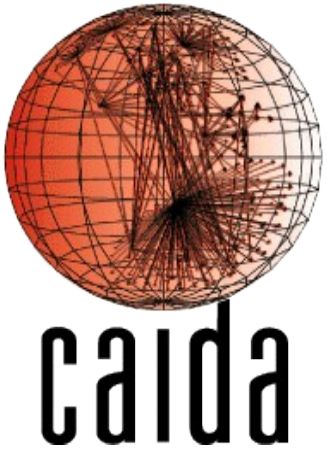
A Day in the Life of the Internet

Analysis and results of four years of DNS data

Sebastian Castro
sebastian@nzrs.net.nz

NZRS / CAIDA



Overview

- What is DITL and its motivation
- Evolution of the collection
- DNS Root Servers Data Analysis
- Operations useful data
- A glimpse of the NZ traffic to the roots
- Lessons learned
- Ideas for the future

What is DITL and its motivation

- U.S. National Academy of Science, 2001
  - Challenge to the research community
    - Network measurement
- CAIDA and OARC coordinated and organized the collection of data in the DNS root servers
  - Could be considered a prototype
  - Based on a trust relationship with operators
  - Been done yearly since 2006
  - Includes other sources of data

Evolution of the collection

	DITL 2006	DITL 2007	DITL 2008	DITL 2009
Participants	3 root servers	5 root servers 2 alternative root servers 1 AS112 instance 5 passive traces	8 root servers 2 alternative root servers 2 RIR 5 TLD 7 AS112 instances 6 passive traces 2 caching DNS servers	8 root servers 2 alternative root servers 3 RIR 8 TLD 6 AS112 instances 5 passive traces
Duration	46.2 hours	48 hours	48 hours	72 hours
Collection times	January 10-11	9-10 January	18-19 March	30 March – 1 April

DNS Root servers data

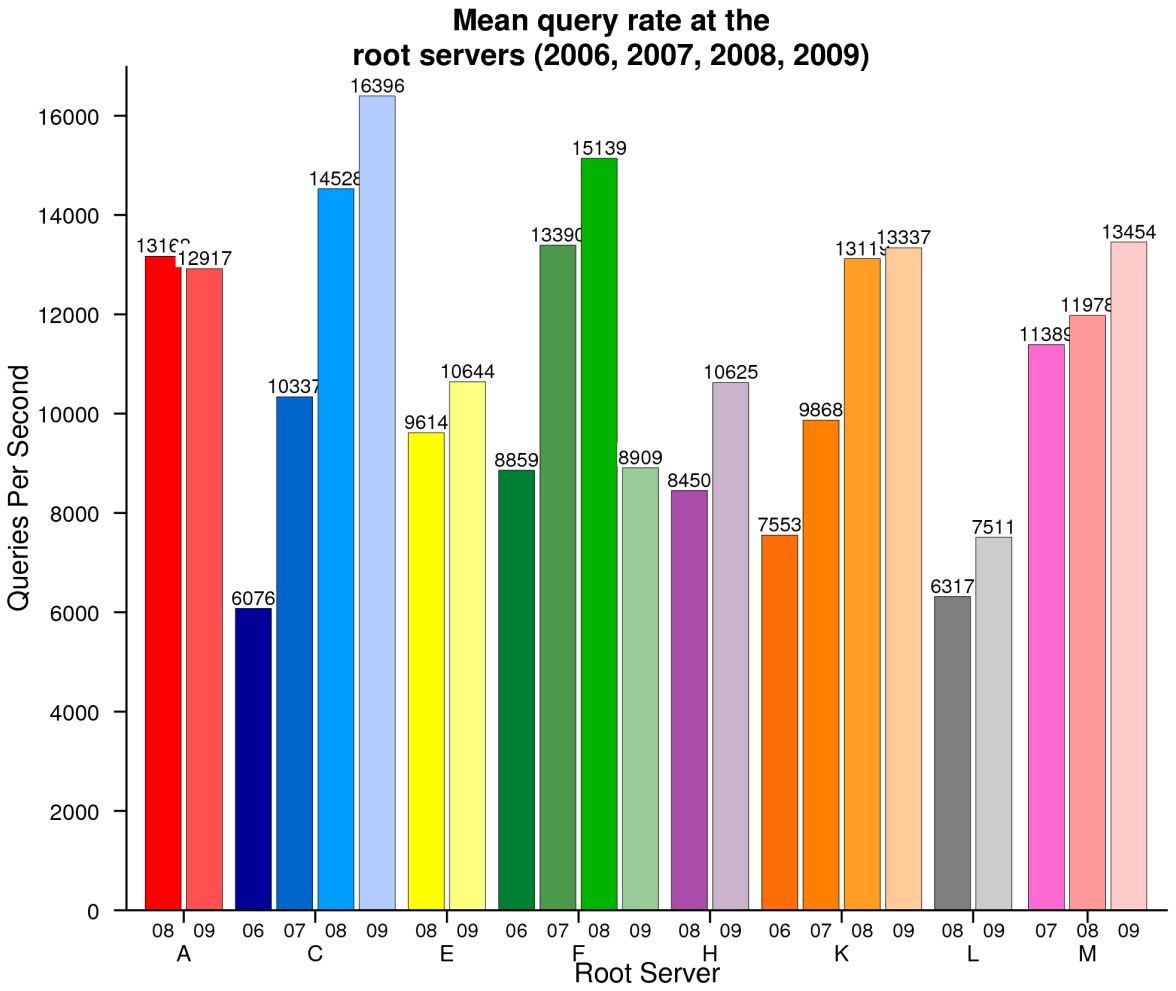
	DITL 2006	DITL 2007	DITL 2008	DITL 2009
Dataset duration	47.2h	24h	24h	24h
Dataset start (UTC)	January 10, midnight	January 9, noon	March 19, midnight	March 31, midnight
Number of instances (collecting/total)	C: 4/4 F: 34/37 K: 17/17	C: 4/4 F: 36/40 K: 15/17 M: 6/6	A: 1/1 C: 4/4 E: 1/1 F: 35/41 H: 1/1 K: 15/17 L: 2/2 M: 6/6	A: 1/1 C: 6/6 E: 1/1 F: 36/48 H: 1/1 K: 16/17 L: 2/2 M: 6/6
Query count	3.86 billion	3.84 billion	7.99 billion	8.09 billion
Unique clients	~2.8 million	~2.8 million	~5.6 million	~5.8 million
Recursive queries	4.02%	17.04%	11.99%	9.76%

Analysis: General Statistics

	DITL 2006	DITL 2007	DITL 2008	DITL 2009
TCP Bytes (1)	1.40%	1.65%	0.86%	0.74%
TCP Packets (1)	2.26%	2.67%	1.45%	1.20%
TCP Queries (1)	~221K	~700K	~2.07 million	~3.04 million
Queries over IPv6	0	~228K	~23 million	~29 million
Number of instances providing IPv6 traffic	0	6	16	31
Queries from RFC 1918 (2)	2.73%	4.26%	1.31%	1.57%

(1) L-root did not collect TCP traffic
(2) A, E, K and L-root did not see any traffic

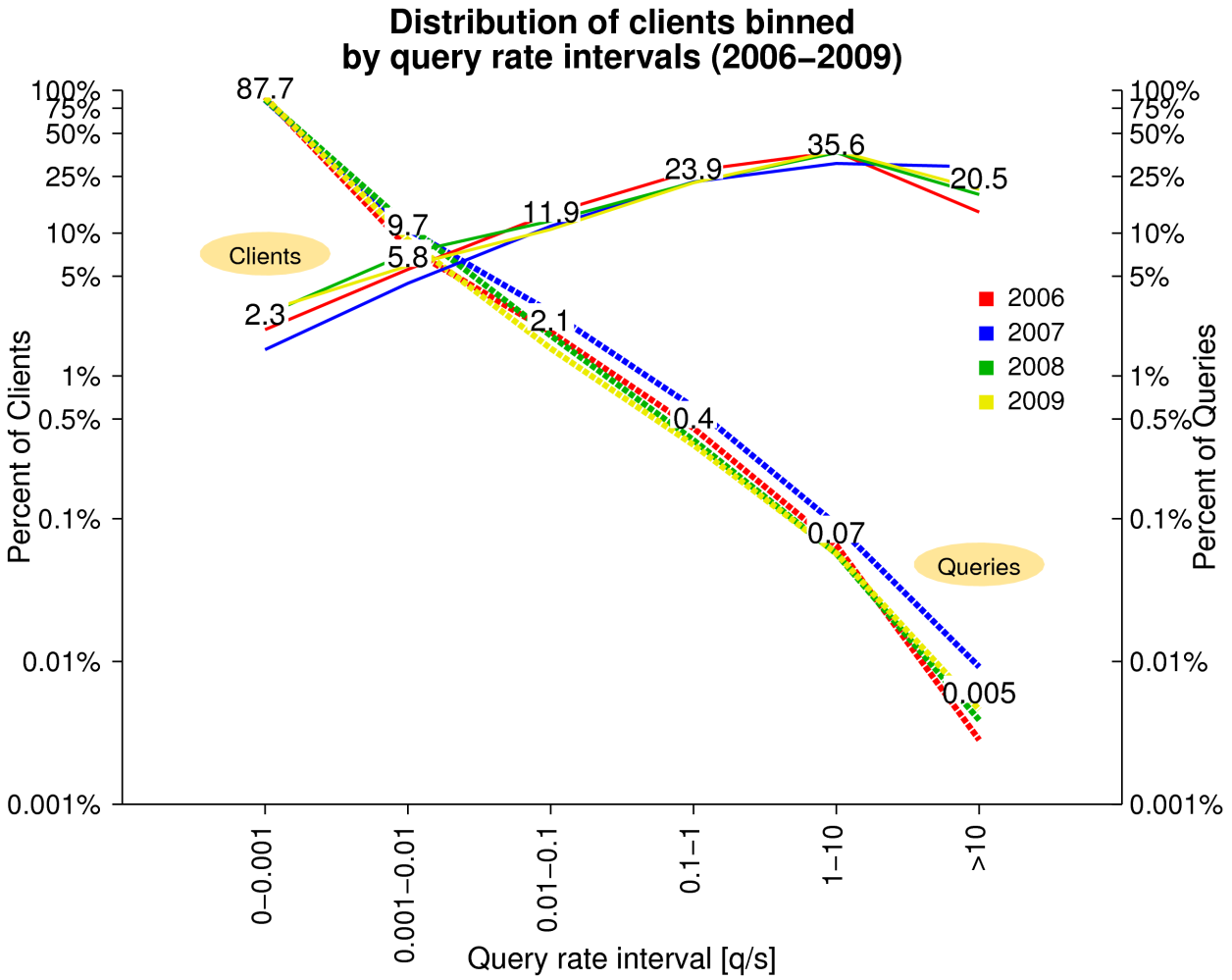
Mean query rate growth



Root	Growth 2007-2008	Growth 2008-2009
A		-1.91%
C	40.54%	12.86%
E		10.71%
F	13.06%	-41.15% *
H		25.74%
K	32.94%	1.66%
L		18.90%
M	5.17%	12.32%

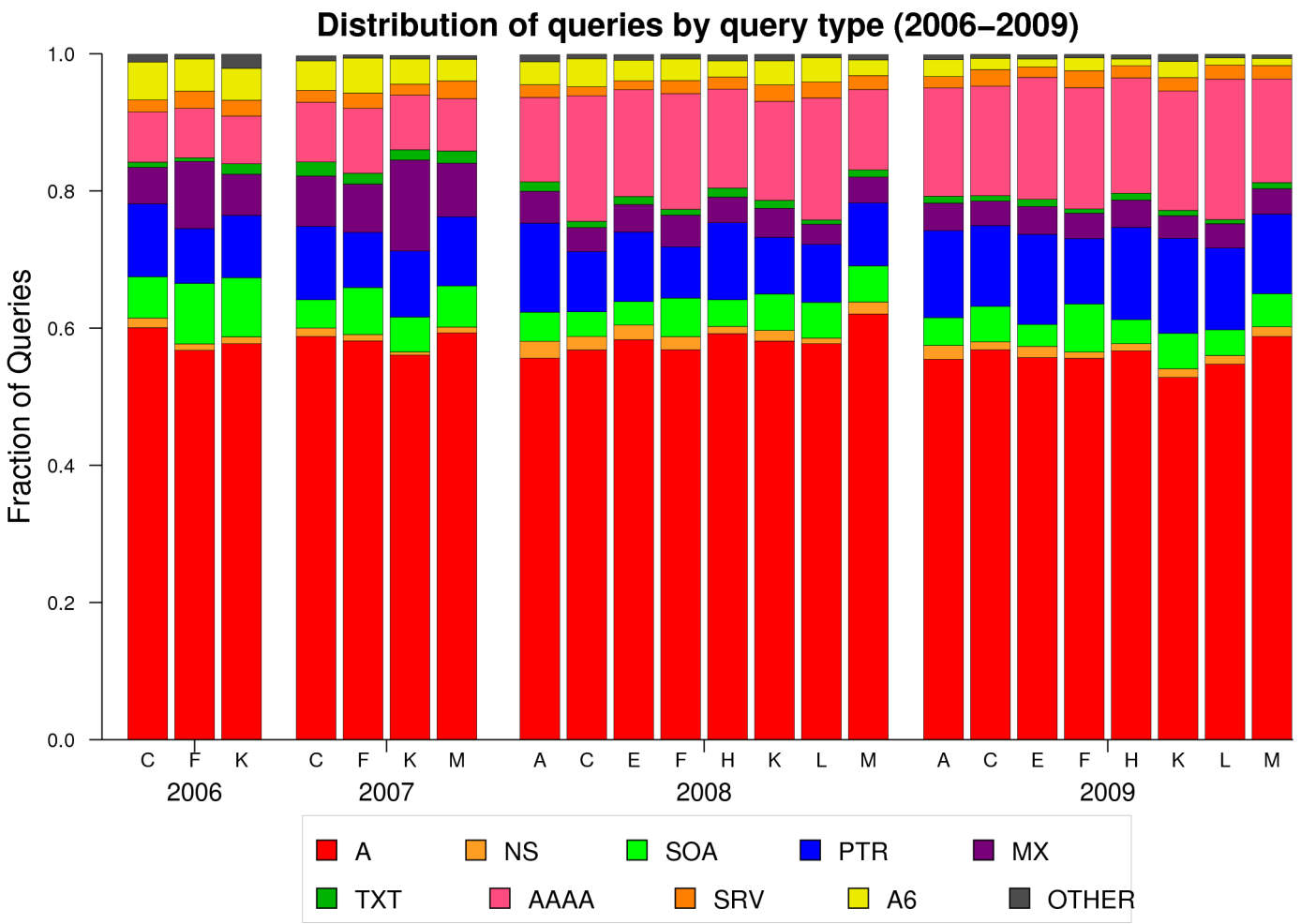
* Data from f-sfo (global instance) was not collected

Sources and their query rate



- Source addresses are grouped in bins by their query rate
- Leftmost column
  - 87.7% of the sources generated 2.3% of the traffic
- Rightmost column
  - Few sources generate more than 50% of the traffic
- Looks familiar?

Query type distribution



- A-queries the most predominant: around 60%
- Increase of AAAA-queries due to AAAA glue records added to the root zone in Feb 2008
- A6 queries still observable

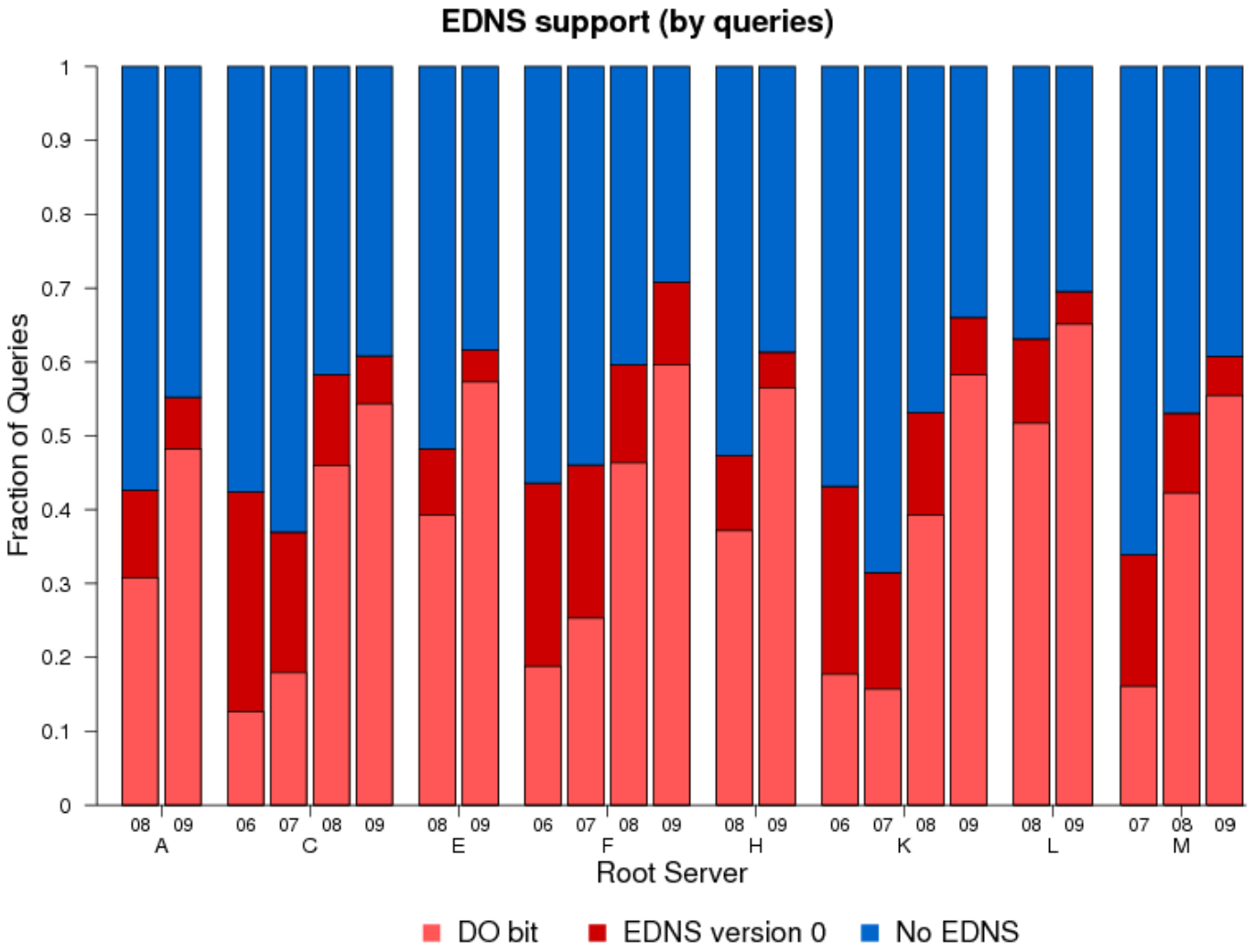
Operational useful data

- EDNS
  - Extension to allow signaling larger UDP buffer sizes and other flags
  - Essential to DNSSEC
- DNSSEC capable clients
  - By using the DO (DNSSEC OK) flag a client can signal if it wants to receive DNSSEC related records
- Invalid TLDs
  - Any query for an invalid TLD reaches the roots

EDNS analysis

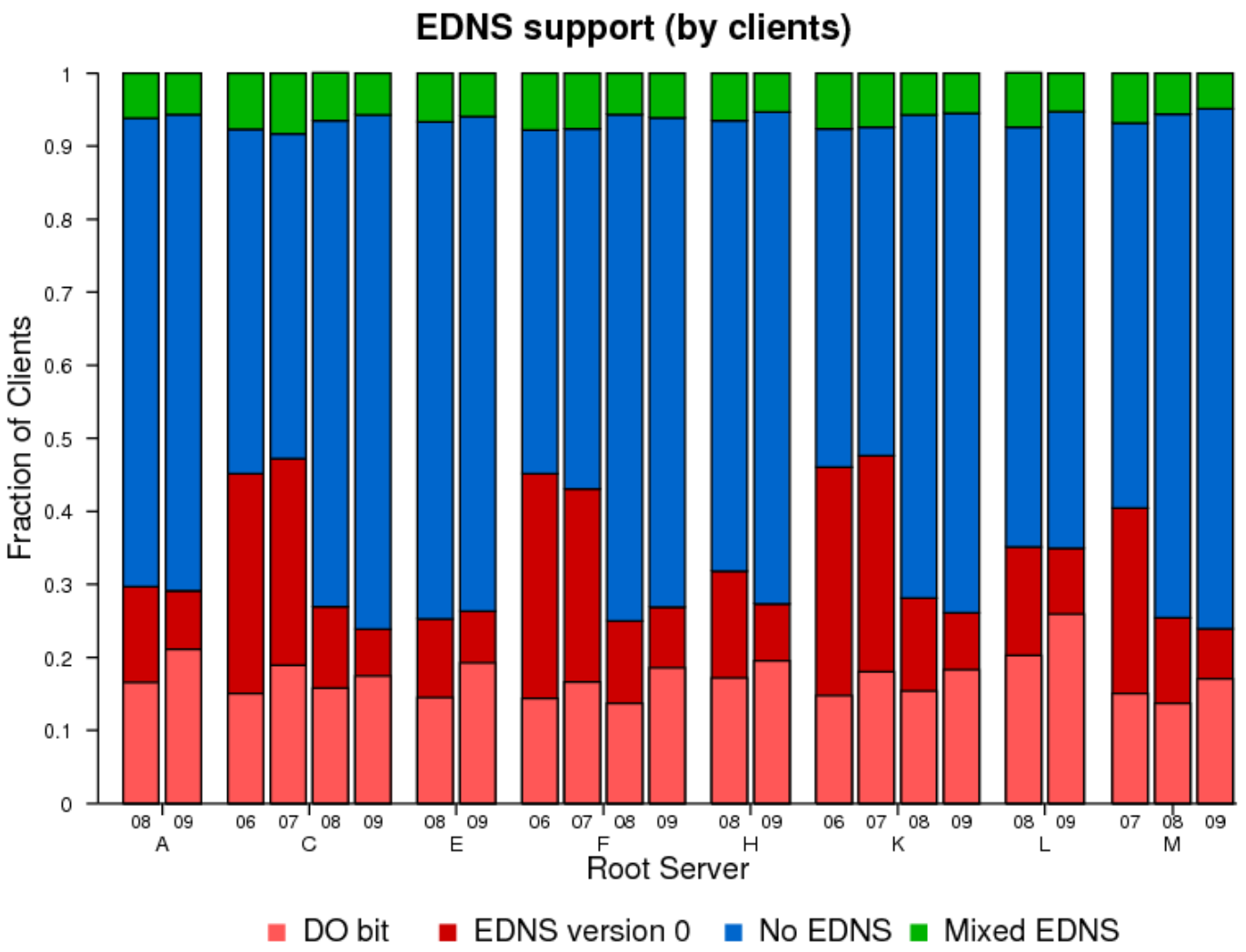
- We provide two metrics to estimate the presence of EDNS support on the traces
  - At the query level
    - By checking the presence of the OPT RR and analyze their values.
    - Possible values: No EDNS, EDNS0 (with or without DO bit)
  - At the client level
    - by checking all the queries coming from the same source address
    - Values: No EDNS, EDNS0, mixed (not all queries include EDNS support)

EDNS at the query level



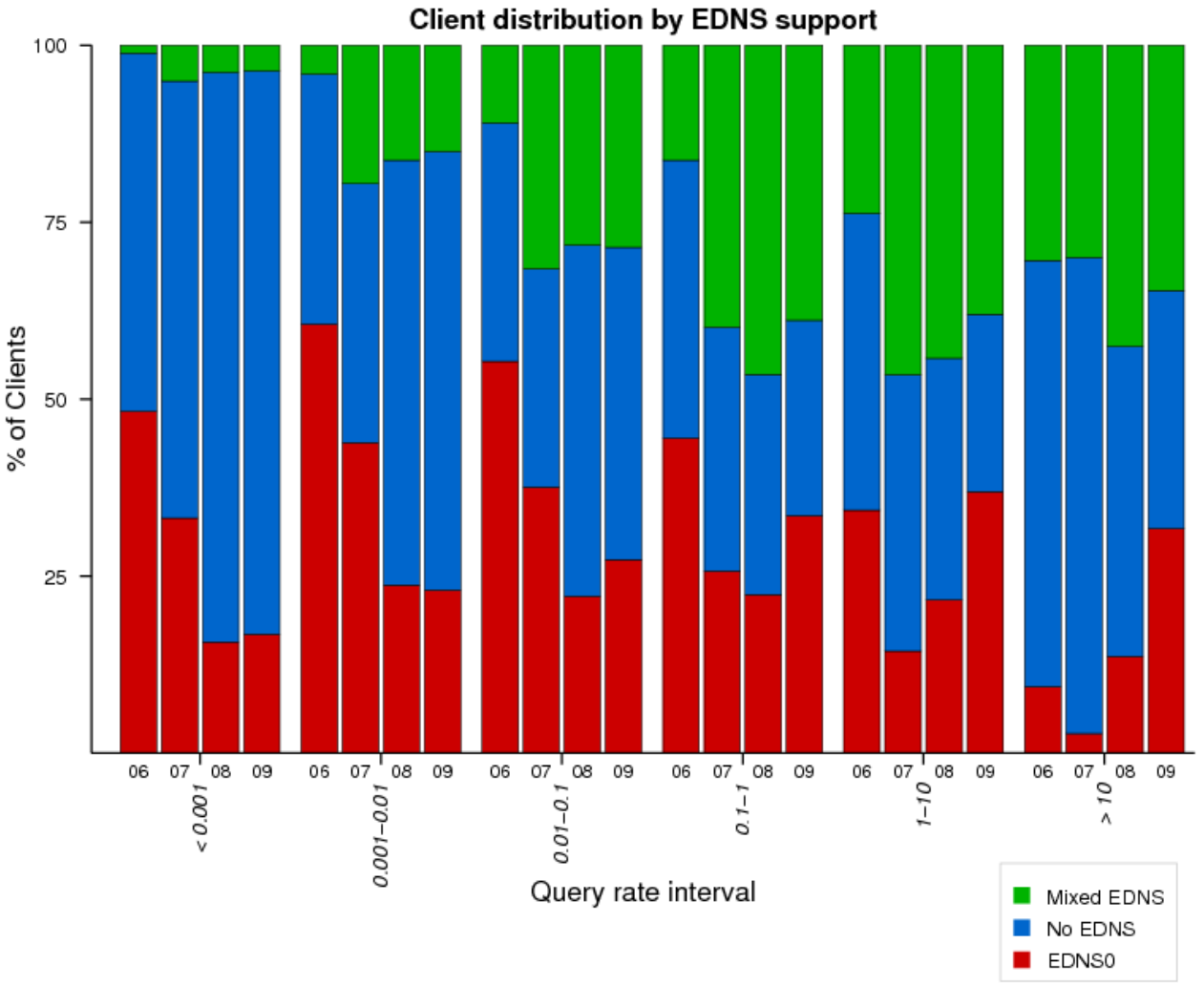
- Clear growth of EDNS, with a jump from '07 to '08
- Over 90% of the EDNS capable queries are DO enabled in 2009
- Good news, right?

EDNS at the client level



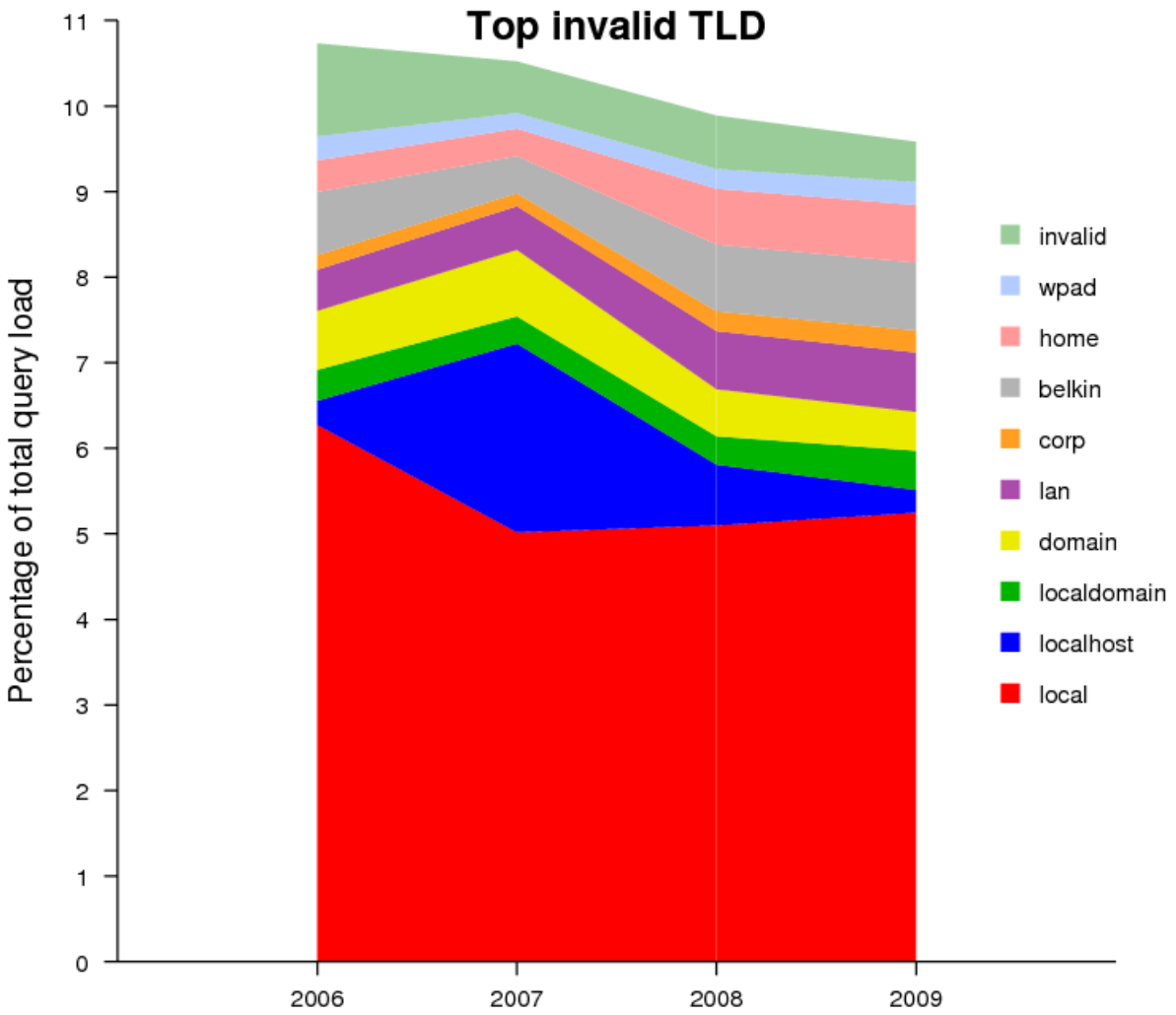
- At the client level, the situation is totally the opposite!
  - Reduced along the years
  - Around 30% support
  - The DO enabled/EDNS capable queries ratio is in the 60-70% range
  - How is this explained?
    - **The heavy hitters**

EDNS per query rate



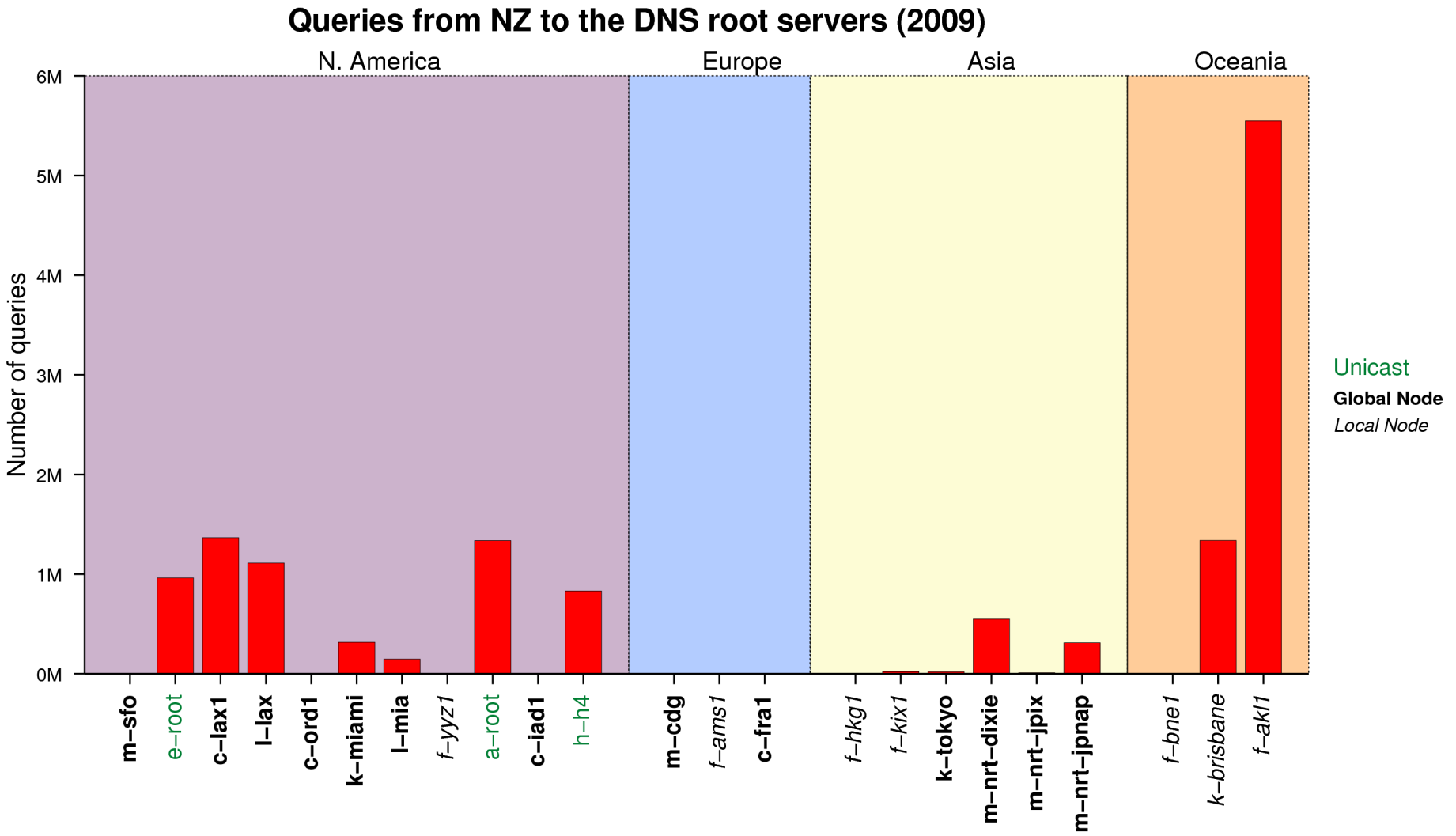
- How we can explain the difference?
  - We grouped the clients by their query rate
  - Clients sending few queries present less EDNS support
    - And they represent most of the clients
  - Client sending lots of queries present more EDNS support
    - Most of the queries (>50%) are generated by the two rightmost categories
    - Most of their queries are pollution :(

Traffic for invalid TLDs



- 10 invalid TLDs represent 10% of the **total** query load at the root servers
- The TLD has not changed in the last four years (only the ranking)
- If all invalid TLDs are included, the percentage moves from 18% to 26% (not shown)

NZ traffic to the roots



Lessons learned

- Data Collection
  - HARD!
    - Clock skew, data loss, wrong command line options, dysfunctional network taps
    - We rely in pcap format
- Data Management
  - More data, more participants, more formats!
  - Big efforts to normalize the datasets before start analyzing
    - Adjust/account clock skew, fill the gaps, create manageable file sizes and consistent time boundaries, separate sources, strip VLAN tags,

Ideas for the future

- The DITL collection and data is likely to be the right place to look for answer to deployment questions
  - DNSSEC for example?
- More brain and computing power is needed to extract particular information
  - Contact OARC if you are interested on access to the data
- May be we should share the analysis code?
  - We have solved some issues (or found workarounds) to handle large datasets
- Add some level of active measurement