Notes on the Holt-Winters Procedure

Sally Floyd

October 4, 1993

1 Holt-Winters

The RED gateway uses a simple EWMA procedure when calculating the average queue size. In this section, we discuss the Holt-Winters procedure, a variant of the EWMA procedure, that responds somewhat more quickly than the EWMA procedure to a sustained increase in the queue. For our purposes, the Holt-Winters procedure is a modest but not essential improvement to the EWMA procedure.

The Holt-Winters procedure calculates the average slope sl of the average queue size as well as the average queue size ave itself [H89, p.27]. Each estimate of the average queue size incorporates the last estimate for the slope. The Holt-Winters procedure is as follows:

$ave_{old} = ave$
$ave \leftarrow (1 - w_q) * (ave + sl) + w_q * q$
$sl \leftarrow (1 - w_q/2) * sl + (w_q/2) * (ave - ave_{old})$

The averaging procedure must balance the goals of filtering out short bursts in the queue size and responding reasonably promptly to sustained congestion. Figure 2 compares the EWMA and the Holt-Winters procedures. The network, shown in Figure 1, consists of a fast line feeding into a slower line, with one connection. The connection has a maximum window of 120, somewhat larger than the pipe size of 82 packets. Figure 2 shows the queue at the gateway.

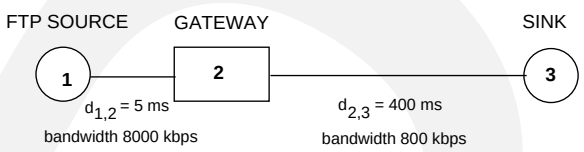

Figure 1: Network with one connection.

In Figure 2, the solid line shows the queue, line #1 shows the EWMA average, and line #2 shows the Holt-Winters average. The top figure shows both averages with $w_q = 0.001$, and the bottom figure shows both averages with $w_q = 0.002$. The x-axis shows time, and the y-axis shows the queue size. During the 'slow-start' phase of TCP, the window doubles in each roundtrip time, increasing from 1 packet to 2, 4, 8, 16, 32, and 60 packets respectively. The *bottleneck time* is the time required by the gateway to transmit a packet to the sink. During each slow-start window increase, two packets arrive at the gateway during each bottleneck time, resulting in a transient queue. For example, when the window increases from 16 to 32 packets, the 32 packets arrive at the gateway spaced over 16 bottleneck times. The result is a temporary queue of 16 packets. After the slow-start phase, the window increases by roughly one packet each roundtrip time, and a significant queue does not form again until the window exceeds the pipe size in packets. At this point the queue increases slowly

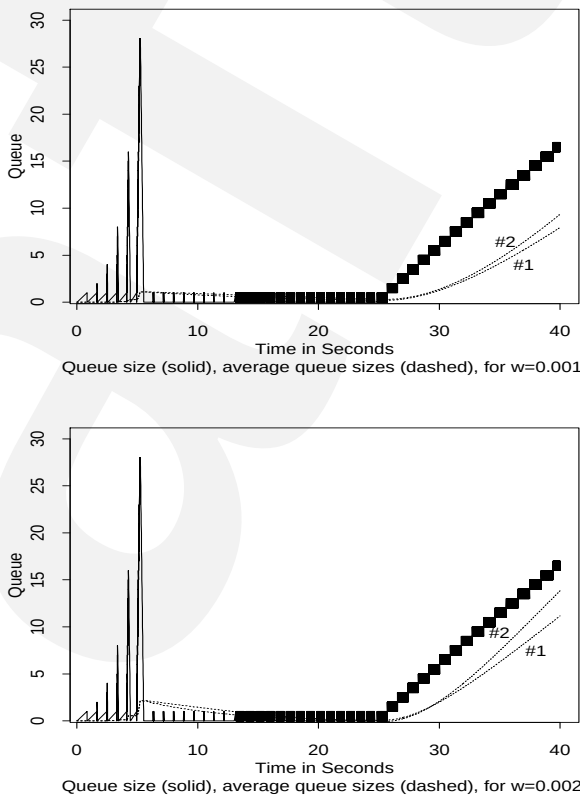

Figure 2: The queue and the calculated average queue size, for $w_q = 0.001, 0.002$.

Draft

Draft

but steadily. The initial transient bursts in the queue result in a small increase in the average queue size, while later in the simulation the steady increase in the queue results in a larger increase in the average queue size. The EWMA and the Holt-Winters averages have a similar performance in filtering short bursts in the queue. The Holt-Winters procedure performs slightly better than the EWMA procedure in responding to the slow but steady increase in the queue at the end of the simulation.

The advantage of the Holt-Winters procedure over the EWMA procedure is more evident when there is a sudden large increase in the queue size, as in Figure 3. Figure 3 shows a queue that increases from empty to 50 packets, maintains a queue of 50 packets over 900 packet arrivals, and then returns to an empty queue. The x-axis shows the packet number n and the y-axis shows the queue size. The solid line shows the queue size when the nth packet arrives at the gateway. The three dashed lines #1, #2, and #3 show the average queue size calculated by the EWMA procedure with $w_q = 0.001$, 0.002, and 0.003, respectively. The dashed line #4 shows the average queue size calculated using the Holt-Winters procedure with $w_q = 0.002$. For $n \leq 100$, the average calculated with EWMA with $w_q = 0.002$ and the average calculated with Holt-Winters with $w_q = 0.002$ are similar. However, for $n \geq 100$, the Holt-Winters average increases significantly faster that the EWMA average with the same parameter w_q. The Holt-Winters average overshoots the actual queue size of 50 packets somewhat. However, if the actual queue size had remained constant at 50 packets, all four procedures would converge to an average to 50 packets.

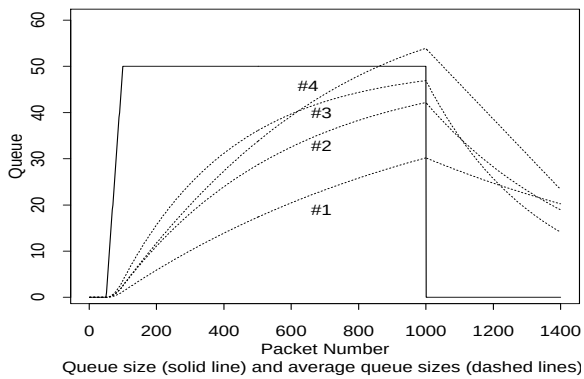

Figure 3: EWMA and Holt-Winters queue averages.

One detail concerns the calculation of the average queue size after the queue has been empty for m "packet arrival times". Assume that the average queue size is a_0 and the average slope is s_0 when the queue becomes empty. Assume that a packet first arrives again after m "packet arrival times". If the EWMA procedure is used, then the average queue size a_m at this point is

$$(1 - w_q)^m * a_0.$$

A simple approximation, when the Holt-Winters procedure is used, is simply to set the average queue size to $(1 - w_q)^m * a_0$ and to set the slope to 0 after the queue has been empty for m "packet arrival times".

References

[H89] Harvey, A., Forecasting, structural time series models and the Kalman filter, Cambridge University Press, 1989.