Inside the Social Network's (Datacenter) Network – Public Review

Srikanth Kandula
Microsoft
srikanth@microsoft.com

Good design and management of datacenter networks is becoming increasingly important since it often translates to a competitive advantage in delivering services and analytics.

This paper presents a first look into the traffic characteristics of the workloads in datacenters operated by Facebook. In doing so, the paper extends the publicly available knowledge-base on production workloads. As well, it offers some novel and interesting measurements. First, the Hadoop clusters operated at Facebook exhibit substantially smaller amounts of rack locality than previously reported. That is, more than 80% of the traffic crosses *racks*. Second, in spite of the above lack of locality, the overall network link utilization is quite small– an average of less than 10% on all potential bottlenecks. Links that support Hadoop traffic are much more likely to have a higher load relative to other links in the datacenter. And, diurnal changes in request load from users accounts for only a modest variation in network traffic (about 2×). Third, the other dominant application at Facebook is memcached-style request-response workload which comprises primarily of small packets and has some specific traffic patterns. Finally, the paper posits lack-of-stability-in-the-traffic-matrix, poor-predictability-of-heavy-hitters and good-application-level-load-balance as reasons that make fine grained traffic engineering not needed or not easily doable.

In all, the results are fascinating and raise some interesting takeaways. The direct takeaway is that such holistic network-wide instrumentation is incredibly useful to have. The reasons behind why the traffic looks as it does are not as clear: specifically, whether the cross-rack traffic is because their cluster scheduler for Hadoop fails to achieve locality (i.e. place tasks within the same rack as their input) or something more to do with the nature of the jobs at Facebook (e.g. jobs resemble Terasort more than Wordcount in that shuffle≈input). The most fundamental perhaps is whether there exist datacenter workloads that are consistently limited by network bandwidth. This is clearly true across datacenters but the jury appears still out for *within* a datacenter. Further, at utilizations that are routinely below 10%, low latency is relatively easy to achieve for the *average* flow. The reasons, one suspects, are 10Gbps or more available network capacity at servers and datacenter switch interconnects that already have a low amount of over-subscription. Can applications be engineered better to drive up their network usage? Or are they limited by other resources? Lacking network-heavy applications, are we done with intra-datacenter networking as a problem? Of course, there will always be some application that crucially depends on the network, as there always have been, but the debate is whether a majority of datacenter workload exhibits such a profile.

Finally, a note of caution: traffic characteristics derive from the applications– Hadoop's traffic profile is different from that of memcached and that from enterprise applications. In some cases traffic from the same application may well depend on how the application is configured or used. For example, Terasort jobs pose more demands on the network compared to data exploration jobs especially when they run with a good query optimizer. It is our hope that this work would spur reports from other production workloads to broaden our understanding of datacenter traffic characteristics.