Day In The Life of the Internet 2008 Data Collection Event

http://www.caida.org/projects/ditl

Duane Wessels
The Measurement Factory/CAIDA

k claffy
CAIDA

NANOG 42
February 19, 2008

Short Story

- 2008 “Day In the Life of the Internet” data collection event this March.
  - We got some good data in 2006 and 2007
- You have interesting data. Your data can help researchers answer interesting questions. We would love for you to participate.
- We don’t necessarily want you to send your data to us.
  - Although we can store it for you if you prefer.
- After you collect data, we’d like you to annotate/describe it and tell others how they might be able to get it.

Motivation

- In 2001 a National Academy of Science report challenged the research community to ‘capture a day in the life of the Internet’ with as much scientifically grounded methodology as possible, and with resulting data as widely accessible as possible.

- In 2006 CAIDA and OARC coordinated 48-hour data captures from as many DNS root nameservers as possible.

- In 2007 we expanded participation, and developed supporting infrastructure, documentation, and analyses.

- Results, e.g., “The DNS Root Name Servers: a 2007 Snapshot”

- Inspired to further expand scope in 2008

2007 Participants & The Data

1. OARC DNS Root Nameservers: C, E, F, K, M dns packet traces with payload.

2. NaMeX Internet Exchange: AS112 data.

3. Open Root Server Network (ORSN): dns packet traces with payload.

4. Japan: WIDE, GigE campus/transit packet traces.

5. Korea: universities (3), R&E backbone access link (1) and commercial GigE link packet traces.

6. AMPATH: Miami IXP, anonymized OC12 trace (.edu link).

7. CAIDA: topology, ucsd telescope, IRcache squid logs.

8. Other available data: Internet2 netflows and Routeviews.

What did we learn from DITL 2007 data?

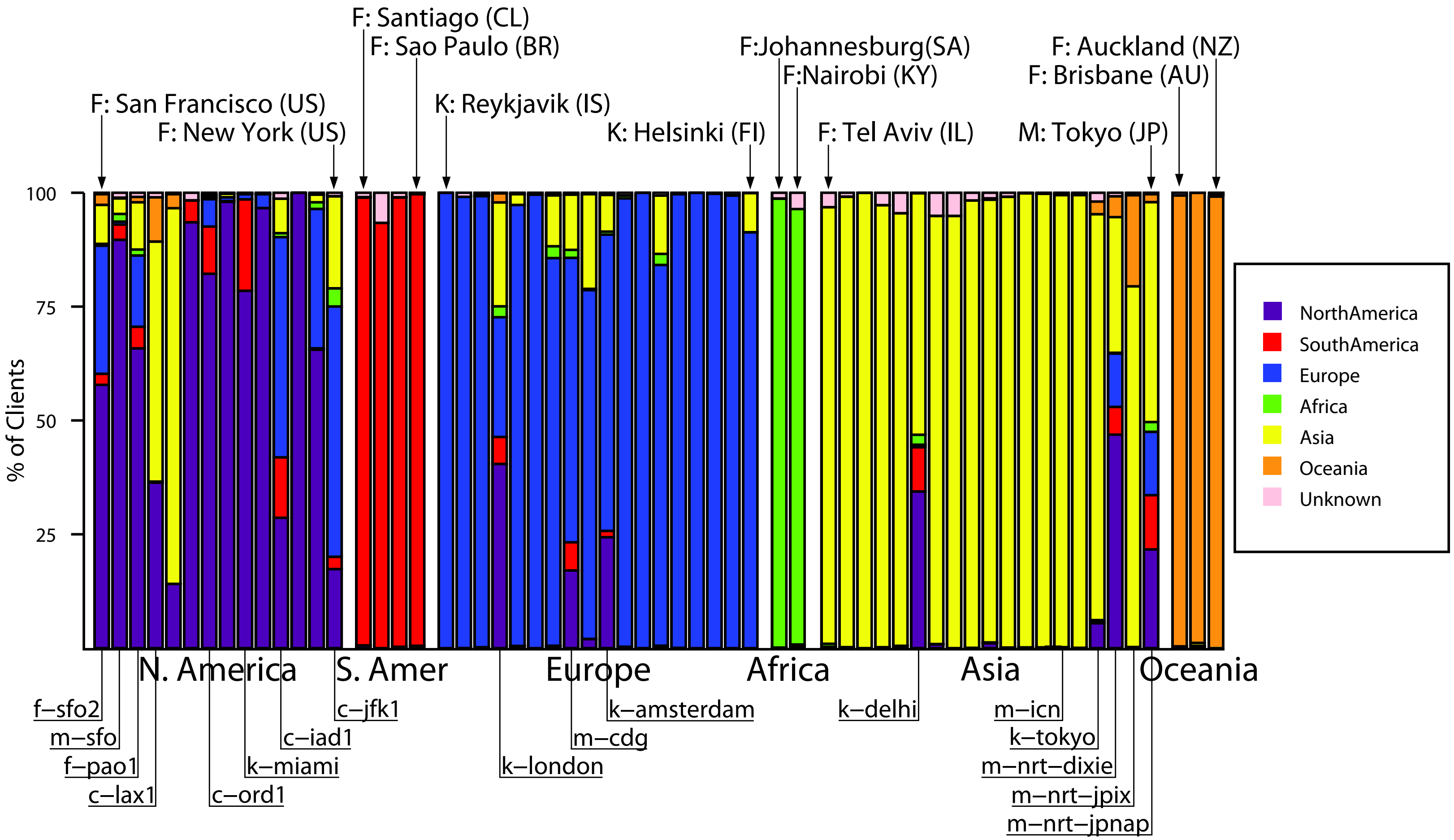
- 5% of clients contribute 95% of the query load (5 roots)
  - diverse query rate distribution

- 1–2% of queries to roots are legitimate
  - repeated, identical, referral-not-cached queries are 69% of total
  - the higher the query rate, the lower fraction of legitimate queries
  - for low rate queriers, similar distribution of query types
  - quite a few A6 (deprecated) queries

- Query rates at observed servers approximately doubled from 2006 DITL.

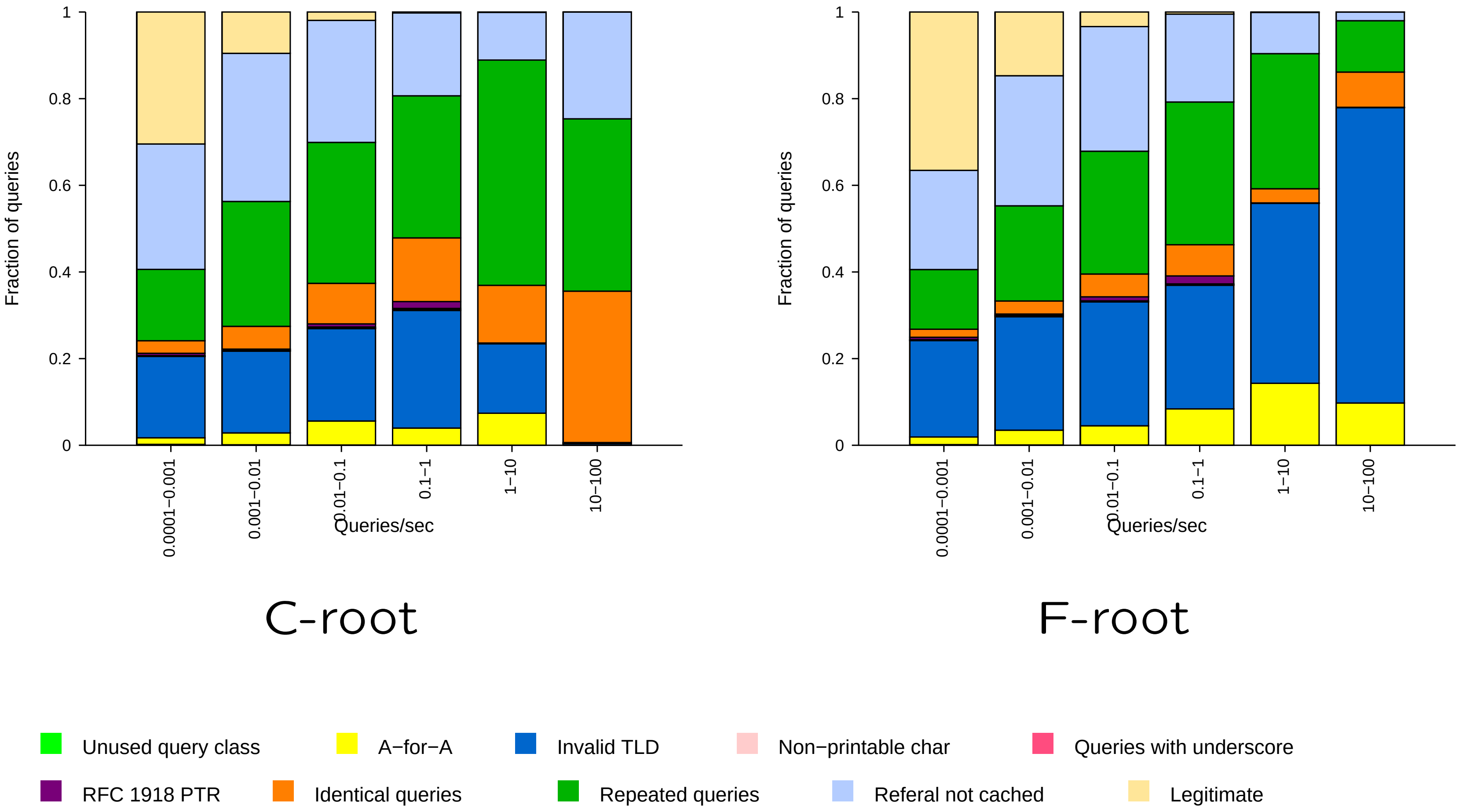
- Anycast at roots is effective at localizing resource consumption
  - and limiting impact of DOS attacks
  - stable: 98–99% of clients use only one instance of an anycast cloud

Where do root queries come from? (DITL 2007)



Server locations (longitude) on X-axis, Client geography by color.

What is all the pollution at the roots? (DITL 2007)



C-root

F-root

What did we learn from DITL 2007 collection process?

- You need more disk than you think.
- Some remote collection sites don't have enough bandwidth to upload data.
- Collecting data is easy; Indexing and annotating data is hard.
- Some clocks will be skewed, despite best efforts.
- Why use *tcpdump* when you can use *dnscap*?

We're addressing all of these issues for the 2008 collection:

`http://www.caida.org/research/dns/roottraffic/dnsroot_measurement_recommendations.xml`

Questions to pursue with DITL data: DNS

- Who/what is the source of root server garbage traffic?
- What does root server data suggest about trends in IPv6, DNSSEC, DNS packet sizes, prevalence of TCP-based DNS queries, and use of unallocated or unassigned IP address space?
- Can we characterize workload and performance of IDN deployments?
- Why are millions of clients querying old IP addresses of roots?
- How prevalent are misconfigurations, e.g., lame delegations?
- Is there correlation of DNS query target to geographic origin?

Questions to pursue with DITL data: traffic/performance

- What observable behavior is attributable to botnets?
- How can we identify applications (web, VoIP, video, p2p) and estimate their share of traffic and responsiveness to congestion?
- Do IPv6 traffic characteristics differ from IPv4?
- Is latency and jitter on the Internet increasing or decreasing?
- How are flow and packet size distributions changing, including bandwidth symmetry?
- How are TCP characteristics changing: flags, retransmits, buffer sizes, new versions?
- How is R&E traffic different from commercial traffic?
- How much web data is unnecessarily uncacheable?

Questions to pursue with DITL data: routing/topology

- What are the convergence properties of current and proposed routing protocols?
- Which ASes control how much of the Internet address space?
- What percent of Internet links block ICMP or other probing traffic?
- Estimate the distribution of hosts behind NATs.
- What percentage of users on open wireless networks use VPNs?
- How much allocated but “unused” IPv4 space remains?

Types of data needed for the above questions

- DNS query packet traces and/or logs from various places (roots, TLDs, IN-ADDR.ARPAs, ISP resolvers).
- Other dns-related measurements (passive and active).
- IPv4/IPv6 topology probing data.
- BGP feeds/updates, ideally with adjacent topology probes.
- Web cache logs.
- Anonymized report generator, e.g, coralreef, netflow-based.
- Router-level topology (anonymized), with event log per link.
- Consistent macroscopic ping data over years.
- Packet traces, appropriately anonymized.

Measurement methodology questions

- How can we measure host-to-host clock skew and NTP pool drift characteristics?
- How can we probe IPv4/IPv6 in a better way?
- When I ping a cable modem, how can i tell if I am reaching the modem or the host behind the modem?
- How dynamic are dynamic address assignments? What is the distribution of how long a single customer uses the same one?

Example Data Access Policies

- Unrestricted: anonymized versions w/o payload publicly available
- Restricted: access via OARC access agreement requires that the data and analysis must remain on OARC servers.
- Restricted: contact via email for access
- Restricted: access requests accepted for collaborative agreements to share analysis, implementation code, and results.
- Restricted: researchers may submit analysis code for staff to run on data
- Restricted: available to academic, government and non-profit researchers and members upon request

DatCat

- Though most have access restrictions, these datasets are indexed in the Internet Measurement Data Catalog (IMDC) nicknamed *DatCat*.

- DatCat provides metadata and annotations and information about raw data availability.
  - collection methodology, e.g., configurations, options
  - basic attributes: filenames, sizes, long and short descriptions, keywords, md5 hash, etc.
  - anomalies or peculiarities in the data, clock skew, etc...

- Allows researchers to learn about datasets before they request direct access, as well as annotate what they discover

Accommodating legal/privacy concerns: 2008 participation modes

- Make data selectively available under terms of your choosing.
- Index data (anonymized/aggregated) into datcat.caida.org. (Submit only meta-data or statistics about the data.)
- Contributing more “cooked” data, e.g., logfile summaries, RRD graphs, reports based on data collected internally by networks during the DITL dates.
- Contribute questions, feedback, analysis.

Conclusions

- We hope DITL data can inform discussions relevant to the operational future of the Internet, in the face of now anticipated growth for which the Internet has not been designed.

- But we can't do it without your help. Please consider participating in DITL 2008 with data, metadata, summary files, reports, analysis of resulting data, or your questions and thoughts on empirical science of the Internet.

Related Links

- A Day In The Life of the Internet
  `http://www.caida.org/projects/ditl/`
- A Summary of the January 9-10, 2007 Collection Event
  `http://www.caida.org/projects/ditl/summary-2007-01/`
- Internet Measurement Data Catalog (DatCat)
  `http://imdc.datcat.org/`
- Day in the Life of the Internet, January 9-10, 2007 (DITL-2007-01-09)
  `http://imdc.datcat.org/collection/1-031B-Q`

To participate, contact:

- k :: *kc@caida.org*
- Duane :: *wessels@measurement-factory.com*

The End