1 Introduction

A fundamental aspect of protocols that operate in TCP/IP networks is the different congestion control mechanisms utilized throughout the network. Various forms of congestion avoidance algorithms designed to prevent congestion have been proposed in an effort to improve end-to-end performance. The addition of slow-start and congestion avoidance to TCP in the late 1980's was crucial to ensuring the stability of the Internet. More sophisticated end-to-end congestion avoidance algorithms that monitor round trip time (RTT) and use increases in RTT samples as an indication of congestion have been proposed (e.g., TCP/Vegas, TCP/Dual) [BRAK94, WANG92]. The benefits claimed by end-to-end congestion avoidance algorithms are increased TCP throughput (i.e., due to fewer retransmissions and timeouts) and lower network buffer usage. Another claimed advantage of algorithms such as TCP/Vegas and TCP/Dual is that they do not require anything from the network and consequently could be incrementally deployed in today's best effort Internet.

In this dissertation, we study the performance of a class of TCP end-to-end congestion avoidance algorithms which uses an increase in a packet RTT as an indicator of congestion and future packet loss. We refer to this class of algorithm as *delay-based end-to-end congestion avoidance* (DCA). The following describes the attributes of DCA:

- DCA augments the TCP/Reno protocol.
- DCA monitors TCP packet round trip times allowing the algorithm to reside entirely within the TCP sender.
- DCA assumes that RTT variations are due to changes in queueing delays experienced by the packet that is being timed.
- DCA makes congestion decisions based on the RTT variation and reduces the transmission rate by some percentage by adjusting the TCP congestion window (*cwnd*).

The importance of an *incrementally deployable* enhancement cannot be overstated. It can take years for Internet protocols to standardize and even longer for Internet standards to be widely deployed. An

incremental enhancement to TCP that meets the following requirements will have the best chance for deployment:

- It must improve the throughput of the TCP connection that employs the enhancement.
- It should not reduce the performance of other competing TCP flows on the same path where the "enhanced" TCP flow travels.
- Ideally it requires changes only to a TCP sender.

Furthermore, due to the economics associated with deploying change within an existing network, we assume that an incrementally deployable change must provide benefits with a single deployment. If this is not true, the chances of deployment drop drastically. Therefore, we study the benefits associated with the deployment of a single DCA connection. To be complete, we also show how the "global" benefits of DCA (i.e., reduced buffer requirements within the network) might scale as the percentage of DCA traffic increases. However, the focus of the dissertation is to study a single DCA deployment.

We limit the scope of our study of DCA to Internet environments where the lowest link capacity along the path is 10 mbps. Previous studies of DCA (i.e., [BRAK94,WANG92]) have focused on low speed Internet paths. A significant percentage of traffic that flows over the Internet consists of high speed flows. For lower speed flows, the bottleneck is generally the access link to the Internet (i.e., the first Internet Service Provider over the path). Technologies such as Asymmetric Digital Subscriber Line (i.e., ADSL) and Cable modems are quickly driving these traditionally slower links to higher speeds [ADSL, CABLE].

We assume that a DCA flow will consume only a fraction of the total bandwidth. [THOM97] confirms that the Internet is dominated by HTTP traffic. [CROV96] shows that Web transfers are generally small (15Kbytes or less although some transfers will be much larger) with significant wait times in between document retrievals. A measurement analysis that studied the traffic arrival characteristics of a high volume web server found that 60% of TCP connections used a maximum window (i.e., the TCP receiver's maximum advertised window) of 12Kbytes [BALA96]. Furthermore, only 14% of the observed flows were limited by the maximum window. Finally, recent measurement studies indicate that Internet

backbone switches are subject to thousands of active TCP flows at any time [CLAF97, THOM97]. These findings confirm the widely held belief that Internet traffic is dominated by many low bandwidth, ON/OFF TCP flows.

The practicality and benefits of DCA have not been studied in high speed Internet environments. Recent measurement studies of the Internet suggest that the level of congestion (i.e., packet delays and packet loss) is increasing [PAXS97]. An incrementally deployable enhancement that can improve end-to-end TCP performance without increasing the load placed on the network is highly desirable. Our work shows that DCA cannot achieve this objective.

The thesis of this dissertation is to show that RTT-based congestion avoidance cannot be reliably incrementally deployed over high speed Internet paths. There are two reasons for this:

- A TCP sender that is extended with DCA will generally result in degraded throughput primarily because the congestion information contained in RTT samples cannot be reliably used to predict packet loss. We utilize a method based on analysis of TCP traces to show that:
  - ♦ A TCP constrained RTT congestion probe is too coarse to accurately track the bursty congestion associated with packet loss over high speed paths.
  - ♦ A DCA algorithm cannot differentiate between the significant number of increases in RTT that are not associated with packet loss from those increases that do lead to packet loss.
- By design, a DCA algorithm will limit the number of packets allowed in the network during periods of congestion (as compared to a TCP/Reno connection). If we assume that the level of traffic generated by a single TCP connection over a high speed path represents a small fraction of the total traffic that flows over the path, the reaction of a single DCA flow will have minimal impact the congestion level over the path.

We claim, therefore, that there are no significant benefits gained by incrementally deploying a DCA algorithm over high speed Internet paths. If a DCA algorithm reacts to the frequent RTT fluctuations and the reactions are not able to significantly reduce the loss rate, the connection will experience degraded

throughput as compared to a comparable TCP/Reno connection. Over a high speed Internet path, where routers are subject to thousands of active flows at any given time [CLAF97,THOM97], if a DCA flow reacts to an increase in RTT by reducing the TCP congestion window by half, assuming that demand for network buffers exceeds capacity, the buffers that would have been consumed by the flow over the next several RTT's will be consumed by other competing flows who are not as responsive as the DCA flow. In other words, the queue levels will continue to rise and packet loss will still occur regardless of the DCA reaction. Therefore, future increases in RTT are not affected by DCA congestion reactions.

We validate these ideas through extensive measurement and simulation. Below, we discuss the summary of our experiments and their findings.

Measurement Analysis:

We are interested in the relationship that exists between packet loss events and increases in the per packet round trip time samples that precede each loss. Using the *tcpdump* IP packet trace tool [JACO89], we trace TCP connections between a host located at North Carolina State University and a set of hosts over the Internet connected by high speed links. We extract a per-packet RTT time series from the traces. Encoded within each sample of the time series is a loss event indication which flags the RTT samples that precede the transmission of a dropped segment. We refer to this data as the *tcpRTT* time series. By examining the values of the *tcpRTT* samples that immediately precede the transmission of a dropped segment (which we call a "*loss conditioned*" delay), we can assess the level of correlation that exists between packet loss and increases in RTT samples.

We present a set of metrics that helps to assess and quantify the level of correlation between increases in *tcpRTT* samples and packet loss events that exists for a given connection. Our results are based on data obtained by tracing TCP connections over seven high speed Internet paths taken over the course of five days. The statistical results of the metrics applied to the aggregate data provide the following results:

- While a TCP constrained per-packet, RTT probe generally can detect long term congestion, it is able to detect the RTT increases that precede packet loss only 30-50% of the time. Furthermore, the

percentage of loss events that are preceded by an increase in RTT that exceeds a moving window average of previous RTT samples by a standard deviation is (on average) 7-18%. The implication is that it is very difficult for an end-to-end congestion probe to detect the queue buildup associated with packet loss over high speed Internet paths.

- For the loss events that are accompanied by an increase in RTT, the magnitude of the RTT increase associated with loss is not easily distinguishable from other more frequent increases not associated with packet loss. The implication is that in its efforts to avoid packet loss, a DCA algorithm will be reacting frequently to increases in RTT that are not associated with packet loss.

When operating over high speed Internet paths, the findings above prevent a DCA algorithm from improving the TCP throughput. To demonstrate this, we run a hypothetical DCA algorithm on the *tcpRTT* time series extracted from the TCP traces. We modify an analytic TCP throughput model [PADH98] to measure the throughput of the DCA algorithm under the same transmission environment. We show that if DCA were subject to the congestion dynamics reflected in the traced TCP connections, it leads to reduced throughput as compared to a similar TCP/Reno connection. We show that this is true even if the DCA algorithm is able to avoid a high percentage of the actual packet loss events.

This measurement analysis described above relies on the following conjectures:

- The measurement analysis assumes that the reaction of a single DCA flow to an increase in RTT does not significantly impact the congestion process that is active over the path. This assumption allows us to run a DCA algorithm and the analytical model on the RTT time series extracted from a TCP/Reno trace.
- The measurement analysis is limited to a specific algorithm that assumes a DCA congestion reaction that is equivalent to a TCP reaction to packet loss (i.e., a 50% window reduction). We believe that the results hold for less aggressive reactions to congestion and also for other variations to the algorithm.

We validate the above conjectures through simulation discussed below.

Simulation Analysis:

Using the *ns* simulation package [NSSIM], we develop a series of simulation models that emulate the characteristics of two of the Internet paths that we used for the measurement analysis. Using these models, we perform similar experiments as we did in our measurement work. An experiment consists of running a TCP/Reno connection over the simulated Internet path with a particular level of background traffic. Using the same metrics that we used in the measurement analysis, we show that the end-to-end dynamics associated with the simulated paths are similar to the measured results. Then, we provide additional simulation analysis that provides a deeper understanding of why a DCA algorithm cannot reliably avoid packet loss.

We develop a simulation model of the hypothetical DCA algorithm used in the measurement analysis and validate our assumption that a DCA reaction has minimal impact on the congestion processes that are active over the path. The statistical results of multiple runs over the simulated Internet paths indicate if we replace an end-to-end TCP/Reno flow with a version of TCP/Reno that is enhanced with DCA (we refer to this as TCP/DCA) the congestion dynamics at the bottleneck link are roughly the same. We also perform simulation runs with modified versions of the TCP/DCA algorithm and show that our thesis holds across a range of DCA algorithm design choices. Finally we evaluate the TCP/Vegas and TCP/Dual algorithms using the high speed Internet path simulation models and confirm that our findings apply to these DCA algorithms.

We stress that the focus of the thesis is to assess the effectiveness of DCA as an *incrementally deployable* improvement to TCP. Therefore, our analysis evaluates the performance of DCA when competing traffic consists primarily of most common TCP flows. It seems reasonable to assume that DCA will consume fewer buffers within the network during times of congestion compared to a similar TCP/Reno connection. The "global" benefits associated with this might become more apparent as the percentage of DCA traffic flowing through the congested routers increases. To confirm this, based on limited simulation experiments, we find that the loss rates at a bottleneck do not significantly drop until the percentage of DCA traffic exceeds 10%. Aside from this experiment, we focus our research on a single DCA deployment.

This dissertation consists of introductory sections, two main parts that constitutes the research contribution and concluding sections. The introductory material consists of a background chapter and a related research chapter. The first part of the thesis presents measurement and subsequent analysis of TCP flows over high speed Internet paths. The second part presents simulation analysis that supports the measurement work. We end the thesis with a chapter summarizing the key conclusions and contributions of this work followed by a chapter identifying future work.