BwE: Flexible, Hierarchical Bandwidth Allocation for WAN Distributed Computing– Public Review

Srikanth Kandula
Microsoft
srikanth@microsoft.com

This paper represents the next step in a series of recent papers on more holistic management of the bandwidth *between* geographically distributed datacenters.

Bandwidth on the WAN is an expensive resource. However, the applications using WAN bandwidth have varying amounts of rate and time elasticity. That is, some can make do with less bandwidth (e.g. serve video at a lower bit rate) and others can make do with some delay (e.g. transferring data for archival purposes). Hence, this line of work aims at global utility maximization while meeting application constraints.

The specific contributions in this work are three-fold. First, whereas prior work requires applications to specify a fixed demand, this paper allows specifying demand as a *(concave) function of bandwidth*. This allows an application to encode the fact that its marginal value from receiving a higher rate decreases after a point. Second, the paper offers an algorithm that takes the above demands as input and outputs a global max-min fair allocation that also allows for *hierarchical fairness*. That is, fairness can be defined among sites or clusters rather than per demand. Finally, the paper offers some details from an extensive deployment. Some interesting aspects include (a) scalability of the algorithm: the algorithm converges within 3s for their largest setup and (b) decoupling of the overall problem into a traffic engineering system called TE that is not described in this paper and the BwE system that is described in this paper which fairly distributes the rate allocated by TE.

The MPFA algorithm that handles concave demands and hierarchical fairness is nifty. Especially, the trick to compute the *bandwidth function* of a link by aggregating that of non-frozen demands (Section 5.3). It is also useful to learn about systems that have seen extensive deployment. A few questions could benefit from more exploration, however.

- Failure tolerance and worst-case behavior are perhaps the most significant concerns behind the practical use of such centralized traffic management solutions. The system not only has to be robust to failures in the software components and the network paths between these components but also has to move traffic rapidly in response to faults on the WAN network. It would be interesting to understand the uptime (or other reliability and failure responsiveness) statistics of this system as well as the techniques used to improve robustness.
- Decoupling rate allocation into two systems (TE and BwE) can lead to sub-optimal allocations relative to a single optimization that handles both aspects. It would be interesting to characterize this *gap*. Of course, the single optimization may not scale as well and so the gap may be a necessary evil to achieve scale. Characterizing the trade-off however would open up the possibility for other solutions in this space.
- Finally, it is apriori unclear how well the offered notions of hierarchies and rate elasticity match the needs of users. Do user's request fairness at some other granularity that does not naturally fit within a hierarchy? Are there demands that do not have concave utility functions such as (soft) deadlines on network transfers? Exploring use cases better could lead to a deeper understanding of this space.