Delay Characterization of Cable Access Networks

Neil Barakat, *Member, IEEE,* and Thomas E. Darcie, *Fellow, IEEE*

***Abstract*— This letter presents a detailed characterization of the transmission delay in cable modem (CM) access networks. We analyze data obtained from measurements on operational access networks, examining both a moderately-loaded and a heavily-loaded CM network. We find that the medium access control algorithm used in CM networks results in a multimodal delay distribution, with measured delays clustered around a few discrete values. In the heavily-loaded CM network, a significant fraction of packets experienced delays that were multiple times larger than the average. The results imply that mean delay is a poor metric for measuring the performance of CM networks, especially when the network is heavily loaded.**

***Index Terms*— Access network, cable modem, communication network, delay, jitter, medium access control, network measurement.**

I. Introduction

OVER the past few years, cable modem (CM) networks have become one of the leading technologies for providing broadband Internet access to home users. There are currently over 26 million users in North America alone [1], and this number continues to grow.

The past few years has also seen the emergence of a number of real-time, interactive multimedia Internet applications such as voice-over-ip (VoIP) telephony, online gaming, and online music collaboration [2]. These applications have strict end-to-end delay requirements, so network latency is as important as bandwidth in determining the quality of users' experience. Providing low-latency connectivity to such applications is, therefore, an important and challenging problem facing today's network designers.

The problem of access-network latency is particularly important because packets tend to experience relatively large delays in traversing these networks. For example, in examining the delay between a CM used in this study and a remote Internet server located 8-hops ($\sim$ 150 km) away, we found that the access delay was responsible for approximately 75% of the overall round-trip delay and almost all of the experienced jitter. Thus, understanding of the delay and jitter characteristics of access networks is important for providing users with low end-to-end delay.

Although there have been a number of studies on CM access network delay (e.g., [3]–[5]), all past studies have only examined the average or standard deviation of delay. While such these metrics are useful, they give little insight into the variability of delay that is experienced over short time scales. Short-term delay variations can strongly impact the performance of CM networks, especially in the context of real-time applications, for which the untimely reception of packets results in data loss and poor user experience.

In this letter we provide a detailed characterization of transmission delay in CM networks. In addition to average delay, we also examine the delay distribution of CM networks and time of day variations of delay. We find that the medium access protocol used in CM networks has a significant impact on their delay profile and results in a uniquely shaped delay distribution that shows evidence of significant jitter and delay variability, especially in overloaded or poorly provisioned CM networks.

II. Cable Modem Network Medium Access Control

The hybrid fibre-coaxial cable infrastructure upon which cable-modem access networks are built is inherently broadcast in nature, so multiple CMs share upstream and downstream bandwidth. A cable modem termination system (CMTS) resides at the head-end of a CM network. The CMTS serves as a gateway for all upstream and downstream packets in the CM network and coordinates the CM transmissions using the medium access control (MAC) algorithm defined in the Data Over Cable Service Interface Specification (DOCSIS®) [6].

The CM MAC protocol relies on a request-grant process to share upstream bandwidth between CMs and the CMTS. Bandwidth in a CM network is divided into time slots. The CMTS acts as the principal controller in the CM network and specifies how each time slot is to be used. For example, certain slots are allocated for data transmission, others are allocated for bandwidth requests, and others for initialization and channel maintenance. When a CM is ready to transmit a packet, it generates a *bandwidth request* message, which it transmits to the CMTS in a future bandwidth-request slot. After receiving this message, the CMTS performs a scheduling operation and transmits a *bandwidth grant* message back to the CM indicating the future slot in which the CM should transmit its packet.

The CMTS generally allocates each data slot to only one CM, so upstream packet transmissions are not subject to contention. By contrast, multiple CMs may attempt to transmit a request in the same time slot. When this occurs, the requests collide and none of them are successfully received by the CMTS. CMs do not implement direct detection of bandwidth-request collisions, but instead rely on a timeout-like mechanisms to determine that a collision has occurred. After sending bandwidth requests, each CM monitors the downstream channel and waits for the *absence* of a reply from the CMTS in order to implicitly detect that a collision has occurred. Then each CM attempts to retransmit its request

Manuscript received November 6, 2006. The associate editor coordinating the review of this letter and approving it for publication was Dr. Jae Kim.

The authors are with the University of Victoria, PO Box 3055 STN CSC, Victoria, British Columbia V8W 3P6, Canada (e-mail: nbarakat@uvic.ca).

Digital Object Identifier 10.1109/LCOMM.2007.061821.

1089-7798/07$25.00 © 2007 IEEE

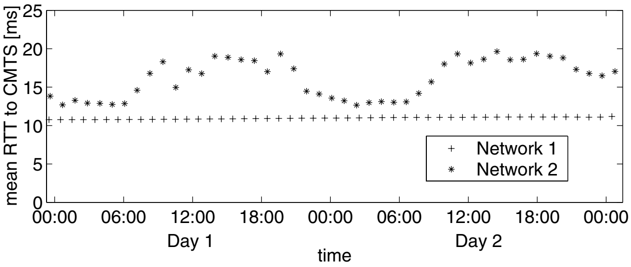


Fig. 1. Average (per hour) round-trip-delay to the CMTS over a 48-hour period in two commercially deployed CM networks. For the more lightly-loaded of the two networks (Network 1), the average delay remains almost constant, while for the more heavily-loaded network, time-of-day variations are apparent.

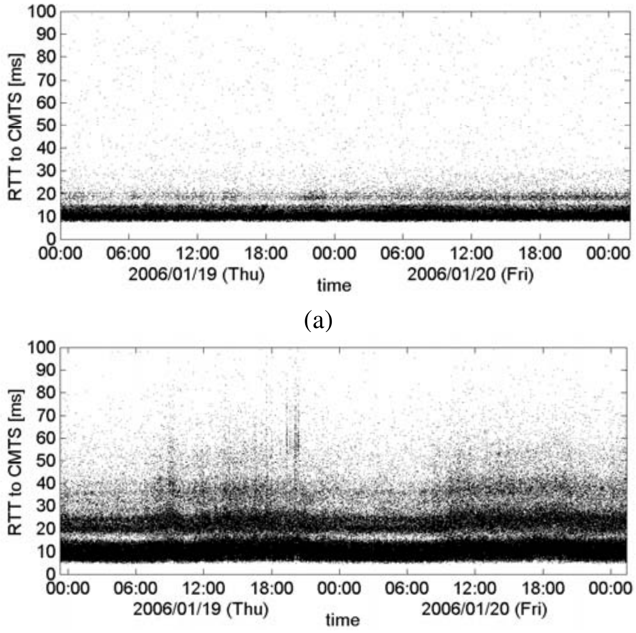
in a later time slot after waiting a random amount of time determined by an exponential back-off algorithm.

The above timeout mechanism for collision detection, and the use of the request/grant cycle for upstream transmission has a dramatic effect on the delay distribution of CM networks. In particular, the additional delay caused by each contention and subsequent back-off includes a large constant component that is approximately equal to the duration of the request/grant cycle of the system. As we show in the next section, this fixed-delay contribution has a large impact on transmission delay and results in a delay distribution with multiple, evenly-spaced peaks. In moderately-loaded systems, the vast majority of packets lie in the first peak. However, for overloaded or poorly provisioned networks, excessive request-channel contention causes a large fraction of packets to experience delays that are multiple times larger than the average.

In general, high delay variability in CM networks can be reduced by reducing the amount of traffic (i.e. number of users) on a CM network. Since this represents a reduction in the overall capacity of the network, an alternate approach would be preferable. Having recognized that the additional delay introduced by contention and retransmission in the CM MAC protocol might be prohibitive for future real-time applications, alternate mechanisms for providing contention-free bandwidth to CMs were integrated into DOCSIS 1.1 (and DOCSIS 2.0). For example, via the unsolicited grant service (UGS) model, a CM can be allocated a periodic train of upstream transmission slots thereby avoiding the delay caused by bandwidth-request contentions. However, cable operators have been slow to integrate these mechanisms into their service offerings, and we are unaware of any currently deployed CM networks that make these enhanced quality of service mechanisms available to CM users.[1] Thus, all data packets transmitted via CM networks still rely on the request/grant CM MAC protocol described at the start of this section.

III. Experimental Results and Discussion

The experiment in this study was conducted on commercially deployed CM networks. In order to sample the delay on the first hop of the CM networks, we periodically transmitted ICMP echo requests ('pings') from a host computer to its CMTS for a period of two days. We then used the measured delay between the transmission and reception of each ICMP request-reply pair as an estimate of the round-trip transmission time (RTT) between the host computer and the CMTS. Messages had a payload of 50 bytes and were transmitted approximately every 100 ms. The 100 ms inter-transmission time was small enough to provide us with a relatively dense sampling of the network delay and large enough to avoid significant correlation in adjacent samples.



(a)

(b)

Fig. 2. Raw data trace of round-trip delay measured for the two networks examined in Fig. 1. The request/grant cycle used in the CM MAC protocol results the pronounced horizontal bands in the structure of both traces.

Average delay is the most commonly studied delay parameter for CM networks [3]–[5]. Fig. 1 plots the measured average delay versus time for each hour in the experiment duration for both a moderately-loaded and heavily-loaded CM network. For the moderately-loaded system, the average latency is equal to approximately 10.75 ms with little time-of-day variation. By contrast, a clear time-of-day dependency is visible in the heavier-loaded CM-network data, as average delay values range from 20 ms in the daytime to 12 ms in the nighttime. This diurnal variation can be attributed to reduced network loading and congestion during the nighttime hours.

In addition to average delay, variation of delay (jitter) is of critical importance to many real-time applications. Fig. 2 plots the raw trace data obtained from the experiment for both CM networks. The traces show an interesting banded structure, as most data points are quite noticeably clustered around multiple distinct delay values. These bands are likely a direct result of the upstream MAC algorithm used in CM networks, with the first, second, and third bands corresponding to successful transmissions on the first, second, and third bandwidth request-attempts respectively.

For the moderately-loaded network in Fig. 2a), the majority

[1] In the past year, cable companies have begun to use some of the advanced quality-of-service features of DOCSIS 1.1 to provide cable telephone service to customers. However, the telephony service is separate from the CM data service, relying on dedicated channels in the network and separate hardware at the user premises.

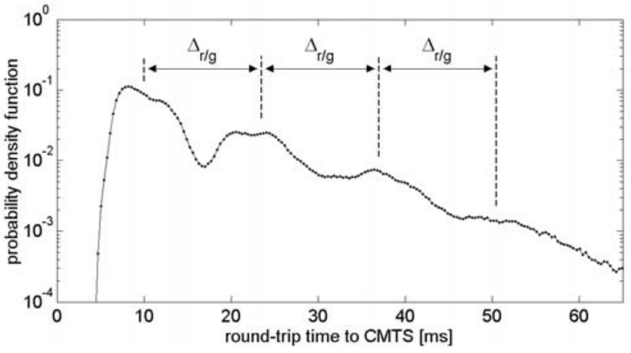


Fig. 3. Probability density function of the measured round-trip delay between a cable modem and the CMTS for the heavily-loaded (under-provisioned) CM network data shown in Fig. 2b). The contention resolution procedure of the CM MAC protocol results in a broadly-shaped distribution with multiple, evenly spaced peaks, each separated by the request-grant cycle duration $\Delta_{r/g}$ of the CM MAC protocol.

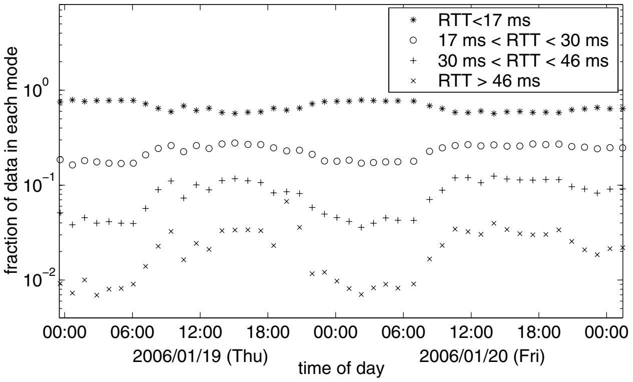


Fig. 4. Fraction of packet delays that lie in each of the modes of the distribution in Fig. 3. The jitter and delay in the network increases significantly during the daytime, when network congestion leads to a higher number of contentions on the bandwidth-request channel, resulting in one or more retransmissions being required.

of packets experience delays close to the mean RTT of the system, as approximately 96.5% of data points lie in the lowest band. By contrast, in the heavily-loaded network in Fig. 2b), a significant fraction of data points also reside in the higher bands. While this has a relatively minor effect on the average delay in the system (as seen in Fig. 1), the pronounced presence of these higher bands could be quite problematic for real-time, low-latency applications, because many packets will experience delays much higher than the average.

To examine this unique multi-banded delay structure more closely, Fig. 3 shows the probability density function (pdf) of the trace data from Fig. 2b). The overall delay distribution appears to be multi-modal, with a separate mode or peak corresponding to each band in the trace in Fig. 2. As expected from the discussion in Section II, the peaks are equally spaced with a spacing equal to the request-grant cycle duration ($\Delta_{r/g}$) of the CM MAC protocol.

To better understand the potential performance of low-latency applications deployed over CM networks, one would like to quantify the fraction of packets that lie in each band of Fig. 2b). This is plotted using a log scale in Fig 4 as a function of the time of day. To compute each curve, we used the local minima between the peaks in Fig. 3 as delineation points between modes, then counted the fraction of points that lay in each mode.

Examining the lowest mode (RTT< 17 ms) in Fig. 4, we see that the number of packets requiring only one bandwidth-request transmission varies from as high as 80% in the nighttime to as low as 60% in the daytime. As one might expect, we also observe that the weights of the three other modes follow an opposite pattern. The mirrored fluctuations between the lowest mode and the three highest modes in Fig. 4 implies that the diurnal fluctuations in the average delay seen in Fig. 1 are due not to a relatively small and uniform increase in the packet delay, but are rather due to a large increases in the delay of a minority of packets. This has significant implications for the performance of real-time and low-latency applications because it means that a non-negligible fraction of their packets will experience delays that are significantly larger than the mean delay of the network. The number of packets experiencing these large delays depends strongly on the traffic load and level of congestion in the CM network.

IV. Conclusions

In this letter, we examined the transmission delay characteristics of CM access networks. We found that the request/grant cycle and contention detection mechanism used in the CM network caused some packets to experience round-trip delays that were multiple times larger than the mean delay of the network. The resulting delay distribution was multimodal in shape with multiple distinct peaks, each corresponding to a different number of transmission attempts. The results also imply that average delay is a poor metric for the performance of CM networks. The results highlight the importance of adequately provisioning resources for upstream transmission in CM networks, and provide additional motivation to the cable community to hasten their adoption of alternate bandwidth sharing mechanisms, such as those defined in the DOCSIS PacketCable™Multimedia specifications [7], so that they can better support emerging low-latency and real-time applications

References

[1] (2005, Dec.) Baby bells trump cable on data front. [Online]. Available: http://www.cabledatacomnews.com/dec05/dec05-5.html

[2] G. Xiaoyuan, M. Dick, Z. Kurtisi, U. Noyer, and L. Wolf, "Network-centric music performance: practice and experiments," *IEEE Commun. Mag.*, vol. 43, no. 6, pp. 86–93, June 2005.

[3] T. T. Nguyen and G. J. Armitage, "Experimentally derived interactions between TCP traffic and service quality over DOCSIS cable links," in *Proc. IEEE Globecom '04*, pp. 1314–1318.

[4] W.-T. Lee, K.-C. Chu, K.-C. Chung, J.-Y. Pan, and P.-C. Chung, "Scheduling delay minimization for non-UGS data in multi-channel HFC network," *IEICE Trans. Commun.*, vol. E88-B, no. 2, pp. 623–631, 2005.

[5] V. Sdralia, C. Smythe, P. Tzerefos, and S. Cvetkovic, "Performance characterization of the MCNS DOCSIS 1.0 CATV protocol with prioritised first come first served scheduling," *IEEE Trans. Broadcast.*, vol. 45, no. 2, pp. 196–205, 1999.

[6] Docsis specification. [Online]. Available: http://www.cablemodem.com/

[7] Packetcable multimedia specification. [Online]. Available: http://www.packetcable.com/specifications/multimedia.html