The effect of layer-2 switches on pathchar-like tools

Ravi S. Prasad
Georgia Tech
ravi@cc.gatech.edu

Constantinos Dovrolis
Georgia Tech
dovrolis@cc.gatech.edu

Bruce A. Mah
Packet Design
bmah@packetdesign.com

Starting from Jacobson's *pathchar* [1], there has been a significant interest in techniques that can measure the *link capacities*, a.k.a. *link rates*, of an IP path. Follow-up research papers and variations of *pathchar* are Downey's *clink* [2], Mah's *pchar* [3], and Lai's *tailgating* technique in *nettimer* [4]. The motivation for such *per-hop capacity estimation tools* is that they can be used, mainly by network managers, to debug, monitor, and characterize network paths. These tools use a common underlying measurement methodology, which we refer to as *Variable Packet Size (VPS)* probing. The key assumption in VPS is that each hop of a path increases the one-way delay of a packet by a *serialization latency*, given by the ratio of the packet size over that hop's capacity. If this is true, the VPS technique can estimate the capacity of a hop i based on the relation between the Round-Trip Time (RTT) up to hop i and the probing packet size [2]. The required RTTs for different packet sizes can be measured using *Time Exceeded* ICMP messages [5], as done by *traceroute* [6].

Unfortunately, the capacity estimates of VPS tools are sometimes wrong. The conventional wisdom is that routers generate ICMP replies through slow forwarding paths, or that it is impossible to avoid the measurement noise due to queueing delays. The main contribution of this report is to show that *store-and-forward Layer-2 (L2) devices, such as switches, in a Layer-3 (L3) hop can cause significant and consistent underestimation of that hop's capacity*. The reason is that L2 store-and-forward devices introduce additional serialization latencies, which are not taken into account by the VPS model, as those devices do not decrement the TTL field of the IP header, being effectively "invisible" to IP and higher-layer protocols.

Let us briefly describe the VPS measurement methodology first. Consider a path from a sender $\mathcal{SND}$ to a receiver $\mathcal{RCV}$ that consists of $H \geq 1$ hops. The capacity of each hop is C_i, for $i = 1, \ldots H$. The RTT from $\mathcal{SND}$ to hop I for a packet of size L can be measured setting the TTL of the packet to I. The RTT $T_I(L)$ of the packet is

$$T_I(L) = \sum_{i=1}^{I} \left(\frac{L}{C_i} + \tau_i^f + d_i^f + \frac{L_{ICMP}}{C_i^r} + \tau_i^r + d_i^r \right) \quad (1)$$

This work was supported by the SciDAC program of the US Department of Energy (award # DE-FC02-01ER25467), and it was completed while the first two authors were with the CIS department of the University of Delaware.

assuming, for simplicity, that the forward and reverse network paths go through the same set of routers. In the previous equation, the fraction L/C_i is the *transmission*, or *serialization*, *latency* of a packet of size L at hop i. The terms τ_i^f and d_i^f are the propagation and queueing delays, respectively, of the packet at hop i of the *forward* path, from $\mathcal{SND}$ to hop I. The terms L_{ICMP}/C_i^r, τ_i^r, and d_i^r, are the serialization, propagation, and queueing delays, respectively, of the ICMP reply at hop i of the *reverse* path, from hop I to $\mathcal{SND}$.

The VPS technique makes the following assumption: *if we measure the RTT $T_I(L)$ up to hop I with several probing packets, it is likely that the minimum RTT measurement $\tilde{T}_I(L)$ resulted from a packet, and a corresponding ICMP reply, which did not experience any queueing delays*. When this is the case for a certain RTT measurement $\tilde{T}_I(L)$, we have that $d_i^f = d_i^r = 0$. Consequently, the minimum RTT measurement $\tilde{T}_I(L)$ can be expressed as

$$\tilde{T}_I(L) = \sum_{i=1}^{I} \left(D_i + \frac{L}{C_i} \right) \quad (2)$$

where $D_i = \tau_i^f + \tau_i^r + \frac{L_{ICMP}}{C_i^r}$ is independent of L.

Suppose now that we measure the minimum RTT for different packet sizes L, with the constraint that L is not larger than the path Maximum Transmission Unit (MTU). From Equation 2, it is easy to see that the relation between the minimum RTT and the packet size L is linear. Thus,

$$\tilde{T}_I(L) = \alpha_I + \beta_I L \quad (3)$$

where $\alpha_I = \sum_{i=1}^{I} D_i$, while β_I *is the slope of the minimum RTT measurements $\tilde{T}_I(L)$ as L varies*. The RTT slope at hop I is given by

$$\beta_I = \sum_{i=1}^{I} \frac{1}{C_i} \quad (4)$$

VPS probing tools, measure the RTT-slope β_I at each hop I of the path, and then calculate the per-hop capacity C_I from the RTT slope differences $\beta_{I+1} - \beta_I$ between successive hops. In more detail, suppose that we have measured the slope β_1 at the first hop. The capacity C_1 can be then computed as $C_1 = 1/\beta_1$. Applying induction, suppose that we have already estimated the capacity of the first I hops using the slope measurements $\beta_i, i = 1, \ldots I$. The capacity of hop $I+1$ can be then estimated as

$$C_{I+1} = \frac{1}{\beta_{I+1} - \beta_I} \quad (5)$$

from the RTT slopes β_I and β_{I+1}

In the following, we show that L2 store-and-forward devices can cause significant capacity underestimation in VPS probing.

First, we distinguish between IP layer, or L3 links, and data-link layer, or L2 links. To differentiate between the two, we refer to L3 links as *hops* and to L2 links as *segments*. Each L3 hop includes at least one L2 segment, since the L3 connectivity is provided on top of L2 connectivity. Different links are interconnected with *packet forwarding devices*, or simply *devices*. We characterize devices as L3 or L2, depending on whether they operate at the IP layer. Consequently, *L3 devices decrement the Time-To-Live (TTL) field of the IP header before forwarding the packet to the next hop, while L2 devices do not*. So, L2 devices cannot be detected using ICMP, and they are essentially invisible to VPS probing.

Our model of a network path, including both L3 and L2 devices, is now revised as follows. An L3 hop i consists, in general, of $M_i \geq$1 L2 segments. The first segment of a hop is the outgoing interface in the corresponding L2 device. Additional L2 devices, such as switches of various technologies, can follow in that hop, however, before the next L3 device. The presence of such intermediate L2 devices (and segments) is common in LANs and campus networks. L2 switches are also used in some backbone networks to implement a fully-connected topology at the IP layer.

Each segment j of hop i has a capacity of $C^{L2}_{i,j}$ bits-per-second (bps), meaning that it can transmit a packet of size L bits in $L/C^{L2}_{i,j}$ seconds. The capacity C^{L3}_i of hop i is defined as the minimum of the L2 segment capacities in that hop.

$$C^{L3}_i = \min_{j=1\ldots M_i} \{C^{L2}_{i,j}\} \tag{6}$$

Given that L2 devices are invisible to IP and higher-layer protocols, it is clear that VPS probing cannot measure the capacity of each segment, at least with IP and ICMP packets. Instead, the objective is to estimate C^{L3}_i for each hop $i = 1, \ldots H$ of an end-to-end path.

Suppose that the first hop of the path consists of M_1 segments, or equivalently, M_1 store-and-forward L2 devices. Each segment $j = 1 \ldots M_1$ in that hop introduces a latency of $L/C^{L2}_{1,j}$ to a packet of size L, because of the serialization latency that store-and-forward devices cause. Consequently, the minimum RTT for a packet of size L is expected to be

$$\tilde{T}_1(L) = D_1 + \sum_{j=1}^{M_1} \frac{L}{C^{L2}_{1,j}} \tag{7}$$

Thus, the RTT slope of the first hop is

$$\beta_1 = \sum_{j=1}^{M_1} \frac{1}{C^{L2}_{1,j}} \tag{8}$$

and so the capacity estimate for the first hop, according to VPS, is

$$\hat{C}^{L3}_1 = \frac{1}{\sum_{j=1}^{M_1} \frac{1}{C^{L2}_{1,j}}} \tag{9}$$

as opposed to $C^{L3}_1 = \min_{j=1\ldots M_1} C^{L2}_{1,j}$, which is the correct value. It is easy to see that if $M_1 > 1$ then $\hat{C}^{L3}_1 < C^{L3}_1$, meaning that *the capacity of the first hop will be underestimated*. For the special case that all L2 segments have the same capacity ($C^{L2}_{1,j} = C^{L2}_1$ for $j = 1 \ldots M_1$), we have that

$$\hat{C}^{L3}_1 = \frac{C^{L2}_1}{M_1} \tag{10}$$

Even though the previous derivations were made for the first hop in a path, it is straightforward to apply them inductively to every hop in the path. In [7], we also show that L2 devices at a hop i do not affect the capacity estimate of subsequent hops. In other words, the underestimation effect of L2 devices is limited to only the L3 hop that includes those devices.

To illustrate the negative effect of L2 devices on VPS tools, we next show some experimental data that resulted from the publicly-available VPS tools *pathchar*, *clink*, and *pchar*. The network path that we performed the following measurements on is from *orion* to *sirius*, two hosts that are located at a CS lab at Univ-Delaware. Both machines have Fast Ethernet network interfaces (100Mbps), and they are connected through an HP4000 L2 Fast Ethernet switch. Both L2 segments are Fast Ethernet interfaces, and so the capacity of the first hop is $C^{L3}_1 = C^{L2}_{1,1} = C^{L2}_{1,2}$=100Mbps.

Tool	Capacity estimate
pathchar	49.0±1.5Mbps
clink	47.5±1.0Mbps
pchar	47.0±1.0Mbps
Nominal capacity	100.0 Mbps

Table shows the capacity estimates for this path from the three VPS tools. The tools fail to measure the nominal capacity of this path, due to the presence of the Fast Ethernet L2 switch. According to our previous model, the (incorrect) capacity that a VPS tool would estimate in this path is $\hat{C}^{L3}_1$ = 100/2 = 50Mbps (see Equation 10). This theoretical value is actually quite close to the values that *pathchar*, *clink*, and *pchar* produce. For additional experimental results, please see [7].

The effect of cut-through L2 devices, as well as the effect of segmentation-based switches (such as ATM), is investigated in the extended version of this paper [7]. In that paper, we also investigate a few other possible sources of error, such as error propagation along the path, limited clock resolution at the probing host, queueing delays, and variability in the generation delays of ICMP messages.

REFERENCES

[1] V. Jacobson, "Pathchar: A Tool to Infer Characteristics of Internet Paths," ftp://ftp.ee.lbl.gov/pathchar/, Apr. 1997.

[2] A.B. Downey, "Using Pathchar to Estimate Internet Link Characteristics," in *Proceedings of ACM SIGCOMM*, Sept. 1999.

[3] B. A. Mah, "pchar: a Tool for Measuring Internet Path Characteristics," http://www.employees.org/∼bmah/Software/pchar/, Feb. 1999.

[4] K. Lai and M.Baker, "Measuring Link Bandwidths Using a Deterministic Model of Packet Delay," in *Proceedings of ACM SIGCOMM*, Sept. 2000.

[5] J. Postel, *Internet Control Message Protocol*, Sept. 1981, IETF RFC 792.

[6] V. Jacobson, "Traceroute: A Tool to Show The Route Packets Take to a Network Host," ftp://ftp.ee.lbl.gov/.

[7] R. S. Prasad, C. Dovrolis, and B. A. Mah, "The Effect of Layer-2 Store-and-Forward Devices on Per-Hop Capacity Estimation," Tech. Rep., College of Computing, GATech, 2002.