Characterizing Nature and Location of Congestion on the Public Internet

Zehra Cataltepe and Prat Moghe
StreamCenter Inc.
571 Central Ave.
Murray Hill, NJ 07974
zehra@cataltepe.com, pratmoghe@yahoo.com

Abstract

We address the following question in this study: What is the nature of Internet congestion and where does congestion really occur on a Public Internet path? Answering this question will help service providers and content providers better engineer emerging services on the Internet. Our large-scale path measurement and analysis study indicates that congestion on the Internet exhibits a wide variety of packet loss and delay characteristics. Based on our classification using "congestion signatures", we find four dominant "types" of congestion which may be related to macroscopic behavior. A particularly frequent type of congestion we observe, is "flash congestion", which creates significant bursty packet loss on a fairly long time-scale.

Additionally, our study suggests that flash congestion predominantly occurs at the access provider network within the "last mile" on an Internet path. The Internet "cloud" does not contribute heavily to congestion. Consequently, prevalent approaches to bypass the cloud using "edge-based" content delivery networks and caching may not be effective in reducing congestion.

1. Introduction

A typical Internet path is made up of at least three networks: the hosting (or originating) network, the backbone & peering network, and the access ISP network. Commonly, these networks are referred to as the "first mile", "middle mile", and "last mile" networks. The term, Internet "cloud", usually refers to the combination of the first and middle mile networks. In this work, we are interested in understanding better the phenomenon of end-to-end "Internet congestion" that a typical user perceives. This understanding can benefit Service Providers, so that they may engineer economically without sacrificing high performance. It can also help Content Providers determine how to offer services with acceptable end-to-end quality.

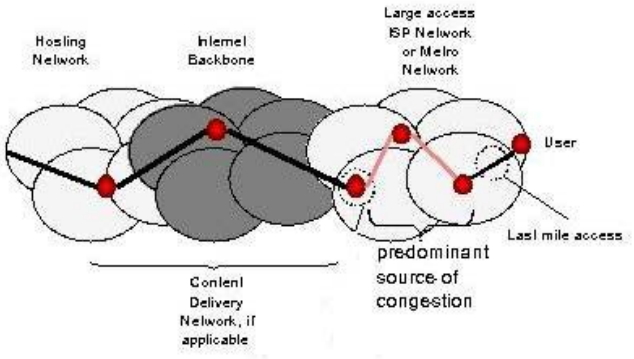

Figure 1. A typical Internet path

Our monitoring and analysis study at StreamCenter was initiated by the need to understand why "video" falters so badly on the public Internet. However, we believe our finding has significantly broader implications, across many applications, service providers, networks, protocols, and services. Based on the data collected and analyzed from a large-scale national monitoring effort (100 narrow-band monitoring clients were used to collect path-level congestion data from several hosting centers between 2000 to 2001), we find several interesting observations:

Nature of Congestion: We find that Internet congestion exhibits a wide variety of packet loss and delay variations. To better understand these variations, we define "congestion signatures", which are statistical groupings of packet loss and delay variations. Our data indicates that Internet congestion can be classified into four "types" of congestion signatures, based on the nature of degradation they create. One particularly dominant type of congestion is "flash congestion", which creates high levels of packet loss, with bursts lasting several tens of seconds. Flash congestion is harmful to most applications, but it is especially detrimental to reliable viewing of video and audio streaming, where

Proceedings of the Eighth IEEE International Symposium on Computers and Communication (ISCC'03)
1530-1346/03 $17.00 © 2003 IEEE

it causes frequent interruptions.

Location of Congestion: We find that the majority of the flash congestion occurs in the access ISP network (i.e., the last mile network). The actual extent of congestion in the access ISP depends on the specific networks, but we find that averaged over large number of paths and networks that we monitor, approximately 80% of severe congestion occurs within the access ISP network. The access ISP network is typically made up of the last mile access and two or three highly shared network hops.

The implications of our observations are interesting. First, the prevalence of "flash congestion", a serious form of congestion, within the access networks, suggests that this phenomenon is economically driven, and consequently, is here to stay. Access ISP networks are operated at significantly higher utilization levels than other networks to be able to recoup higher bandwidth and operating costs [11]. Higher utilization leads to higher sharing and higher congestion. In fact, we expect to see this problem eventually appear and become even more serious on the broadband access side, since broadband sessions incur significantly higher bandwidth costs relative to the revenue increase on the access side. Second, we believe new technological approaches are required to address the congestion problem in the access networks. Today, several technological solutions address the middle and first mile problems, such as CDN's, managed peering, and caching, however, they do not address the dominant source of congestion that is observed to occur in the access ISP.

2. Prior work

Internet congestion has been the subject of several important investigations into the nature of Internet traffic. [5] and [4] analyzed the nature of internet delay. There have been other approaches to establish link-level traffic properties, notably [6][7][8] and [9]. [2] and [3] clearly established the non-Poisson behavior of wide-area traffic.

In this study, we choose to focus on the large-scale congestion behavior of Internet paths, particularly modeling types of congestion, and analyzing its hop-level characteristics. Additional key characteristics of this work include:

- Detailed characterization of a path, including narrowband last mile effects of slow dialup lines and top-tiered access ISPs. Considering that over 75% of Internet users are still behind narrowband dialups[10], we believe that characterizing these effects are important to understand the overall impact of congestion on network and user behavior. We also include the effect of specific hosting points at different Tier 1 data centers.

- Large scale collection and sampling. We sample the congestion behavior from a wide number of end-points (100 monitors) over the duration of a year. Over 2 million data samples were collected in all, and subsequently analyzed with statistical techniques. In this paper, we report on a sub-set of findings observed over 5 months (August to December 2000).

3. Congestion Monitoring Framework

To realistically capture the end user behavior, the study deployed 100 "client" monitoring PCs at dial-up points throughout the continental United States. We used dedicated dial-up last-mile access, which is the predominant form of Internet connectivity today. Client locations were selected based on geography and population. To realistically sample the behavior of "best-of-breed" delivery systems, content was hosted in three different locations at top tier hosting data-centers.

The data collection method worked as follows. At scheduled times, the clients (receivers) dialed out using specific access ISP and received test data from the servers (sources). The test data consisted of 10-minute constant rate UDP sessions to passively sample end-to-end network behavior. The clients measured congestion based on packet loss and delay on the delivery path. Periodically, clients repeated this procedure using different access ISPs to receive the test data via different Internet routes. At the beginning of each session, a traceroute is made from the host to the client so as to store the route for the session.

Overall, 375,351 test sessions were collected and analyzed between August and December 2000.

4. Congestion Metrics

In this section, we first define basic congestion parameters. We also introduce the concept of congestion "signatures", which help characterize congestion.

4.1. Congestion Parameter Definitions

We define three basic congestion parameters that are used throughout the study:

- *bottleneck bandwidth* is the sustained throughput of the slowest hop on an Internet path. We infer bottleneck bandwidth by sending back-to-back packets and measuring the spread between received packets. To reduce the error in this technique, we send multiple batches of back-to-back packets, with a spacing interval between batches. One of the received batches is selected so as to reduce the error. For narrowband dialup lines, bottleneck bandwidth is typically the bandwidth of the last mile link. Bottleneck bandwidth is an important parameter for such sessions, since we want

to distinguish between packet loss caused by exceeding bottleneck bandwidth (which is not congestion, but really because source overflows the bottleneck bandwidth) versus packet loss caused by congestion from competing users. In our study, we require sessions to restrict peak UDP source rate to less than 75% of the bottleneck bandwidth; this eliminates the possibility of causing packet loss due to source overflow.

- *packet loss* is sampled every second by the client and measured as a running average, typically over a duration of 100ms.
- *delay* is measured as the end-to-end transit delay of a packet, reported relative to the transit delay of the first packet delivered from the server to the client. At the end of the session, the delay values are shifted so that they have a minimum of 0 seconds.

4.2. Congestion Signatures

We now define "congestion signatures" that will help classify congestion characteristics. Congestion signatures are mathematical functions of the time-varying packet loss and delay measurements, designed to capture various types of congestion. The congestion signature of a session is defined to be a triple that includes:

- *Baseline packet loss (pb)* is the "sustained" or "nominal" value of packet loss in a session. Baseline packet loss is conceptually determined by distinguishing the boundary between "sustained" packet loss and "bursty" packet loss. Sustained packet loss creates steady levels of packet loss for long duration (over tens or hundreds of seconds), while bursty packet loss creates packet loss "hills" (The concept of packet loss hills is similar to loss episodes reported in [9]) that range from hundreds of milliseconds to tens of seconds.

  To algorithmically compute baseline packet loss, we use the following procedure: 1. Initially set baseline packet loss to the mode of the packet loss process. 2. Define "sustained" loss to be any hill of duration 50 seconds or higher. 3. Incrementally increase baseline packet loss level until there is no sustained loss "hill" above it.
- *Area under the largest packet loss hill (pamax)* determines the largest amount of loss due to a hill during a session. The area of largest packet loss hill is the backlog in seconds of the amount of information that gets lost during one hill. This is an important construct, as we shall see, since it provides us a distinction between instantaneous congestion due to tiny bursts, versus flash congestion due to significant bursts.
- *Maximum delay (dhmax)* is the maximum of the relative transit delay values measured during the session, in seconds.

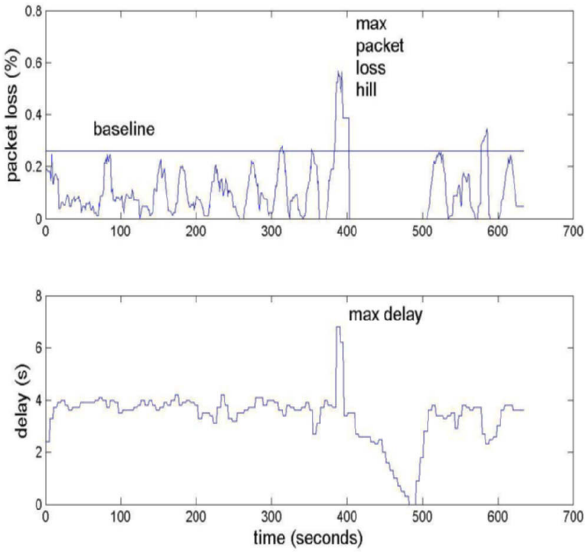

Figure 2. Signature example.

5. Characterizing Nature of Congestion

In this section, we characterize the nature of monitored congestion using the signatures defined earlier. We observe that four major behavioral "types" or "modes" of congestion are seen. We report on the relative incidence of each congestion type. The relative incidence of each type depends on the nature of traffic, utilization, the network design and applications on the specific path. We will demonstrate this variation for the Top 4 ISP providers within the US. Finally, we examine time of day effects on the overall probability of congestion.

5.1. Congestion Classification into Four Types

Based on the signatures described earlier, we classify congestion into four easily different types of congestion. The four types of congestion are summarized in the table below, along with their mathematical definitions. Wildcard (*) means the specific value is not significant to classification.

Congestion Type	pb	pamax (seconds)	dhmax (seconds)
Instantaneous Congestion	$\leq$0.05	$\leq$1	$\leq$10
Baseline Congestion	$>$ 0.05	*	*
Flash Congestion	$\leq$0.05	$>$ 1	$\leq$10
Spiky Delay	$\leq$0.05	*	$>$10

We believe that the four types described above capture congestion in its different modes. For instance, we believe that instantaneous congestion is caused by mild bursts, created naturally by burstiness of IP traffic. Baseline congestion appears to be caused by systematic under-engineering of network or hop capacity (or alternatively due to simple source overflow described earlier). Flash congestion suggests frequent but momentary periods of overload in a highly utilized network, where bursts from individual sources add up to create significant packet loss hills. Spiky delay is a condition where no packets are transferred for a long duration of time - the transit delay of packets shoots up from few milliseconds to tens of seconds during this period. At this point, we have inadequate intuitive explanation for what causes spiky delay, other than to surmise that it may be caused by router queueing policies in the access ISP network.

Figure 2 was an example of baseline congestion. Figures 3 and 4 are examples of flash congestion and instantaneous (i.e. no) congestion.

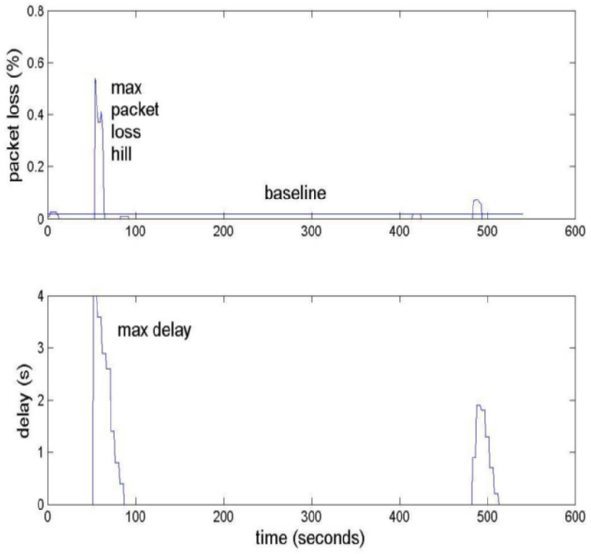

Figure 3. Flash congestion example.

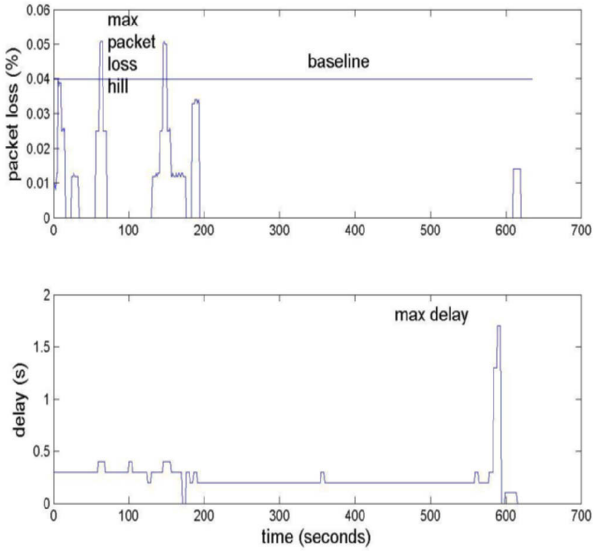

Figure 4. Instantaneous congestion example.

5.2. Characterizing Type of Congestion for Top ISP Providers

In this section, we demonstrate the relative incidence of the four congestion types described above. We will present results taken from monitoring congestion of a large number of sessions. Sessions are run between: 1. Servers hosted at three data-centers, in NJ, CA, and IL. 2. Clients accessing Internet via Top 4 ISPs within the US.

The sessions were run from August to December 2000. For each month, we computed the probability of congestion type as follows. First we partitioned the sessions in that month according to the time of day of the session. We used four time periods: 0-6am, 6-12am, 12-6pm and 6-12pm Next, we computed the probability of congestion in each time period. Finally, we used the average of the four time-periods to compute the congestion for each month, and then averaged the monthly results over 5 months.

Following is the overall probability of congestion and relative incidence of congestion types for each ISP:

Prob	[Cong]	[Fl.loss]	[Base.loss]	[Spi.delay]
ISP_A	0.34	0.20	0.11	0.03
ISP_B	0.36	0.15	0.16	0.05
ISP_C	0.16	0.04	0.03	0.09
ISP_D	0.16	0.06	0.02	0.08

As we see above, congestion types reveal significant insight into the impact of congestion. For instance above, ISP_A suffers from high incidence of flash congestion. Flash congestion is particularly harmful for applications that require steady, guaranteed throughput, such as

streamed video. While ISP_A and ISP_B exhibit similar overall incidence of congestion (36% versus 34%), the nature of congestion on ISP_B is quite different. The incidence of baseline loss is much higher on ISP_B. This can degrade most applications, but particularly reliable transaction-oriented and bulk transfer applications. Finally, ISP_C and ISP_D have low overall congestion incidence, but high incidence of spiky delay. Spiky delay can degrade all applications.

5.3. Time of Day Variation of Congestion

We also examined our data to understand how time of day affects the overall level of congestion. We partitioned each day into four time periods: 0-6am, 6-12am, 12-6pm and 6-12pm. For each month, we computed the probability that a session would be congested in each time slice. Then we took the average of these monthly probabilities for each time slice.

Following table indicates the probability of congestion for each time period and ISP. It appears that for all isps, 6-12am was the least congested time of the day. For most ISPs, it also appears that 6-12pm is the most congested time of day.

Time Period	0-6am	6-12am	12-6pm	6-12pm
ISP_A	0.34	0.31	0.32	0.37
ISP_B	0.38	0.22	0.38	0.45
ISP_C	0.21	0.09	0.14	0.20
ISP_D	0.11	0.09	0.15	0.21

6. Characterizing Location of Congestion

As pointed out earlier, a typical Internet end-to-end path consists of three major networks: the hosting network, the backbone & peering network and the access ISP network. We are now interested in understanding "where" congestion occurs on the Internet path, or in other words, how is congestion distributed between the networks on the Internet path. The focus of our attention will be the access ISP network. We will formulate a metric that measures the "fraction" of path-level congestion that occurs within the access ISP network based exclusively on path-level measurements. We will refer to this metric as "extent" of access ISP congestion. If extent of access ISP congestion is "X", that means, the access ISP contributes to X% of the path-level congestion.

6.1. Formulation of Extent Metric

The ISP addresses from the content host to the first ISP address form the *cloud network*. The ISP addresses from the second ISP address to the last ISP address form the access ISP network. In figure 5, R_0 is the cloud network shared by two clients. R_1 and R_2 are the access ISP networks for clients 1 and 2 respectively.

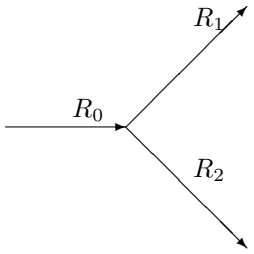

Figure 5. Cloud network R_0 and access ISP networks R_1 and R_2.

Let a_i be flash congestion (i.e. pamax) of a session on path i, $i = 0, 1, 2$. Let a_{0i} be congestion measured for a session at client i, $i = 1, 2$. We assume that $a_{0i} = a_0 + a_i$, which amounts to saying that no two packet loss hills on R_0 and R_1 or R_2 coincide.

Let $a_{01} \geq A^*$ for some $A^* \geq 0$ ($A^* = 1$, in our computations.) We define E_{01}, extent of congestion for a session with cloud network R_0 and client path R_1, as follows:

$$E_{01} = \frac{a_1}{a_1 + a_0} = \frac{a_1}{a_{01}} \quad (1)$$

The problem is that we cannot exactly measure E_{01}, since we cannot measure a_0 and a_1 individually. We can only assume measurement of a_{01}. Under the circumstances, we will formulate a lower-bound to E_{01}.

6.2. Lower-Bound to Extent Metric

We will now derive a lower bound to extent of access ISP congestion, using the concept of "divergent" sessions. Divergent sessions are sessions that share the same "cloud" network, but diverge in terms of the "access ISP" network paths.

Let there be two simultaneous divergent sessions, one that goes through cloud network R_0 and access ISP network R_1, and another that goes through cloud network R_0 and acccess ISP network R_2. Let their respective path-level congestion be a_{01} and a_{02}.

Let $a_{01} \geq a_{02}$. Now, by our assumption above, $a_{01} = a_0 + a_1$ and $a_{02} = a_0 + a_2$. We can use the two simultaneous divergent sessions to find a lower bound on the extent of congestion E_{01} as follows:

$$a_{01} - a_{02} = (a_0 + a_1) - (a_0 + a_2) = a_1 - a_2 \leq a_1 \quad (2)$$

Therefore we define $\hat{E_{01,2}}$, the lower bound for E_{01}

computed using second divergent session as follows:

$$E_{01} = \frac{a_1}{a_{01}} \geq \frac{a_{01} - a_{02}}{a_{01}} = \hat{E_{01,2}} \quad (3)$$

While this analysis assumed "simultaneous" divergent sessions, in reality, we do not expect to be able to find statistically significant instances of simultaneous divergent sessions. In practice, we will relax the requirement of simultaneity to require divergent sessions to be within the same "time of day". The lower bound will now be the relative difference between the average behaviour of divergent sessions, across a single time of day. The mathematical formulation of the lower bound follows logic similar to the formulation above.

6.3. Results: Extent of Access ISP congestion

We now report on the extent of access ISP congestion using the lower bound described above. We compare the relative difference between the time-averaged flash congestion metrics observed by divergent client sessions across the same time of day. The data is based on 1763 sessions that were delivered to 30 clients across 16 states over a period of four months. We find that the lower bound on average extent of congestion in the access network is 83% with an error of $\pm$-6. This means that congestion within the access ISP network dominates the path-level congestion.

7. Conclusions

We report on the findings from a large-scale monitoring study, intended to improve the understanding of congestion as it occurs on the public Internet. The focus of our work was to characterize the "nature" of congestion, and better isolate its "location" within end-to-end Internet paths. We introduced "congestion signatures" to classify congestion during a session.

8. Acknowledgements

This work has been based on efforts of a multi-year, large team effort at StreamCenter. We gratefully acknowledge everyone involved, and specifically appreciate significant contribution by the following team leads: Joe Argiro, Natesh Bhargav, Ashish Haruray, Danny Johnson Kanchan Mhatre, Ravi Narayan, Meetul Patel Adam Porter, Sindhu Xirasagar, Dr. Gang Zhou. StreamCenter's effort was funded by private and Institutional venture investors. We acknowledge their support and extended backing. Finally, we appreciate feedback and detailed comments on our work from Dr. Tanju Cataltepe, Prof. Doug Comer, Dr. Vern Paxson and Dr. Binay Sugla.

References

[1] W. Leland, M. Taqqu, W. Willinger, and D. Wilson, On the self-similar Nature of Ethernet Traffic, IEEE/ACM Transactions on Networking, 2:1-15, 1994

[2] V. Paxson and S. Floyd, "Wide-area Traffic: The Failure of Poisson Modeling", IEEE/ACM Transactions on Networking, 3:226-244, 1995.

[3] K. Park, G. Kim, and M. Crovella, "On the Relationship Between File Sizes, Transport Protocols, and Self-similar Network Traffic", In the Proceedings of the IEEE International Conference on Network Protocols, 1996.

[4] Qiong Li, David Mills, "On the Long-Range Dependence of Packet Round-trip Delays in Internet", Proceedings of IEEE ICC'98, Vol 2, pp 1185-1191, Atlanta, 1998.

[5] David Mills, Internet Delay Experiments, ARPANET Working Group RFC, DDN Network Information Center, SRI International, Menlo Park, CA, Dec 1983. RFC-889.

[6] A. Adams, T. Bu, R. Caceres, N. Duffield, T. Friedman, J. Horowitz, F. Lo Presti, S. B. Moon, V. Paxson, and D. Towsley, "The Use of End-to-end Multicast Measurements for Characterizing Internal Network Behavior", IEEE Communications, 38(5), May 2000.

[7] R. Caceres, N.G. Duffield, J. Horowitz and D. Towsley. "Multicast-based inference of network-internal loss characteristics", IEEE Transactions on Information Theory, vol. 45, No. 7, pp. 2462 - 2480, Nov. 1999.

[8] Nina Taft, Supratik Bhattacharyya, Jorjeta G. Jetcheva and Christophe Diot. "Understanding Traffic Dynamics at a Backbone POP", Sprint Technical Report TR01-ATL-020201. February 2001.

[9] Y. Zhang, V. Paxson, and S. Shenker, "The Stationarity of Internet Path Properties: Routing, Loss, and Throughput", ACIRI Technical Report, May 2000.

[10] Morgan Stanley Dean Witter Internet Research, December 2000.

[11] Metro Networks, The New Hotspot, Infonetics Research, April 2001