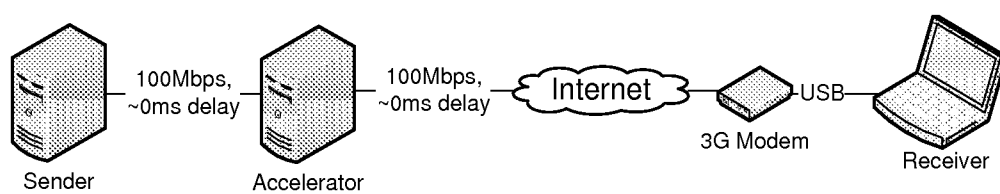


FIG 23

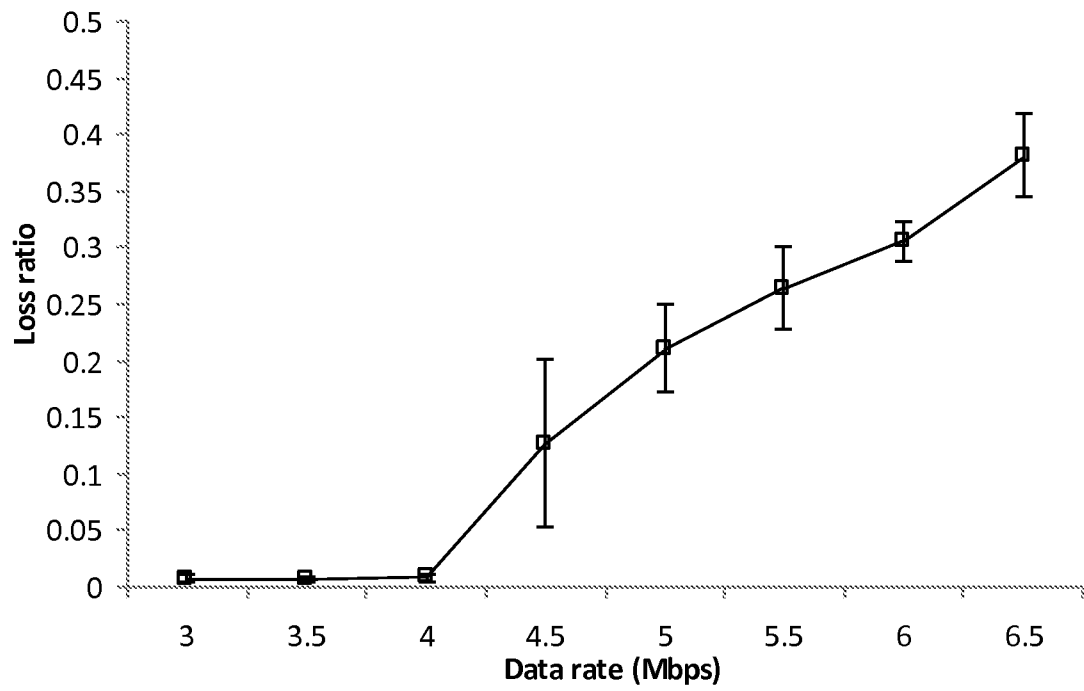


FIG 24

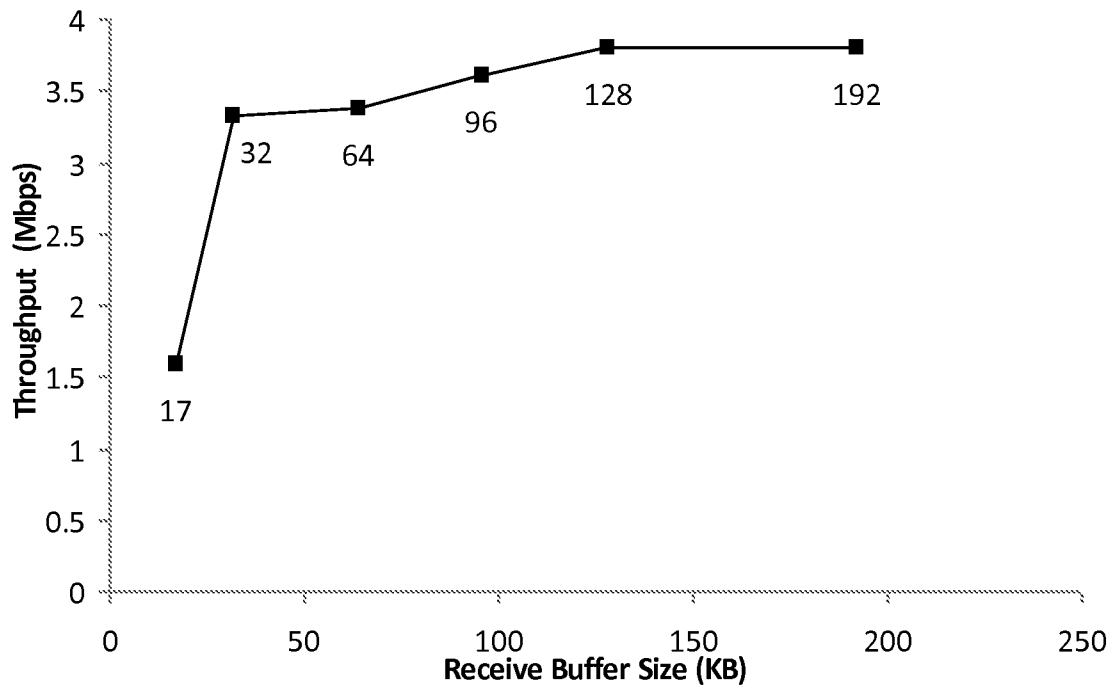


FIG 25

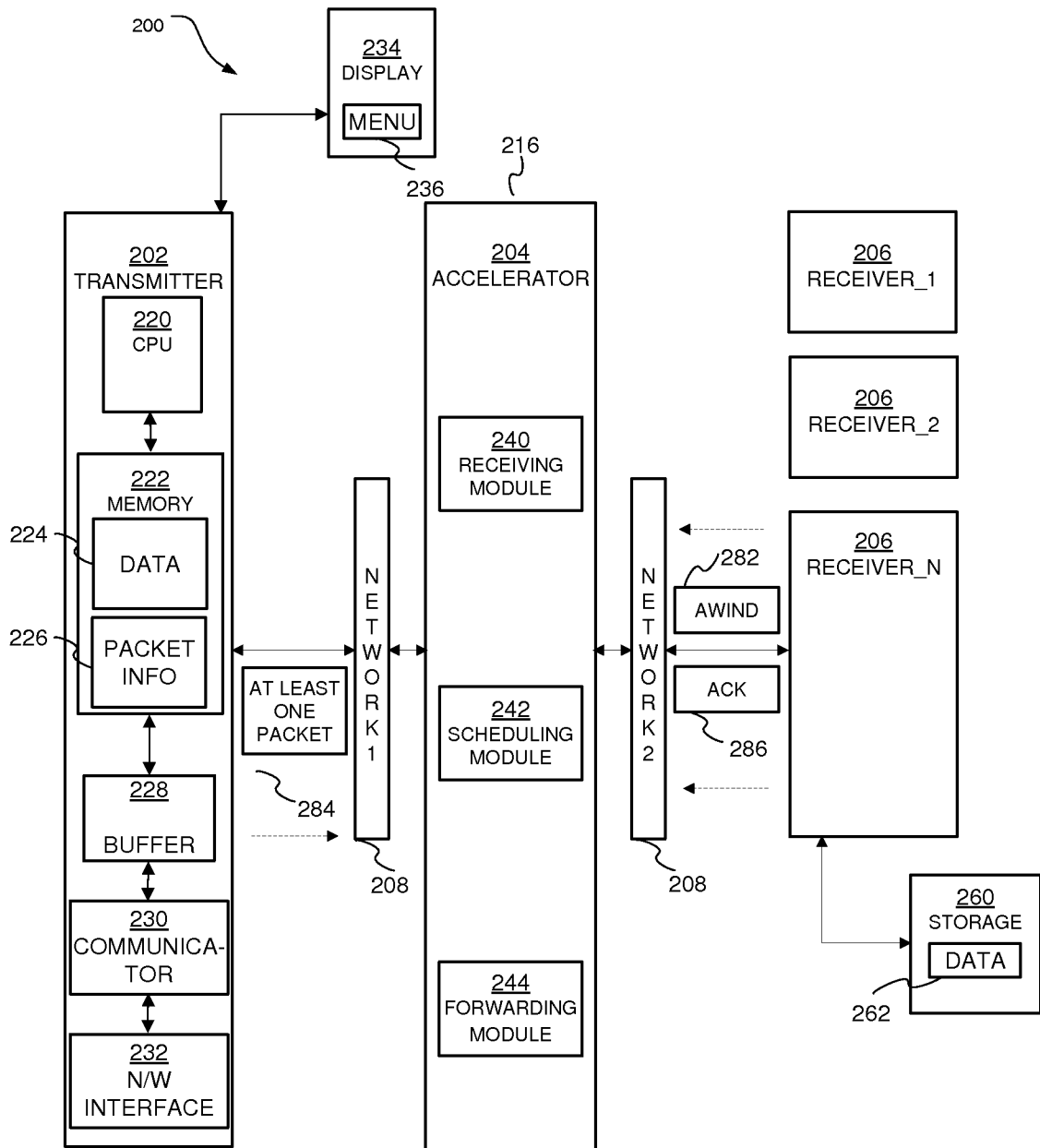


FIG 26

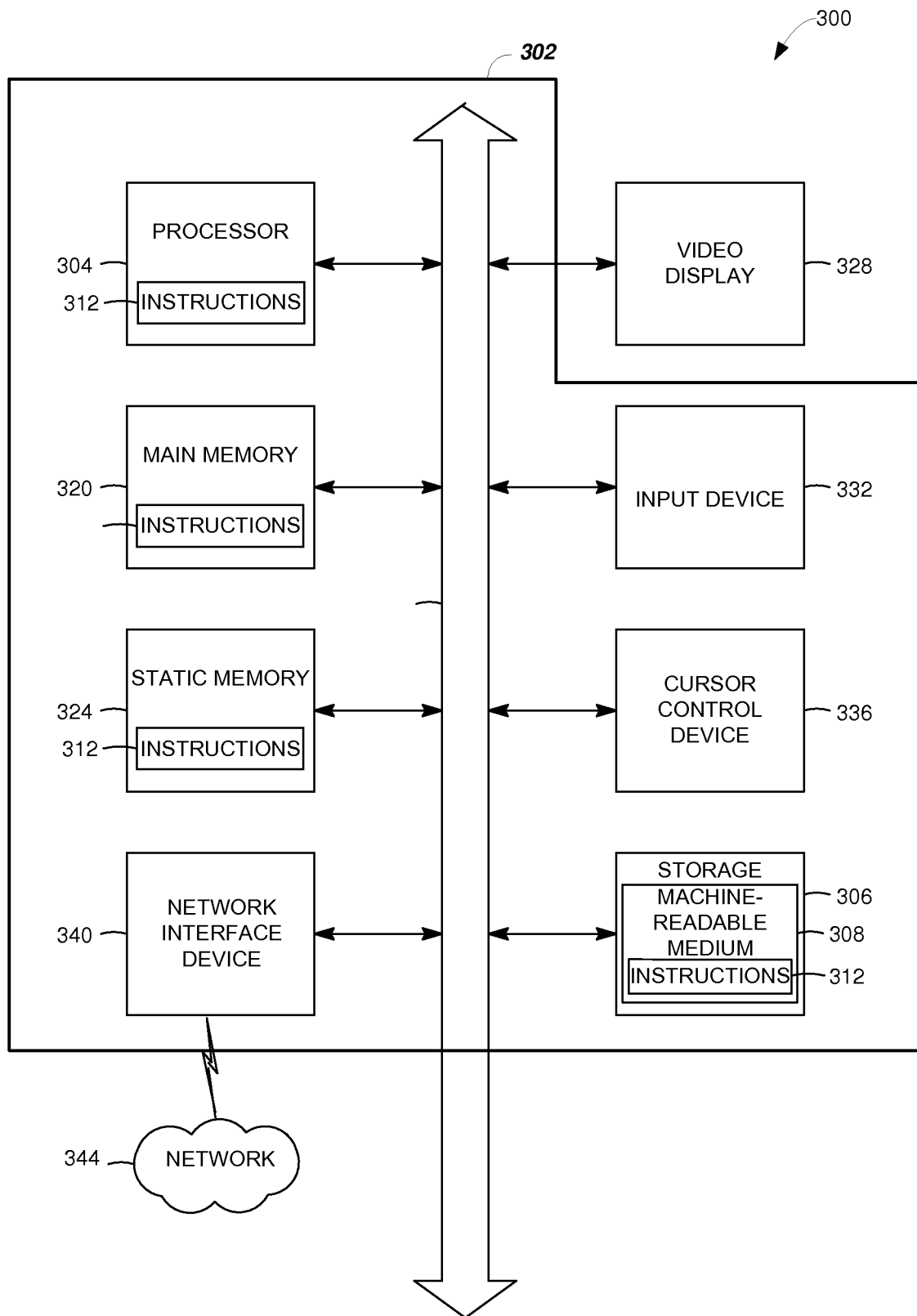


FIG 27

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2011/084762

A. CLASSIFICATION OF SUBJECT MATTER

See extra sheet

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

IPC: H04L;H04W; H04Q

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNPAT, CNKI, WPI, EPODOC, IEEE: network, first, second, flow, data, packet, stream, over, over w flow, over w run, congest+, buffer, receiver, time+, delay, window, transmit, transmission, forward, schedule, coordinate

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	CN1778079A(CISCO TECHNOLOGY, INC.) 24 May 2006(24.05.2006) page 7 line 2- page 8 line 13, in the description, figures 4-5	1-32
Y	CN1787487A(INTERNATIONAL BUSINESS MACHINES CORPORATION) 14 Jun. 2006 (14.06.2006) page 4 line 16-page 8 line 29, in the description, figures 1, 3	1-32
A	CN101040278A(INVENSYS SYSTEMS, INC.) 19 Sep. 2007(19.09.2007) the whole document	1-32
A	US6438101B1(AT&T CORP. et al.) 20 Aug. 2002(20.08.2002) the whole document	1-32

☐ Further documents are listed in the continuation of Box C.☒ See patent family annex.

* Special categories of cited documents:

“A” document defining the general state of the art which is not considered to be of particular relevance

“E” earlier application or patent but published on or after the international filing date

“L” document which may throw doubts on priority claim (S) or which is cited to establish the publication date of another citation or other special reason (as specified)

“O” document referring to an oral disclosure, use, exhibition or other means

“P” document published prior to the international filing date but later than the priority date claimed

“T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

“X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

“Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

“&”document member of the same patent family

Date of the actual completion of the international search
05 Mar. 2012(05.03.2012)Date of mailing of the international search report
05 Apr. 2012 (05.04.2012)Name and mailing address of the ISA/CN
The State Intellectual Property Office, the P.R.China
6 Xitucheng Rd., Jimen Bridge, Haidian District, Beijing, China
100088
Facsimile No. 86-10-62019451

Authorized officer

LU,Xia

Telephone No. (86-10)62413397

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2011/084762

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN1778079A	24.05.2006	WO2005002155A1	06.01.2005
		US2005141419A1	30.06.2005
		EP1634415A1	15.03.2006
		AU2004253083A1	06.01.2005
		CA2521461C	05.04.2011
CN1787487A	14.06.2006	US2006120289A1	08.06.2006
CN101040278A	19.09.2007	US2006059268A1	16.03.2006
		WO2006031589A2	23.03.2006
		EP1810174A2	25.07.2007
US6438101B1	20.08.2002	NONE	

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2011/084762

CLASSIFICATION OF SUBJECT MATTER

H04L 12/66 (2006.01)i

H04L 29/02 (2006.01)i