Estimation of the Packet Processing Time of Hosts in the Presence of Interrupt Coalescence

Khondaker M. Salehin and Roberto Rojas-Cessa

Abstract: We propose a method to measure the packet processing time (PPT) of remote hosts that use network interface cards (NICs) with interrupt coalescence (IC). The model is based on sending a pair of probe packets. The proposed method first detects whether the host under test uses IC and then measures PPT. The presented experimental evaluation under 100- and 1000-Mb/s transmission speeds shows that the proposed method consistently measures PPT with high efficacy.

Keywords: Active measurement, packet processing time (PPT), interrupt coalescence, compound probe, high-speed link

Classification: Internet

References

1 Introduction

The packet processing time (PPT) of a host (i.e., workstation) is the time elapsed between the arrival of a packet in the host’s input queue of the network interface card, NIC, (i.e., the data-link layer of the TCP/IP protocol stack) and the time the packet is processed at the application layer [1]. The role of PPT becomes more significant in the measurement of different network parameters as link rates continue to increase faster than processing speeds [3].

Considering OWD [2] as an example, Figure 1 illustrates the end-to-end path packet $P$ follows between two end hosts: the source ($src$) and destination ($dst$) hosts. The figure also shows the different layers of the TCP/IP protocol stack that $P$ traverses at both end hosts. As defined in RFC 2679 [1], actual OWD ($OWD$) is the wire time $P$ experiences in the trip from $src$ to $dst$. The wire time includes the transmission time ($t_t$), the queueing delay ($t_q$), and the propagation time ($t_p$), or $OWD = t_t + t_q + t_p$, as Figure 1 shows. However, as time stamping after packet creation at $src$ ($PPT_{src}$) and packet receiving at $dst$ ($PPT_{dst}$) takes place at the application layer, the measured OWD includes these PPTs plus the interrupt coalescence (IC) of the NICs [4, 5, 6, 7] as an apparent OWD ($OWD'$) of $P$ between $src$ and $dst$, or:

$$OWD' = PPT_{src} + IC_{src} + OWD + PPT_{dst} + IC_{dst}$$

At $dst$, IC is the time interval over which a NIC buffers more than one packet at its input queue before sending an interrupt to the central processing unit (CPU) for time stamping them at the application layer [4, 5]. The time stamping at the application layer is considered to be the packet processing function at end hosts [8]. Because of the small transmission rates of legacy systems, PPT and IC have been considered so far negligible in the measurement of OWD and other network parameters.

As transmission rates increase, the contribution of PPT on OWD increases, and the error in the measurement of OWD in high-speed networks can be large if PPTs are neglected [9, 8]. Therefore, PPT and IC must be considered in accurate OWD measurements. Accurate OWDs are required...
for precision-timing applications, e.g., IP geolocation [9] and financial trading [9, 8].

The measurement of PPT of a host is complex. It requires recording the
time when a packet arrives at data-link layer and the time application layer
takes to process the packet. A scheme, based on Internet Control Message
Protocol (ICMP) packets, measures the PPT of a host in a single-hop sce-
nario [9]. This scheme instruments the data-link layer of the host under test
to time stamp the received ICMP echo packets. Another scheme measures
PPT of a remote host in a multiple-hop scenario from link-capacity estima-
tion using pairs of UDP packets, called compound probes [10], without any
instrumentation at the host [8]. The existing schemes for PPT measurement
do not consider the effect of (non-negligible) IC when it is present. IC is
usually present in commodity high-speed NICs, e.g., 100 Mb/s and higher
rates, to reduce the processing load of hosts [4, 6]. Herein, we propose a
scheme to measure PPT of hosts with NICs using IC.

2 Proposed scheme

The proposed method also uses the compound probe, which consists of a
large heading packet ($P_h$) followed by a small trailing packet ($P_t$). Consider
that $s_t$ is the size of $P_t$ and and $l_n$ is the capacity of the end link connected
to the remote end host of an $n$-hop path. According to [8], the intra-probe
gap, $G(s_t)$, of the compound probe at the remote end host shows a linear
relationship with a non-zero slope as:

$$G(s_t) = \frac{s_t}{l_n} + \delta_n + \Delta \text{PPT},$$  \hspace{1cm} (2)

where $\delta_n$ is the cumulative dispersion (i.e., separation) between $P_h$ and $P_t$,
and $\Delta \text{PPT}$ is the differential PPT for time stamping the compound probe.
$\delta_n = 0$ when $P_h$ and $P_t$ arrive back-to-back at the remote host. However, (2)
does not hold if $\frac{s_t}{l_n}$ is smaller than the IC period of the remote host.

Figure 2 illustrates IC on a NIC and the basic principle of the proposed
method, respectively. Figure 2(a) shows the packet processing events con-
sidering a First-In-First-Out (FIFO) queueing policy as a compound probe
arrives at the remote host. Here, $P_h$ and $P_t$ arrive back-to-back inside the
IC period when the NIC buffers packets. After the expiration of the IC, two
time stamps $TS_h$ and $TS_t$, each generated with a constant delay, acknowl-
edge the buffered $P_h$ and $P_t$, respectively. Because the time stamping delay
corresponds to the travelling delay of a packet from application to data-link layer, the intra-probe gap is equivalent to the PPT of the remote host, or:

\[ PPT = G(s_t) \]  

(3)

Figure 2(b) shows the zero-slope linear relationship between the measured intra-probe gap and the trailing-packet size, as stated in (3). This relationship holds as long as both \( P_h \) and \( P_t \) arrive at the same IC period, as Figure 2(a) shows.

The detailed steps of the proposed method for measuring PPT of \( dst \) from \( src \) over an \( n \)-hop path are:

1. Send a train of compound probes from \( src \) to \( dst \) using a \( P_h \) with a size \( s_h = s_{max} \), where \( s_{max} \) is the Maximum Transmission Unit (MTU) of the path, and a \( P_t \) with \( s_1 < s_{max} \) such that the largest possible packet-size ratio, \( \alpha = \frac{s_h}{s_1} \), in the compound probes is obtained. Sizing of \( s_t \) in the compound probe for ensuring \( \delta_n = 0 \) at \( dst \) over a multiple-hop scenario can be determined based on the capacity of the links over the \( n \)-hop path [8, 10, 11].

2. Repeat Step 1 using \( P_t \)s with another \( n - 1 \), where \( n > 1 \), different \( s_t \)s, such that \( s_2 < s_3 < ... < s_n \). The minimum difference between consecutive \( s_t \)s is determined by the resolution of the system clock at \( dst \), e.g., 25 bytes over 100-Mb/s speed [8].

3. Determine the average intra-probe gap, \( G_{avg}(s_t) \), of each compound-probe train from the intra-probe gaps measured at \( dst \) after discarding the gaps with \( \delta_n > 0 \) [8].

4. Check if \( G_{avg}(s_t) \)s for multiple and consecutive \( s_t \)s have a zero-slope linear relationship in order to detect IC, such that

\[ G_{avg}(s_1) = G_{avg}(s_2) = ... = G_{avg}(s_n) \]  

(4)
5. If IC is detected in Step 4, estimate PPT from the constant value of $G_{avg}(s_1)$, e.g.,

$$PPT_{dst} = G_{avg}(s_1).$$  \hspace{1cm} (5)

Else, use the linear relationship between $G_{avg}(s_1)$ and $G_{avg}(s_2)$ to estimate PPT, as proposed in [8].

3 Experimental results

We evaluated the proposed scheme in a single-hop scenario. The performance of the scheme over a multiple-hop scenario is the same if $P_h$ and $P_t$ arrive back-to-back at $dst$. Back-to-back arrivals over multiple-hop paths have been successfully verified in testbed and Internet environments [8, 10].

We used two Linux based workstations, a Dell Optiplex 790 (DO790) and a Dell Studio XPS 435 MT (DS435), as $dst$. The kernel versions of these two workstations are 2.6.18 and 3.11.0, respectively. DO790 is equipped with an Intel Core i3 (3.30 GHz) processor and an integrated gigabit-Ethernet NIC, Intel 82579LM. In case of DS435, the workstation has an Intel Core i7 (2.67 GHz) processor along with an integrated gigabit-Ethernet NIC, Intel 82567LF-2. Both NICs are configured with a default IC period that ranges between 50 and 250 $\mu$s [6].

For PPT measurement, a large $P_h$, $s_h = 1512$ bytes, and seven $P_t$ sizes, $s_t = \{87, 112, 212, 512, 812, 1012, 1212\}$ bytes, are used in the compound probe. The packet sizes include 12 bytes of Ethernet encapsulation [10, 8]; they provide different $\alpha$s in the compound probe. We used 5000 compound probes, each separated by 100 ms.

Figure 3 presents the summary of $G_{avg}(s_t)$s measured on DO790 and DS435, in logarithmic scale and microseconds, for $s_t$s under 100- and 1000-Mb/s transmission speeds. In this figure, the hollow and the solid circles show the theoretical (i.e., $c_{lin}$) and estimated $G_{avg}(s_t)$s, respectively. In case of DO790, the measured $G_{avg}(s_t)$s have a zero-slope linear trend for all $s_t$s under both speeds, according to Figures 3(a) and 3(b), respectively, as $G_{avg}(s_t)$s have a constant value of 3 $\mu$s. The zero-slope line shows the presence of IC on the NIC and after proceeding with measurement, the PPT of DO790 is 3 $\mu$s.

The summary of $G_{avg}(s_t)$ measurements on DS435 under 100 and 1000 Mb/s also show a zero-slope linear trend in Figures 3(c) and 3(d), respectively. Therefore, IC is detected on DS435 where $G_{avg}(s_t) = 11$ $\mu$s with 1-$\mu$s variation for all $s_t$s. The 1-$\mu$s variation is produced by to the limited clock resolution (i.e., 1 $\mu$s) of the Linux system [8]. Based on the constant $G_{avg}s_t$, the measured PPT of DS435 is around 11 $\mu$s.

To verify the proposed method, we set DO790 and DS435 at a static IC period = 125 $\mu$s and measured their actual PPTs using the ICMP-packet-based scheme, as the reference scheme, which estimates the sum of PPT and IC. Other than the reference scheme, there exists no other scheme that measures PPT in the presence of IC according to the best of our knowledge. The actual PPTs are determined after removing the IC from the final estimates.
The theoretical gap without IC and the estimated gap with IC are calculated for different packet sizes under 100 Mb/s and 1000 Mb/s. The results show that the actual PPTs of DO790 under 100 and 1000 Mb/s are $6 \pm 25 \mu s$ and $9 \pm 25 \mu s$, respectively. For DS435, the values are $37 \pm 31 \mu s$ and $20 \pm 30 \mu s$. The large standard deviations in the actual PPTs are due to variations in the IC and ICMP packet generation time of the reference scheme. The PPTs measured by the proposed method are $3 \pm 1 \mu s$ and $11 \pm 1 \mu s$ on DO790 and DS435, respectively, which are within the ranges of the actual PPTs. This verifies the consistency of the proposed method in measuring PPT with high efficacy.

4 Conclusion

We proposed a method to measure the PPT of remote hosts that use IC in their NICs. The experimental results show that the measured PPTs are consistently within the ranges of actual PPTs and confirm the high efficacy of the scheme in detecting IC and measuring PPT.