WOMBIR 2021
Critical Questions Submission

The importance and feasibility of traffic weighting on Internet performance analyses;
and
Planned approaches to obtain massively distributed edge performance observation

Avi Freedman, Kentik, avi@kentik.com
Alistair King, Kentik, alistair@kentik.com

Key Research Topic 1:
The importance and feasibility of traffic weighting on Internet performance analyses

First question: How different are performance rankings when weighted by traffic density? Would traffic-weighting such studies drive better operational, business, and policy conclusions and action?

A key recurring theme that we'd like to see become addressable is establishing as normal practice the traffic-weighted studies of performance, stability, security, and other Internet surveys and analytics.

We believe that conclusions about operational effects, to inform operational, engineering, and policy decisions can be materially improved by understanding how much traffic is affected by such performance or security issues or activity.

Practically, we have seen this in reviewing our own, and external, studies of network performance and BGP hijacking - often the worst performance problems or most active hijacking networks are affecting the least, or sometimes, no traffic. Being able to show customers this correlation has been of huge perceived value in our commercial offerings, and we are interested in finding ways to enable adding traffic matrices based on real Internet measurements, to be used in research addressing network design, management, and policy issues.

Specifically, we would focus analysis on understanding the relative amount of traffic affected by inter-network congestion measurements underway; and questions related to adding traffic to weighting (for example, relative weighting of broadband and wireless access providers in a given region) to policy-supporting analyses of whether regional infrastructure is sufficiently performant to support remote work or education.

Related extension questions might include tagging types of traffic such as educational access, educational platforms, e-commerce, payment platforms, gaming. While would enable different business and policy questions to be asked in a more nuanced way, our suggested first focus would be on overall traffic first before trying to get into more nuanced taxonomies.

We have sufficient data to conduct such traffic-weighted performance analyses on Kentik data, but one challenge in validation and reproducibility of any such analyses is that it would be difficult for others to access the same data set (particularly the traffic matrix data) except under restrictive NDA, unless we can solve for the next question.

Second question: Can we obtain sufficiently privacy-respecting aggregates of combined private traffic data across networks, either inter-AS matrices, or better validated matrix generation methods, to enable multiple use and sharing across studies (and in an open/reusable way across multiple projects)?

Ideally in our view would be the ability to share inter-AS matrices generated from many networks.Or if that is not possible, to at least start by testing methods of generating inter-AS matrices from network structure and/or data from portions of one or multiple participating networks (i.e. Network Syntax from Sanchez, et al or otherwise).

There would still remain a concern to be satisfied that tuning methods and models would not cause them to generate patterns that would violate privacy constraints of the network they were tuned against, but this seems a much lower hurdle than sharing more complete traffic data, even aggregated and/or redacted.

Areas of Concern for These Questions

- Data sharing policy - Traffic information is very sensitive, and many operators have explicit obligation to protect user and customer identification, both regulatory and contractual. While our network customers have expressed perceived value in multi-customer aggregated performance data, there is less incentive in this area.

  However, based on conversations, we believe there are potential contributors who would allow sharing geographical traffic matrix data to be aggregated at a very high level. We think that there is a somewhat larger subset of networks who would be comfortable sharing their view of other networks to be combined with many other views, as long as their view of their customers is aggregated or omitted and prefix-based aggregation occurs, and this is where we would start in a first phase. And even more that would be comfortable with their traffic data to be used to validate models. In many of our conversations, networks would want to see the results to validate that the aggregates don't reveal unique characteristics of their networks/customers, at least in early phases of work.

- Making knowledge useful - We are open to showing specific deltas between performance rankings, as weighted with and without traffic matrices we can use based on our data today. We believe this could help spark discussion and generate interest in this data as a way to help enrich a wide range of network-related performance, security, and policy research.

Ways the NSF Can Help

We believe that the NSF can help by sponsoring:

- Infrastructure - If there are other identified networks that would help for completeness and are willing, NSF sponsorship could be extended to open source tool development to observe, process, and transmit traffic matrices; and to test and validate generating models against actual traffic data. If some of those networks are willing to share network traffic metadata in more detailed form, NSF support would likely be helpful to support the infrastructure or subsidize commercial service for participants.

- Data sharing - We believe that it will be helpful to gain agreement to data sharing if we can show operators collaboration with researchers studying management, integrity, and automation (of great interest to commercial operators), and researchers involved with experience and activity in privacy-respecting aggregation and analysis methods.

  In addition to supporting such related research, NSF sponsorship of workshops focused on policy and technology for data sharing will be very helpful to bring together the operations, research, and government community (the latter to help discuss regulatory principles and concerns), and ultimately to establish and increase operator collaboration in sharing traffic data.

- Making knowledge useful - We are proposing to move beyond traffic matrix data for the purpose of validating and forming synthetic networks and analyses, and hopefully would be enabling a new source of telemetry to be used in a wide range of research. Participation in NSF-sponsored workshops that bring together the commercial as well as research world would be helpful.

Key Research Question Topic 2:
Planned approaches to obtain massively distributed edge performance observation

Questions: What are the performance characteristics of edge user access (broadband and mobile networks), as seen between those networks and major content and application delivery infrastructure? How frequent are performance impairments? How unevenly distributed is usable (consistently performant) access to work, school, and social infrastructure? These are not new questions, but our goal is to achieve tens of thousands or more observation points from which we can perform continuous measurement.

With a vast portion of economic, work, social, and collaborative activity now being performed to and through consumer-connecting networks (broadband, wireless, and satellite), understanding the performance, stability, and ability of users to access critical services is key. While home network and wireless is a critical component, as is the intra-AS network, understanding the end-end and often inter-network performance is critical to both operating and managing these delivery and access networks, and to operational, research, and policy analyses involving home and mobile networking. We operate 250 locations and growing for active performance measurement today, but the goal would represent orders of magnitude increase in our edge locations and visibility.

Proposed Methods of Measurement Network Deployment

With a goal of 100,000+ measurement locations, this would be an order of magnitude larger than the other systems we know of, excluding single-entity passive measurements available to a specific content provider with streaming or other apps or agents deployed, or large enterprises with endpoint measurement at sites/branches/endpoints.

Our focus is on continuous monitoring, so the measurements we are seeking to gather to start are focused on are latency, loss, jitter, and for the immediate term we only consider throughput where we can obtain passive traffic observation.

We are discussing a number of approaches to enabling this scale of distributed measurement, but our focuses in 2021 will be threefold in addition to our current paid deployment model: 1) Subsidized commercial service for enterprise and SP customers in exchange for measurement access of deployed endpoints; 2) Incenting app, content, SaaS, and other web service providers to share and receive aggregated views of passive performance data from RUM and Layer 7 vantage points; and 3) Subsidizing tech-savvy consumers by reimbursing for bandwidth provided, and/or with subsidized online services (the latter perhaps like Luminati but without questionable other uses for their proxy network).

We're not aware of any of these approaches leading to long-lived massively distributed edge measurement networks supporting longitudinal Internet measurement at 100k+ observation point scale, so our hope is to work on all 3 in parallel to try to get traction with at least one, or more.

Our primary focus in those methods will be working with customers on sharing active and/or passive measurements. We have had discussions with many of our customers, both enterprise and SP, about data to get back more global/aggregated data to help them understand if edge, inter-AS, cloud, or SaaS performance problems are localized to them, or to enable 'what if' analyses for informed traffic rerouting or transit purchasing decisions. To fully achieve the stated challenge of both massive edge deployment - and with data that can be shared - we may still require end-user access.

Areas of Concern

The issues we most need to solve include:

- Which data - Our plan is to start with even wider active measurement, but is it possible to get permission to obtain, process, and share passive data?

- Infrastructure - Is it tractable to financially incent true edge deployment of active measurement? Can that be done with a combination of commercial and research/governmental funding? Can we obtain a huge footprint with incentives that won't break the bank for even in a combined private and public sector partnership? Is broadband-edge deployment too noisy with individual and neighborhood usage to support continuous measurement, whether low bandwidth or eventually to add on throughput testing?

- Data sharing policy - Especially with passively observed data, is it possible to address privacy concerns with aggregating Layer 7-observed data? Is it possible to address competitive, marketing, and brand concerns with shared actively measured data across providers of infrastructure and services? Is access to actionable aggregate data sufficient incentive for sharing?

Ways the NSF can Help

We believe that the NSF can help by sponsoring:

- Infrastructure - Bootstrapping a broader active measurement infrastructure will be expensive, and we are prepared to continue to bear significant costs ourselves, but subsidization will help deploy more widely. We also are prepared to stand up analytics infrastructure, but depending on types of collaboration and volume and frequency of produced data, supporting that analytics and distribution infrastructure to participate in a federated or shared infrastructure for subsets of the data may become difficult for us to support internally unless it provides value we can monetize with customers.

- Data sharing - We believe that collaboration with NSF-funded researchers will help develop maturity around communicating about sharing within privacy policy; rigor of aggregation and obfuscation methods; and interest commercial operators who are interested in performing research, contributing to broader community efforts, and using collaboration as an investment in future recruiting pipelines.

- Long-term archive - We keep years of many telemetry sources internally, but depending on research interests of data consumers, it may be beyond our ability to subsidize long term storage and access. There may be benefits to having funded sharing infrastructure outside of the control of us or any one entity. Additionally, in a federated model, if one or more hosts of critical data are acquired, change strategy, or shut down, access to that data could be lost.

Background on Kentik

Kentik is a privately-held network analytics company providing real-time ingest, context enrichment, learning, and historical storage and query of network flow, device metrics, BGP routing, and active measurement data. Customers include hundreds of major cloud providers, hosting companies, international transit networks, incumbent telecom operators, broadband networks, WISPs, major wireless networks, gaming companies, content companies, media and entertainment, finance, and other enterprises. We deliver analytics as a service via commonly hosted “Public SaaS” infrastructure in the US and Germany, as well as private clusters for customers who desire them.

Our commercial focus is in real-time management, control, and automation of access and delivery networks, but we are open to participating in shared measurement infrastructure, sharing many of the non customer-specific results of our own measurement, and enabling opt-in customer contribution to such systems. Our focus in these areas is around gaining access to wider measurement vantage points than we deploy ourselves, to enable us to provide better global performance or other dimensions of visibility to correlate and suggest better insights, remediation, and automation.

Kentik Data Sources

The majority of Kentik’s traffic sources are router and switch NetFlow, IPFIX, and sFlow, with an increasing density from cloud VPC Flow Logs, Layer 7 traffic sources, and host eBPF monitoring. Most customers send BGP feeds and device metrics (via SNMP or Streaming Telemetry) from every edge router sending traffic telemetry. Kentik also has over 250 active measurement agents deployed across 80 networks, growing roughly 10 networks and dozens of vantage points per month. We consume across, multiple clusters, tens of millions of telemetry events per second, mostly network flow data, and retain it at full resolution for 45-90 days, depending on commercial arrangement. In 2020 we processed and stored over 100 trillion flows, averaging 273 billion per day, the large majority joined with live BGP data. The majority of Kentik customers start with monitoring of their AS borders, then extend to cloud, backbone, data center, campus, and end user network components. Customers can already opt-in to re-feed normalized and enriched streams of network telemetry to their own systems and other observability platforms. We already provide aggregate Internet-wide measurements to our customers, and are planning to continue to drive community-based opt-in Internet and cloud data sharing.