Massachusetts Institute of Technology
Department of Electrical Engineering and Computer Science

 Lecture 10: Internet Measurements October 13, 1999

Overview. End-to-end delay characterization. Routing behavior and pathologies.

1 End-to-end delay and loss characterization

Jean Bolot's paper from SIGCOMM 1993. Uses *NetDyn,* a tool from Maryland (Sanghi), written to diagnose problems on the Internet.

1.1 Methodology

- Simple experiments: The Internet is complex; complex experiments only serve to complicate. Use a very simple experiment and see how far it can go.
- Send a small UDP packet, get an echo reply. Single bottleneck assumption. s_n = time at which nth probe sent, r_n = time at which echo received, $rtt_n = r_n - s_n$ and $\delta = s_{n+1} - s_n$. Bottleneck has service rate μ bps. [Why not simply use ping?]
- Cute analysis: Clever application of straightforward queueing theory and algebra—the beauty of this work is its simplicity!
- Analysis tool: *phase plots.* Idea: from phase spaces in dynamical systems—eliminate time and deal with other coordinates. Use rtt_{n+1} and rtt_n as space coordinates.

1.2 Delay analysis

- *Goal: Distribution of end-to-end delay.* By-products: bottleneck bandwidth, packet size distribution of Internet packets.
- At bottleneck, waiting time for nth packet is w_n. The question we ask is: how do successive waiting ties vary, *i.e.,* how is w_{n+1} related to w_n?
- *Light load conditions.* Here, $w_{n+1} \approx w_n$. $w_{n+1} = w_n + \epsilon_n$. ϵ_n has mean 0 and small variance. Now, in general $rtt_{n+1} = D + w_{n+1} + P/\mu$ and likewise for rtt_n, which gives us $rtt_{n+1} - rtt_n = w_{n+1} - w_n$. For light load, this is ϵ_n. This explains "$y = x$" line on phase plot.
- Heavy load: And small δ. Probe compression occurs in this case. Suppose two probes are separated by B bits. Before this buffer clears, $k-1$ more probes arrive (so now there are k probes waiting to be serviced). $w_{n+1} = w_n + B/\mu - \delta \approx w_n + B/\mu$. For the remaining packets, $w_{n+i} - w_{n+i-1} = P/\mu - \delta$. This is *probe compression.* RHS is smaller than 0. [Why?] Probe compression explains large number of points on line $y = x + (P/\mu - \delta)$.
- [What happens when δ is large?]

- Bottleneck bandwidth now trivial to estimate! [How?]

  Horizontal intercept gives bottleck bandwidth; if this is R, then R satisfies the equation $R = \delta - P/\mu$ and since R, P, δ are known, plug in to find μ.

- *Principle: Queues at bottleneck (no queues ==¿ not a bottleneck!) cause back-to-back probe packets to get spaced according to the bottleneck bandwidth.* This is the principle (but not the same machinery) used in packet pair, TCP's ACK-clock based congestion control etc.).

- Packet sizes. Use *Lindley's recurrence*. Related waiting times of successive customers into a queue.

  Suppose w_n is waiting time for nth item, δ_n is gap between the nth and $(n+1)$st arrival, and y_n is service time for packet n, then $w_{n+1} = (w_n + y_n - \delta_n)^+$, where z^+ denotes $\max(z, 0)$.

- Apply it twice: once to get waiting time of cross-traffic burst and once to get waiting time of probe $n + 1$.

- End up with $w_{n+1} - w_n = (b_n + P)/\mu - \delta$. This is not hard to see: the $(n + 1)$st packet is separated from the nth by b_n bits, so the time it waits is equal to w_n plus the time it takes to drain P bits (size of nth packet) plus the time it takes to drain the b_n cross-traffic bits, minus δ, which is how much later it arrived in the queue to start with.

- *Idea.* Because $w_{n+1} - w_n = rtt_{n+1} - rtt_n$, we now know what its distribution looks like. Construct a PDF of $w_{n+1} - w_n + \delta$. It peaks at different values of the plotted quantity. Perform a "frequency analysis" of this data.

- Pragmatics: Need to pick "right" value of δ (too small will overwhelm router; too large means too much intervening traffic). Also need to deal with real-world issues like coarse-grained clocks. As it turns out, for small-enough δ, some important conclusions can be made.

- [What do the peaks of this PDF correspond to?]

  Left-most is probe compression, then empty queue.

  Third frequency peak gives packet size estimate. [Treating packet size as an r.v., what statistic does this correspond to? Why?]

  It corresponds to the *mode*.

- $w_{n+1} - w_n > P/\mu - \delta$ implies one or more Internet packets have arrived in the interim. Using curve fit, has evidence of exponential distribution. While consistent with other observations of the time, doesn't gel with subsequent observations of L.R.D and self-similarity.

1.3 Loss analysis

- p_u is uncond. loss probability. p_c is conditional loss probability; probability that nth packet was lost given that the previous one was lost. g is "gap" or size of a burst loss.

- $p_u <= p_c$. Why?

- Looks like there's a typo in Table 3 for ulp when $\delta = 500$ms. (It's probably 0.097 not 0.97!)

- Bolot finds that g is close to 1. Augurs well for FEC. This finding is very inconclusive, though.

- $g = 1/(1 - p_c)$ under certain assumptions (rtt_n is stationary and ergodic). Under stronger assumptions, this is easy to show.

1.4 Summary

- Delays, as expected, arise because of cross-traffic. We can now, from end-to-end measurements, determine packet sizes of Internet cross-traffic. We can also get bottleneck bandwidths.
- Exponential inter-arrival times for Internet packets at a router. Makes sense, from independent users. But more recent work, which we'll study in a few weeks, shows LRD.
- Losses are not as bursty as one might expect. But then again, this part is not as thorough or conclusive.
- *Lesson.* Simple experiments and clever analysis can accomplish a great deal!

2 End-to-end Internet routing behavior

2.1 Motivation

- Large, complex system. How does routing *behave*. Question interesting on several fronts.
- Not performance, but *behavior*. Not diagnosis, but characterization. Not with router support, but end-to-end.
- Three main goals:
  - Characterize routing *pathologies.*
  - Characterize route *stability.*
  - Characterize route *symmetry.*
- Scarce related work until Paxson's thesis work: Chinoy at SDSC and (recently) Labovitz *et al.* at Michigan.
- Major experimental effort with lots and lots of trace data.

2.2 Background, tools, methodology

- The Internet is divided into AS's.
- *traceroute.* Written by Jacobson to diagnose routes. Invaluable tool for diagnosis and route data collection. Uses ttl and ICMP error messages from routers to determine routes. [Q: how does it work?]
- NPD infrastructure. (Traceroute servers exist on the Internet today—check out http://www.uniplus.ch/direct/app_trce.htm) for example.
  37 sites ran NPD in all.
- N hosts implies $O(N^2)$ paths. This quadratic scaling is a good property.
- Two data sets: D_1 and D_2. D_2 had more types of measurements based on observations and analysis of D_1. E.g., symmetric experiments, multiple measurement rates. Learned from experience with D_1.

- *Exponential sampling:* Want to observe "true" state of the Internet, but can only sample. How to do this? Well, memoryless property of the exponential distribution leads to the PASTA principle, which applies in most general situations.

2.3 Routing pathologies

1. **Loops:** Forwarding loop, information loop (e.g., counting-to-infinity), traceroute loop. NPDs obtain traceroute loops.

   Two prevalent kinds of loops: less than 3 hrs and more than 12hrs.

2. **Erroneous routing:** pathology where US to Europe went via Israel! But by and large, it works without errors.

3. **Midstream change:** Does happen, with bimodal recovery times.

4. **Fluttering:** Rapid oscillation in routing, could lead to out-of-ordering. Does happen; in one case it wasn't localized. Aside: out-of-ordering has increased recently, esp in localized regions because of dispersity between tightly coupled routers ("traffic engineering"—but they try and keep TCP flows on the same path).

5. **Infrastructure failures:** Overall network availability rate between 99.5% and 99.8%.

6. **Too many hops:** traceroute default is 30. Median hop length is 16, well-modeled by a discrete Gaussian. Diameter seems to have increased slightly between D_1 and D_2.

7. **Temporary outages:** Do happen, long ones last over 30s. No good explanation for difference between D_1 and D_2.

8. **Outage patterns:** Outages correlated with load.

9. **Trends:** most pathologies getting worse with time!

2.4 Route stability

- *Prevalence:* how likely are you to encounter a given route? This is an unconditional probability of seeing a given route.

- *Persistence:* how long before a route changes? Most interesting for the dominant route. Independent concepts.

- Why are these important? (Path predictability, learning about network, end-to-end path assessment, real-time state migration.)

- In general, Internet paths are dominated by a single route, but there are significant site-to-site variations.

- Persistence at multiple time scales, seconds (load-balancing flutters) to minutes (tightly coupled routers) to 10s of minutes to hours (intra-network changes) to days.

- Two-thirds of all paths had stable routes. (Only D_2 was analyzed for stability.)

2.5 Route symmetry

- Why is this important? (one-way delay, complicates measurement and debugging, asymmetric technologies—not in paper).
- Why does it happen? Technologies and/or metrics different, cold-potato/hot-potato routing between ISPs for cross-country traffic. (E.g., ignoring MEDs in BGP.)
- Paired measurements in D_2 allow this to be studied.
- 49% of all routes had "city-level" asymmetries. 30% of all routes differed in at least one AS on the forward and reverse routes.
- Result: asymmetry exists, is significant (and is likely to increase in future).

2.6 BGP behavior

- Study of pathologies by Labovitz *et al.* SIGCOMM 1997.
- Withdrawing a route previously withdrawn.
- Updating the same route—adds 1-2 orders of magnitude to BGP traffic.
- Route instability correlated woth load.
- Excluding pathologies, 80% of studied routes exhibit high degree of stability.

2.7 Summary

Impressive large-scale analysis of routing behavior. Strong methodology and fairly thorough documentation of pathologies, stability and symmetry characteristics.

Time for another such study, more extensive with performance characteristics too. IPPM working group (ippm-request@advanced.org).

Lesson. The Internet is extremely complex and complicated. There's no such thing as a "representative" Internet path, delay distribution, or workload. It has continued to defy rigorous characterization, despite excellent efforts such as Paxson's. And the research continues...