Congestion balancing using re-ECN

João Taveira Araújo, Miguel Rio and George Pavlou

University College London

Abstract: The emergence of a metric for congestion exposure at the network layer has the potential to significantly impact networking without disrupting the current Internet architecture. Re-ECN [1], a protocol for adding accountability for congestion, has been proposed as a means of aligning incentives between economic stakeholders, securing networks against denial of service and enforcing fairness between users. Our work builds on the information contained in re-ECN packet markings to balance congestion, rather than load, across domains according to expected upstream congestion.

1 Introduction.

The economic nature of interdomain peering has resulted in an architecture where connectivity is assured in accordance to commercial contracts, but at the cost of suboptimal routing and with few attempts at exploring existing path diversity. Domains currently lack both information to route packets according to end-to-end performance metrics and the incentive to carry surplus traffic from customers.

Congestion exposure, as proposed by re-ECN [1], alleviates some of these concerns by including information on end-to-end congestion at the network layer. This allows domains to assert the costs incurred by customers and inflicted on providers, both under the guise of congestion marked packets. Using re-ECN as an explicit cost metric enables providers to explore pricing schemes which more accurately reflect costs of provisioning a network. As such, adding accountability for congestion is a means of progressing interdomain routing beyond current ad-hoc peering arrangements, towards a market model where transit domains compete to attract traffic.

Under such conditions, networks have an incentive to actively minimize congestion charges by balancing traffic over outbound links according to expected upstream congestion. Our work explores how such congestion balancing could be achieved in order to both reduce costs and enhance end-to-end performance.

2. Design Overview

Our preliminary work has focused on balancing congestion on edge domains, which we believe would be natural deployment points for congestion balancing. For one, edge domains are typically multihomed, as both ISPs and businesses increasingly interconnect to improve resilience. Additionally, such domains are not plagued by the scalability concerns of BGP routing tables, since they will often aggregate routes to a default provider. This allows some degree of freedom in manipulating routing without the need to propagate more specific routes downstream and increase BGP churn.

Assuming re-ECN capable traffic, boundary policers monitor both bulk congestion, for each outbound link, and averages of instantaneous upstream congestion for each prefix, allowing insight into how congestion evolves over time for each chosen path.

This data is periodically assembled and analysed, and the output of the decision process is then fed into the route reflector. The decision process itself is the subject of current research, and involves comparing upstream congestion consistently across all outbound links. If a discrepancy in these values is found, the decision process selects a destination network routed through the most congested link with a feasible alternate path and forks a monitor for a subset of the destination network through the use of a more specific prefix or a hash function.

A period of monitoring ensues in order to evaluate whether the sampled traffic is an effective measure of congestion for the announced network prefix, before shifting the traffic through an alternate egress link. Should congestion reduce significantly, we may wish to route the whole address block through the new path, but only if the level of correlation is sufficient to suggest congestion is inherent to a specific AS path.

What defines a discrepancy in upstream congestion, sufficient correlation and update intervals is conceivably a runtime decision. Our current work is focusing on providing guidelines so operators may tweak this decision process to achieve a desired outcome, as well as refining the overall framework.

3. Simulation Results.

A simple setup to illustrate congestion balancing is shown in Figure 1, and was used in simulating congestion balancing using the ns2 simulator [2], extended to support both BGP and re-ECN.

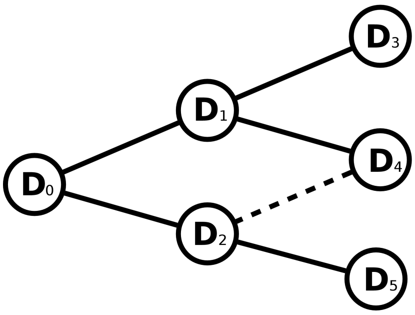

Figure 1: Simulation topology. Each node acts as a single domain. Highlighted linksshow AS paths for each domain from D_0 , as selected by BGP.

Each node, representing a domain D_i, i∈[0,5], announces a network N_i/24 to its immediate neighbours using BGP. Every link has a 10Mb capacity and a 25ms delay and every node uses RED [3] to mark packets. While we will be dealing with re-ECN capable traffic, no extension to ECN is required from routers to perform congestion marking.

Once BGP converges, a total of 30 TCP flows are established from within D_0 towards hosts in D_3, D_4 and D_5, with destination addresses spread evenly across both prefixes and host numbers.

Congestion balancing is exclusively performed within D_0, where an exponentially weighted average of upstream congestion for all network prefixes is maintained and updated every 10 seconds. This is the only network node that need be re-ECN aware.

The ensuing results are shown in Figure 2. Initially, D_0 uses the default routes as calculated by BGP. This results in higher congestion marking rate on flows to N_3/24 and N_4/24, as they both share the bottleneck link through D_1. Once our monitor detects a discrepancy in upstream congestion to outbound links, it looks for networks with alternative paths which are currently routed through the most congested link. Having selected N_4/24, which has two paths with an equal AS hop count, the monitor then applies a new entry to monitor network N_4, but with a larger prefix, thereby sampling the existing entry.

Once the correlation in congestion between N_4 /24 and N_4 /25 has been established, the route reflector triggers an update to the BGP routing process, forwarding N_4 /25 through D_2 at 150s.

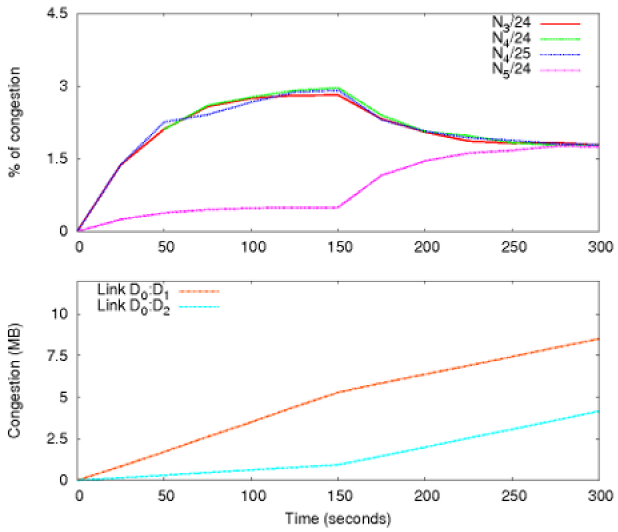

Figure 2: Upstream congestion estimate and congestion volume viewed from D_0. At t=150s, D_0 switches path for destination N_4/25 and through D_2.

A key question yet to be answered is to what extent dynamic congestion balancing can be performed. Congestion pertains to flows, and is therefore best dealt with at a scale of an average round trip time. This is neither feasible nor desirable however, as it may lead to both oscillations in routing and potentially significant packet reordering at the end-hosts. Striking a balance between the timescale in which networks should react to congestion and the subsequent disruption for flows is essential when considering the validity of congestion balancing.

Further ahead, there may be potential in extending this approach into the core in order to assist the edges in selecting an appropriate path. Re-ECN provides an easily verifiable claim of upstream congestion, and therefore may be used alongside interdomain protocols to not only advertise prefixes and paths, but also expected congestion.

4. Conclusions.

We have outlined a mechanism for minimizing bulk congestion across outbound links for edge domains. Balancing congestion is not only necessary to minimize costs when faced with congestion pricing, but, by harnessing information exposed by re-ECN, may also enable domains to improve end-to-end performance beyond myopic load balancing.

Acknowledgments.

This work was funded by the Portuguese government through FCT grant SFRH/BD/37855/2007

References.

[1] B. Briscoe, A. Jacquet, T. Moncaster, and A. Smith, “Re-ECN: Adding accountability for causing congestion to TCP/IP,” Internet Draft, Mar. 2009. (Work in progress).

[2] “The network simulator - ns2,” http://www.isi.edu/nsnam/ns/.

[3] S. Floyd and V. Jacobson, “Random Early Detection Gateways for Congestion Avoidance,” IEEE/ACM Trans. Netw., vol. 1, pp. 397–413, 1993.