

Origins of Microcongestion in an Access Router

Konstantina Papagiannaki[†], Darryl Veitch[‡], Nicolas Hohn[‡]
dina.papagiannaki@intel.com, dveitch@sprintlabs.com, n.hohn@ee.mu.oz.au^{*}

[†]: Intel Corporation, [‡]: University of Melbourne

Abstract. Using an authoritative data set from a fully instrumented router at the edge of a core network, packet delays through an access link are studied in detail. Three different root causes of delay are identified and discussed, related to: unequal link bandwidth; multiplexing across different input links; and traffic burstiness. A methodology is developed and metrics are defined to measure the relative impacts of these separate, though inter-related, factors. Conclusions are given regarding the dominant causes for our representative data set.

1 Introduction/Motivation

Recent studies have shown that backbone networks are highly over-provisioned, and so inflict very little loss or delay on packets traversing them. For example, despite core routers with output buffers capable of holding on the order of 1 second of data, delays rarely exceed millisecond levels [1]. When examined on fine time-scales however, during localised periods of congestion, or ‘microcongestion episodes’, delays can still reach levels which are of concern to core network providers bound by Service Level Agreements (SLAs).

Typically backbone networks are structured in a hierarchy, where link bandwidths decrease as one moves from the long haul links connecting different Points of Presence (PoPs) (currently OC-192), through those interconnecting core routers within a PoP (OC-48 to OC-192), down to access links connecting customers to access routers (OC-3, OC-12 or gigabit Ethernet). The access links, being closer to the edge of the network, are more interesting to study from the delay perspective for two reasons. First, the list of potential causes of delays in a network widens as we move toward the edge. Second, an access link is typically managed by the customer. SLAs therefore do not apply and the link may be run at higher load levels to lower costs, again increasing the potential for congestion.

The aim of this work is to examine in detail the causes of microcongestion episodes in an access router leading away from the core, with a particular emphasis on delays. Although a full separation is not possible, there are nonetheless different generic ‘causes’ or mechanisms of congestion in general, and delay in particular, which can be identified. Briefly, these are related to: i) Reduction in link bandwidth from core to access, ii) Multiplexing of multiple input streams,

^{*} This work was done when K. Papagiannaki, D. Veitch and N. Hohn were with the Sprint Advanced Technology Laboratories, in Burlingame, CA, USA.

iii) Degree and nature of burstiness of input traffic stream(s). To our knowledge a taxonomy of congestion on an access link (or indeed any link) along these lines has not been studied previously. In particular we seek to answer the question, “What is the dominant mechanism responsible for delays?”, in such a context. More generally, a knowledge of the relative importance of different causes of higher than usual delay, and their interactions, gives insight into how delays may evolve in the future, and not only for the access router we study here. Accordingly, one of our main contributions is a methodology and a set of metrics which can be used more generally to this end.

We first flesh out the mechanisms above in more detail in the next section. We then give a description of our data and experimental setup in section 3. We describe our results in section 4 and summarise in section 5.

2 Congestion Mechanisms

Fundamentally, all congestion is due to one thing – too much traffic. The different mechanisms above relate to different ways in which traffic can be built up or concentrated, resulting in a temporary shortage of resources in the router. To explain the mechanisms precisely, we must have a model of router operation, as it is the router which will multiplex traffic arriving from different high speed links, and deliver it (in this case) to the lower speed output link.

In recent work [2] we looked at the modelling question in fine detail, using the comprehensive data set described in the next section. More specifically, we studied the through-router delays suffered by packets destined for a given output interface in a modern store and forward router. A model was developed which consists of two parts: a fixed minimum delay $\Delta(L)$ dependent upon packet size L which models the front end of the router and the transmission across the switch fabric, and a FIFO queue which models the output buffer and serialisation. We showed that for a store and forward router where the output buffer is the bottleneck, predicted through-router delays follow the measured ones very precisely. We also showed that, as expected, the FIFO queue part of the model dominates the delay dynamics. We will use this model below both conceptually and to generate many of the actual results. We ignore option packets here, which can have much larger delays but which are very rare.

In the framework of the model, *microcongestion* can now be precisely understood as the statistics of delays suffered during *busy periods*, which are time intervals where the system is continuously busy, but idle to either side. Here by ‘system’ we mean a given output interface and the portion of the router, leading from the input interfaces, related to it. Note however that packets are deemed to arrive to the system only after they have *fully* arrived to one of the input interfaces involved. For an input link, we will use ‘busy period’ in a different but related sense, to refer to a train of back-to-back packets (corresponding to a busy period of the output link of the router upstream). We can now discuss the three mechanisms.

Bandwidth Reduction Clearly, in terms of average rate, the input link of rate μ^i could potentially overwhelm the output link of rate $\mu^o < \mu^i$. This does

not happen for our data over long time scales, but *locally* it can and does occur. The fundamental effect is that a packet of size p bytes, which has a width of p/μ^i seconds on the input wire, is stretched to p/μ^o seconds at the output. Thus, packets which are too close together at the input may be ‘pushed’ together and forced to queue at the output. In this way busy periods at the input can only worsen: individually they are all necessarily stretched, and they may also then meet and merge with others. Furthermore new busy periods (larger than just a single packet) can be created which did not exist before. This stretching effect also corresponds, clearly, to an increase in link utilisation, however it is the small scale effects, and the effect on delay, that we emphasise here. Depending on other factors, stretching can result in very little additional delay, or significant buildups.

Link Multiplexing For a given output link, input traffic will typically arrive from different traffic streams over different input links. This is particularly the case given the Equal Cost MultiPath (ECMP) routing currently deployed by network providers for load balancing purposes. Whether these streams are correlated or not, the superposition of multiple streams increases the packet ‘density’ at all scales and thereby encourages both the creation of busy periods at the output, and the inflation of existing ones. To first order this is simply an additive increase in utilisation level. Again however, the effect on delays could either be very small or significant, depending on other factors.

Burstiness It is well known that traffic is highly variable or bursty on many time-scales. The duration and amplitude (the highest degree of backlog reached) of busy periods will depend upon the details of the packet spacings at the input, which is another way of saying that it depends on the input burstiness. For example packets which are already highly clustered can more easily form busy periods via the bandwidth induced ‘stretching’ above. To put it in a different way, beyond the first order effect of utilisation level, effects at second order and above can have a huge impact on the delay process.

3 Experimental Setup

We made measurements on a fully instrumented access router inside the Sprint IP backbone network. Of its 6 links, 4 were destined to customers at OC-3 and OC-12 speeds, while the other 2 connected to 2 different backbone routers inside the same PoP at OC-48 rate.

The first 44 bytes of *every* packet seen on *every* link attached to the router were captured, together with a GPS synchronised timestamp accurate to at least $5\mu\text{s}$. Using the methodology proposed in [1], the packets common to any two links were identified, and from the timestamps the through-router delays (with minor exceptions) were determined for each packet.

With one exception, every link on the router had an average link utilisation below 50% and thus experienced low congestion, and in particular low delays (99.26% were below 0.5 ms). The exception, which we study in detail here, was an access link at OC-3 rate fed from the two OC-48 backbone links, where average utilisations measured over 5-minute intervals reached as high as 80%.

Busy periods on this link lasted up to 15 ms, and resulted in maximum through-router delays as high as 5 ms.

In this work we study 13 hours of data comprising more than 129,000 busy periods. The completeness of this data set, and the analysis methodology, allows us to measure in fine detail the evolution of busy periods both on the input links and in the router itself. We can therefore empirically answer essentially any question regarding the formation and composition of busy periods, or on the utilisation and delay aspects of congestion, that we wish. In queueing terminology we have full access to the entire sample path of both the input and output processes, and the queueing system itself. An example of the sample path of the queue workload at the output covering over 3 busy periods is given in Figure 1.

4 Results

As mentioned above, we know from our previous work how to accurately model delays across the monitored router. We therefore use this model to run ‘semi-experiments’ (see [3] for details of this concept) as needed, where virtual scenarios are explored using real input traffic data fed to the physical router model, to help us quantify the contributions of the three mechanisms.

The experiments take the following form. First, a ‘total’ traffic stream S_T is selected. It is fed through the model with output rate μ , and the locations and characteristics of all the resulting busy periods are recorded. Note that even in the case when S_T is the full set of measured traffic, we still must use the model to determine when busy periods begin and end, as we can only measure the system when packets arrive or depart, whereas the model operates in continuous time.

For a given busy period we denote its starting time by t_s , its *duration* by D , its *amplitude*, that is the maximum of the workload function (the largest of the delays $\{d_j\}$ suffered by any packet), by A , and let t_A be the time when it occurred. When we need to emphasise the dependence on the stream or the link bandwidth, we write $A(S_T, \mu)$ and so on.

We next select a substream S_S of traffic according to some criteria. We wish to know the extent to which the substream contributes to the busy periods of the total stream. We evaluate this by feeding the substream into the model, since the detailed timing of packet arrivals is crucial to their impact on busy period shape and amplitude. The focus remains on the busy periods of the total stream even though the substream has its own busy period structure. Specifically, for each busy period of S_T we will look at the contribution from S_S appearing in the interval $[t_s, t_A]$ during which it was building up to its maximum A . Exactly how to measure the contribution will vary depending upon the context.

It is in fact not possible in general to fully separate the congestion mechanisms, as the busy period behaviour is a result of a detailed interaction between all three. The extent to which separation is feasible will become apparent as the results unfold.

4.1 Reduction in Bandwidth

To fix ideas, we illustrate the first mechanism in Figure 1. The two ‘bar plots’ visualise the locations of busy periods on the OC-48 input link (lower bar), and

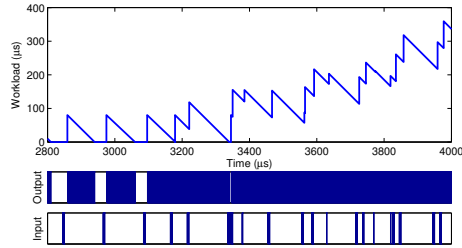


Fig. 1. The bandwidth reduction effect. Bottom Input/Output ‘bar’ plots: busy periods on the OC-48 input and OC-3 output links. Top: resulting queueing workload process. The output busy periods are longer and far fewer than at the input.

the resulting busy periods following transmission to the smaller OC-3 output link (upper bar). For the latter, we also graph the corresponding system workload induced by the packets arriving to the busy period, obtained using the model. We clearly see that the input busy periods - which consist typically of just one or two packets, have been stretched and merged into a much smaller number of much longer busy periods at the output.

In order to quantify the “stretching” effect we perform a virtual experiment where the total traffic stream S_T is just one of the main input OC-48 streams. By restricting to just a single input stream, we can study the effect of link bandwidth reduction without interference from multiplexing across links.

In this case our ‘sub-stream’ is the same as the total stream ($S_S = S_T$), but evaluated at a different link rate. We quantify “stretching and merging” using the normalised *amplification factor*

$$AF = \frac{A(S_T, \mu_o)}{\max_k A_k(S_T, \mu_i)} \frac{\mu_o}{\mu_i},$$

where $AF \geq 1$. The amplitude for the substream is evaluated across all k busy periods (or partial busy periods) that fall in $[t_s, t_A]$.

In simple cases where packets are well separated, so that all busy periods at both the input and output consist of just a single packet, then stretching is purely linear and $AF = 1$. If queueing occurs so that non-trivial busy periods form at the output, then $AF > 1$. The size of AF is an indication of the extent of the delay increase due to stretching. If the utilisation at the output exceeds 1 then theoretically it will grow without bound.

We present the cumulative distribution function for AF in Figure 2 for each of the main input streams separately. Less than 5% of the busy periods are in the ‘linear’ regime with minimal delay detected via $AF = 1$. The majority are significantly amplified by the non-linear merging of input busy periods into larger output busy periods. If instead we had found that in most cases that AF was close to 1, it would have been an indication that most of the input traffic on that link was shaped at OC-3 rate upstream.

To get a feeling for the size of the values reported in Figure 2, note that a realistic upper bound is given by $AF = 240000$, corresponding roughly to a 500ms

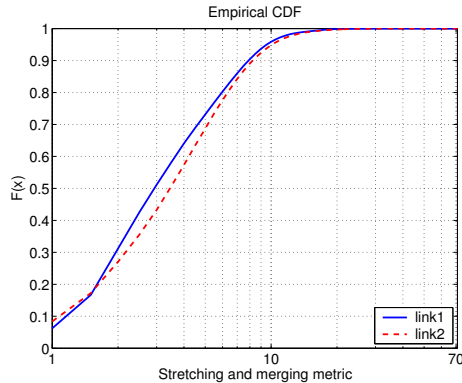


Fig. 2. Empirical distribution functions of AF for the OC-48 input streams.

buffer being filled (in a single busy period) by 40 byte packets well separated at the input, that would induce a maximum workload of $129 \mu\text{s}$ when served at OC-48 rate. A meaningful value worthy of concern is $\text{AF} = 1030$, corresponding to delays of 20ms built up from 375 byte packets, the average packet size in our data.

4.2 Link Multiplexing

To examine the impact of multiplexing across different input links, we let the total stream S_T be the full set of measured traffic. The rampup period, $[t_s, t_A]$, for two busy periods of S_T are shown as the topmost curves in Figures 3 and 4. We select our substreams to be the traffic from the two OC-48 backbone links, S_1 and S_2 . By looking at them separately, we again succeed in isolating multiplexing from the other two mechanisms in some sense. However, the actual impact of multiplexing is still intimately dependent on the ‘stretch transformed’ burstiness structure on the separate links. What will occur cannot be predicted without the aid of detailed traffic modelling. Instead, we will consider how to measure what *does* occur, and see what we find for our data.

Figures 3 and 4 show the delay behaviour (consisting of multiple busy periods) due to the separate substreams over the rampup period. The nonlinearity is striking: the workload function is much larger than the simple sum of the workload functions of the two input substreams, although they comprise virtually all of the total traffic. For example in Figure 3 the individual links each contribute less than 1ms of workload at worst. Nonetheless, the multiplexed traffic leads to a significantly longer ramp-up period that reaches more than 5ms of maximum workload at t_A on the right of the plot.

To quantify this effect we define the “link multiplexing” ratio

$$\text{LM} = \frac{\max_k A_k(S_i, \mu_o)}{A(S_T, \mu_o)},$$

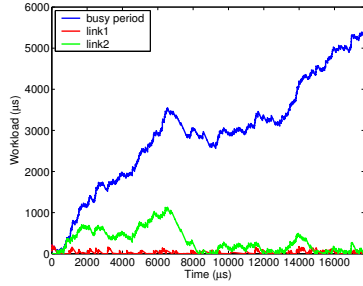


Fig. 3. Effect of multiplexing on the formation of busy periods (from t_s to t_A).

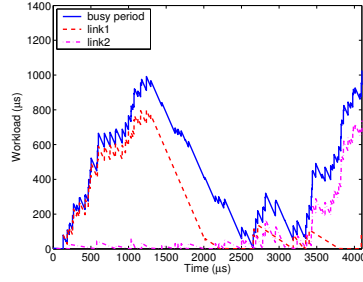


Fig. 4. A ‘bimodal’ busy period, assessing the contribution to A is ambiguous.

which obeys $LM \in [0, 1]$. Values close to zero indicate that the link has a negligible individual contribution. Therefore if all substreams have small values, the non-linear effect of multiplexing is very strong. In contrast, if $LM \approx 1$ for some link then it is largely generating the observed delays itself, and multiplexing *may* not be playing a major role. Large values of LM are in fact subject to ambiguity, as illustrated in Figure 4, where the ratio is large for both links. The busy period has a bimodal structure. The first mode is dominated by link 1, however its influence has died off at time t_A , and so is not significantly responsible for the size of A .

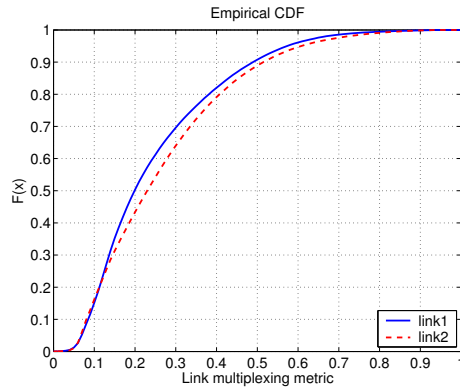


Fig. 5. Empirical distribution functions of LM for the OC-48 input streams.

The results for the data are presented in Figure 5. In more than 95% of the busy periods traffic from each individual link contributes to less than 60% of the actual busy period amplitude. Therefore, it appears that multiplexing is an important factor overall for the delays experienced over the access link.

4.3 Flow Burstiness

There are many definitions of burstiness. It is not possible to fully address this issue without entering into details which would require traffic models, which is beyond the scope of this paper. We therefore focus on burstiness related to 5-tuple flows, to investigate the impact that individual flows, or groups of flows, have on overall delays. We begin by letting the total traffic S_T be that from a single link, to avoid the complications induced by multiplexing.

In order to obtain insight into the impact of flow burstiness we first select as a substream the ‘worst’ individual flow in S_T in the simple sense of having the largest number of packets in the rampup period $[t_s, t_A]$. Two extreme examples of what can occur in the rampup period are given in Figures 6 and 7. In each case the busy period amplitude is large, however the flow contribution varies from minimal in Figure 6, to clearly dominant in Figure 7.

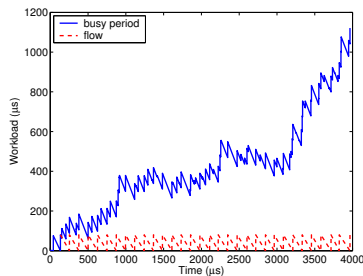


Fig. 6. Flow with multiple packets and no significant impact on the queue buildup.

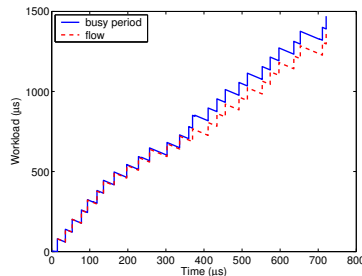


Fig. 7. Flow with multiple packets that dominates the busy period.

To refine the definition of worst flow and to quantify its impact we proceed as follows. For each busy period in the total stream we classify traffic into 5-tuple flows. We then use each individual flow S_j as a substream and measure the respective $A(S_j, \mu_o)$. The worst or “top” flow is the one with the largest individual contribution. We define “flow burstiness” as

$$\text{FB} = \max_j \frac{\max_k A_k(S_j, \mu_o)}{A(S_T, \mu_o)},$$

where as before the inner maximum is over all busy periods (or partial busy periods) of the relevant substream falling in $[t_s, t_A]$. The top flow may or may not be the one with the greatest number of packets.

Our top flow metric takes values $\text{FB} \in (0, 1]$. If FB is close to zero then we know that **all** flows have individually small contributions. Alternatively if FB is large then, similarly to LM, there is some ambiguity. We certainly know that the top flow contributes significant delay but in case of bimodality this flow may not actually be responsible for the peak defining the amplitude. In addition, knowledge about the top flow can say nothing about the other flows.

We present the cumulative distribution function for FB in Figure 8 for each

of the OC-48 links. For more than 90% of the busy periods the contribution of the top flow was less than 50%. In addition for 20% of the busy periods the contribution of the top flow was *minimal* (for example as in Figure 6), that is it was the smallest possible, corresponding to the system time of a single packet with size equal to the largest appearing in the flow.

If the top flow has little impact, it is natural to ask if perhaps a small number of top flows together could dominate. One approach would be to form a stream of the n largest flows in the sense of FB. However, as the choice of n is arbitrary, and it is computationally intensive to look over many values of n , we first change our definition to select a more appropriate substream. We define a flow to be *bursty* if its substream generates a packet delay which exceeds the minimal delay (as defined above) during the rampup period. Note that only very tame flows are not bursty by this definition! We denote by S_b the substream of S_T that corresponds to **all** bursty flows, and compute the new flow burstiness metric:

$$FB' = \frac{\max_k A_k(S_b, \mu_o)}{A(S_T, \mu_o)}.$$

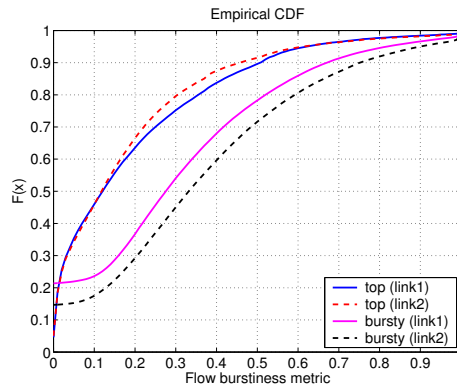


Fig. 8. Empirical distribution functions of FB' for the OC-48 input streams.

As before, $FB' \in [0, 1]$, and its value can be interpreted in an analogous way to before. The difference is that, as the substream is much larger in general, a small value is now extremely strong evidence that individual flows do not dominate. Note that it is possible that no flow is bursty, in which case $FB' = 0$, and therefore that the top flow is not necessarily bursty. This has the advantage of avoiding the classification of a flow as dominant, thereby giving the impression that it is bursty in some sense, simply because a trivial flow dominates a trivial busy period.

Our results are presented in Figure 8. As expected, the contribution of S_b to the busy period is more significant: in 20% of cases it exceeds 60%. On the other hand, 20% of the busy periods had FB' equal or close to zero, indicating

that they had no bursty flows. Indeed, we found that only 7.7% of all flows in our dataset were classified as “bursty” according to our definition. Only in a very small number of cases does the top or the subset of bursty flows account for the majority of the workload (for example as in Figure 7). Consequently, it seems that in today’s network flow dynamics have little impact on the delays experienced by packets in core networks.

5 Summary

We have studied in detail the origins of packet delays flowing toward an access link, and clarified the role of three different mechanisms, related to: unequal link bandwidth; multiplexing across different input links; and traffic burstiness.

Our first contribution was methodological. We showed how a router model can be used to investigate the role of the different mechanisms, and defined metrics to help assess their impact. The possible values of the metrics, and how they can be interpreted, was discussed.

Our second contribution was to investigate the actual contributions in today’s access networks, via a comprehensive and representative data set. We focused on an OC-3 access link fed mainly by two OC-48 links carrying roughly 50% of the traffic each. The link was not highly congested (no packet drops over 13 hours and packet delays all under 6ms), however it was much more congested than typical core links. We found that the link bandwidth reduction factor of 16 (from OC-48 to OC-3) played a significant role in delay buildups (non-trivial amplification factor AF), indicating that traffic is bursty, and not significantly shaped at OC-3 rates upstream. Multiplexing was also found to be significant (small multiplexing fraction LM), as in most cases the traffic on the individual links could not individually induce delays which were a large fraction of the observed delays. Finally the effect of individual 5-tuple flows, and sets of ‘bursty’ flows, was found to be small in most cases (small flow burstiness ratio FB), leading to the noteworthy conclusion that 5-tuple flow dynamics are not responsible for excessive packet delay in today’s core networks. These conclusions are strongly traffic dependent. The methodology and metrics we define can be used to monitor traffic evolution, and are especially effective when one wishes to confirm a hypothesis that a given effect (such as individual flow impact) is negligible.

References

1. Papagiannaki, K., Moon, S., Fraleigh, C., Thiran, P., Tobagi, F., Diot, C.: Analysis of measured single-hop delay from an operational backbone network. In: IEEE Infocom, New York (2002)
2. Hohn, N., Veitch, D., Papagiannaki, K., Diot, C.: Bridging router performance and queuing theory. In: Proceeding of ACM Sigmetrics Conference on the Measurement and Modeling of Computer Systems, New York, USA (2004)
3. Hohn, N., Veitch, D., Abry, P.: Cluster processes, a natural language for network traffic. IEEE Transactions on Signal Processing, special issue on “Signal Processing in Networking” **51** (2003)