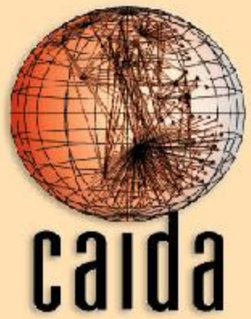


modeling the Domain Name System (DNS)

duane wessels
marina fomenkov
kc claffy

28 may 2003
nms pi meeting san diego, ca

kc@caida.org

DNS is critical infrastructure

refresher: how DNS works

DNS utilizes a hierarchical name space divided into zones that are distributed among the name servers. Each zone has one or more authoritative name servers responsible for answering queries for names within their zone(s).

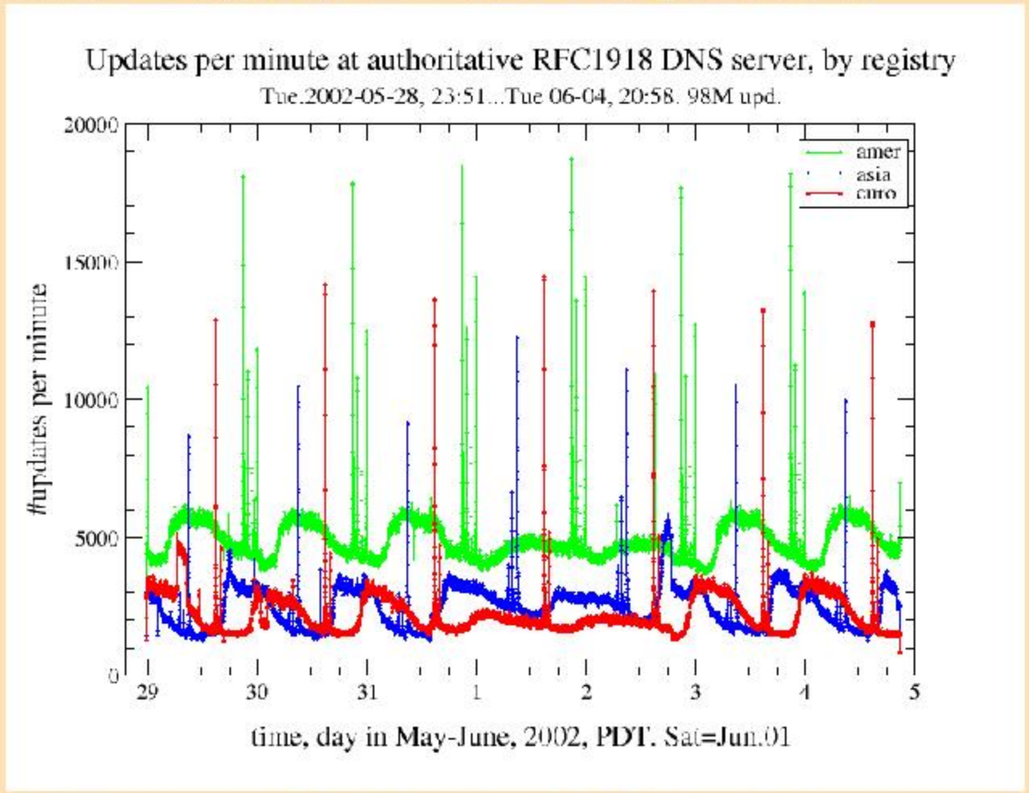
In order to reach a machine with the name not.invisible.net, one must send a query to the DNS server responsible for machines and/or sub-domains in the domain .invisible.net.

To find this DNS server, one must send a query to the server authoritatively responsible for .net. Such a server is called a global top-level domain (gTLD) server. To find the appropriate gTLD server, one must query one of the root servers. Currently there are 13 gTLD servers and 13 root servers.

last time: egregious macroscopic dns damage

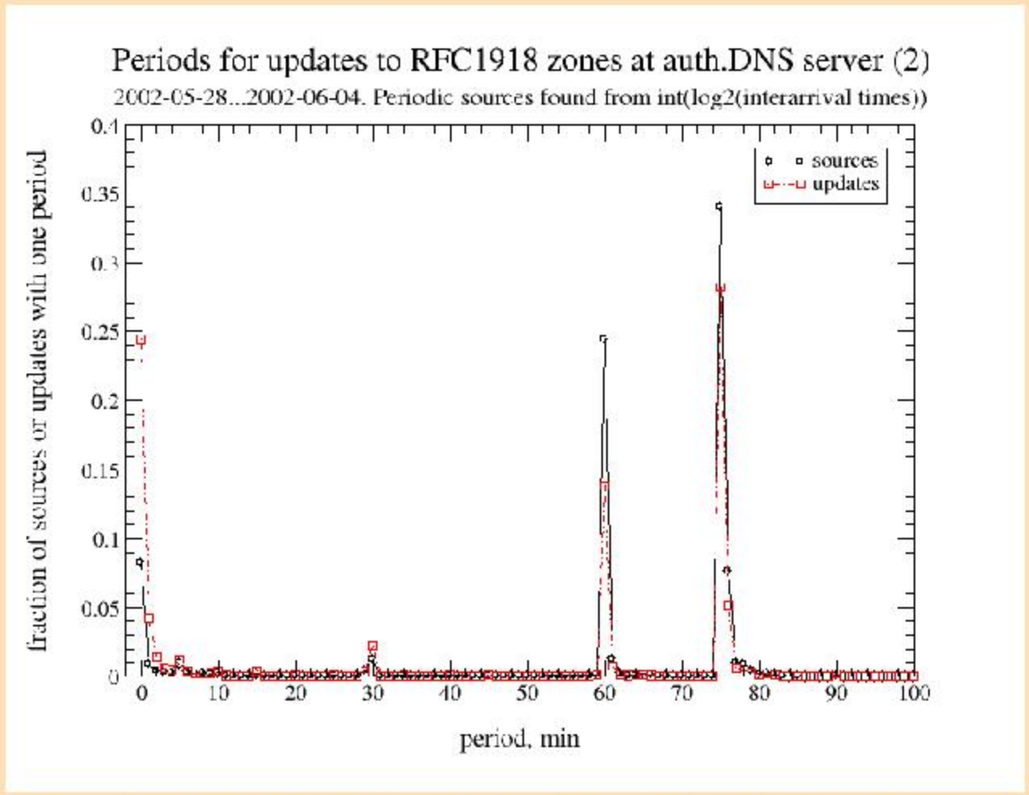
dns updates for private address space leaking up to roots

- spectroscopy analysis of RFC1918 updates
- RFC1918 updates coming from DHCP/nameservers
- millions a day getting to root name servers (whee)
  - **51.4M updates in 86.5 hr = 10,000 per minute = 165 per second**



- **weekday, weekend patterns; weird spikes at midnight local time**
- **4 in the US, 3 in Asia, 2 in Europe**
- **can see that Asians work on the weekend**
- **can see that Europeans and Asians get to work on time**

... global RFC1918 damage in DNS system



- rare to get macroscopic Internet data so radically broken
- who is trying to update the roots anyway?
  - **dsl, cablemodem, small population providers, developing countries**
- verified that vast majority derive from two OSes: Windows 2000 and Windows XP
  - **majority of updates from sources that send them constantly**
  - **bulk of workload from contributions of medium size, not mice/elephants**
  - **most source IP addresses are of home and small business users (owned by individuals, not organizations)**
  - **connected to the Internet via cable, DSL or phone-based ISPs**
  - **majority using software with default vendor settings**
  - **academic, corporate, backbone networks contribute little rfc1918 update traffic**

...global threat arising from single vendor

- combination of Microsoft software features & misconfigurations essentially causing a slowly paced massive distributed denial of service (DDOS) attack on the root name server system

- current state of fielded desktop software poses substantial & increasing burden on (if not threat to) the robustness of the global Internet

- software and setups affecting global systemic Internet stability must be designed more carefully wrt potential effects of:
  - software engineering decisions
  - misimplementations
  - misconfigurations

- measurement can make a huge difference

next step: toward realistic DNS simulation

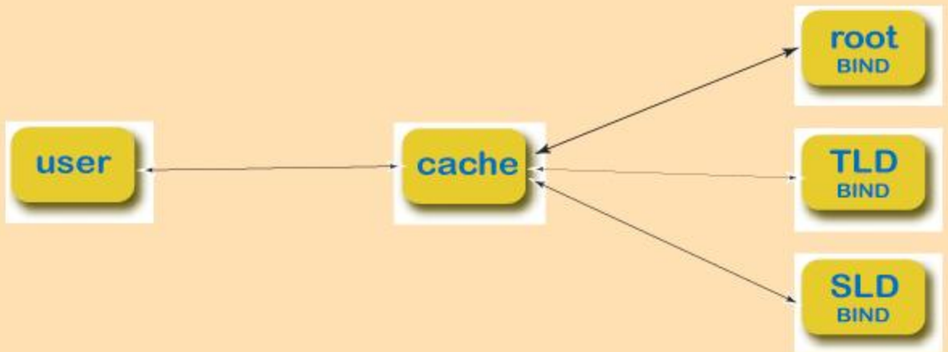
DNS simulation parameters and constraints

- based on real data
- 13 root servers (root zone file)
- 13 gTLD servers (159 zone files)
- thousands of SLDs
- use real SOA resource records where possible
- real user query workload
- test all commonly available nameserver implementations
- variety of realistic experimental conditions

just first step toward global DNS modeling

DNS simulation methodology

requires 5 computers



DNS workload generation

derived from 24 hours of IRCache logs

- 7,507,544 hostnames
- filter invalid data (e.g. query asking to resolve an IP address)
- extract unique SLD zones
- extract valid unique TLD zones
- keep invalid TLDs to model error handling

workload is played back as fast as possible

DNS caching name servers under test

client-side caching dns servers

- configuration files contain only the hints for the root zone

most common DNS implementations today:

- BIND (8.3.4 and 9.2.2)
- djbdns (1.05 w/default 1M cache and w/ 100M cache)
- DNS software in Windows 2000 (v5.0.49664)

runtime < 2 hours

DNS simulation: experimental conditions

- Experiment 1: no loss, no delay (ideal conditions)
- Experiment 2: 10% query loss, no delay
- Experiment 3: no loss, linear delay

Root name servers (and those for .com, .uk, .jp) are given delays starting at 10 msec and increasing by 15 msec for each nameserver. The .org nameserver delays overlap with the .com nameservers, going from 100-210 msec in 15 msec increments.
SLD nameservers have no delays.

- Experiment 4: no loss, constant delays
- Experiment 5: 100% loss, no delays

DNS responses from simulated zones

simulated root zone

- looks much like the real root zone
- same (refresh,retry,expiry,minimum) values in the SOA record
- 13 NS records ([a-m].root-servers.net) and 13 glue (A) records for those nameservers
- TTLs for all records match the real root zone

159 top level domains (TLD) zones, each containing:

- SOA record with values taken from real SOA records for its zone
- some number of NS records, same number of glue (A) records (match real world, e.g. 13 NS+A records for .com, 8 NS+A records for .it)
- values inside NS+A records, however, are fictitious; 24-hour TTLs for all NS and glue (A) records.
- delegations for subdomains

DNS responses from simulated zones (2)

82,891 second level domain (SLD) zones, each contains:

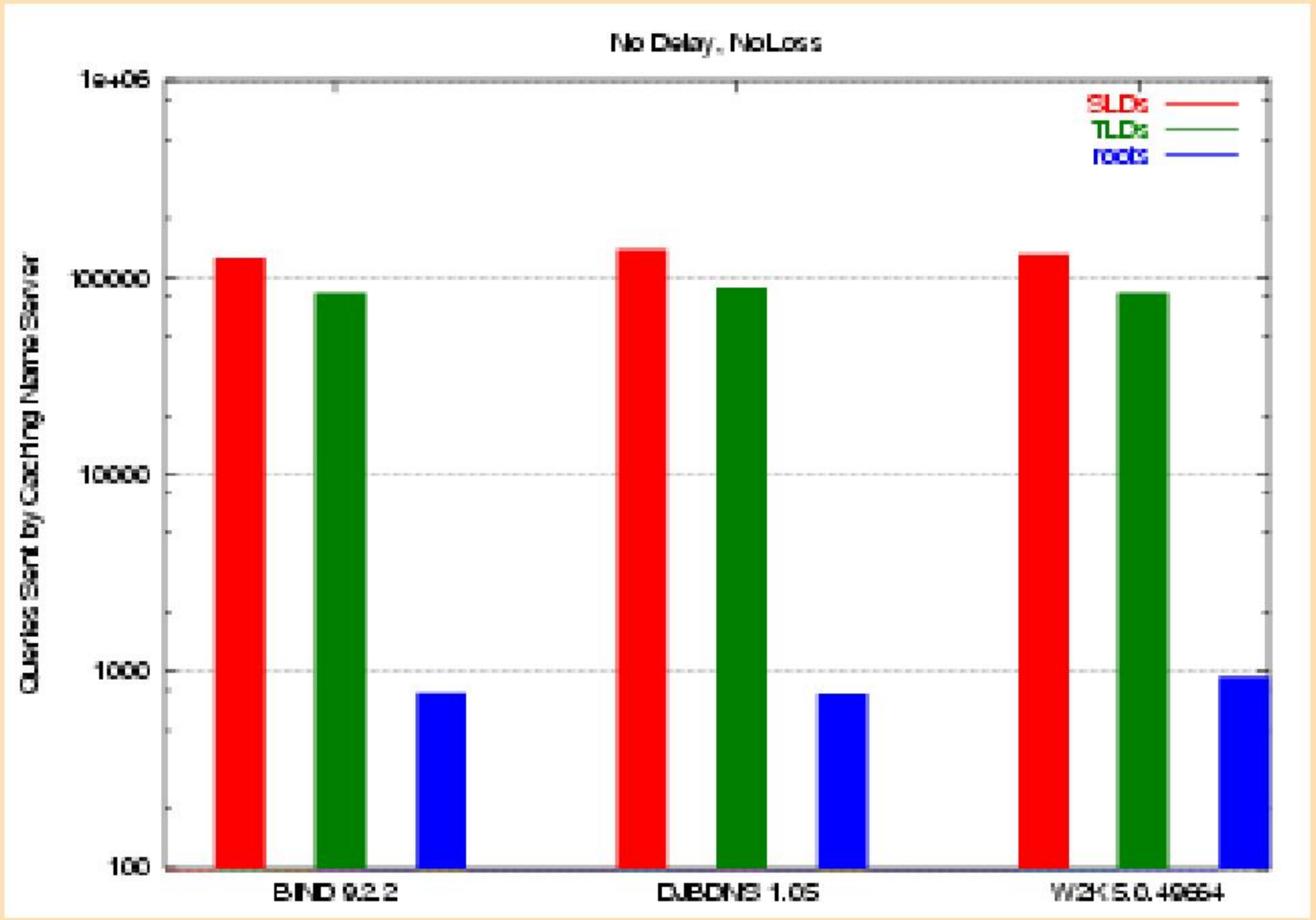
- 1 SOA record with fictitious values
- 2 NS records, 2 glue (A) records
- some number of A records for hosts in the zone
- both NS and A records match those in the parent zone (TLD) file
  - 24-hour TTLs
- name server IP addresses from a pool of 254 addresses
  - allocated sequentially, wrap around as needed
- A records for hosts in the SLD zone have random IP addresses
  - 12-hour TTLs

DNS responses from simulated zones (3)

- small number (254) of TLD and SLD name server IP addresses occurs because BIND binds to each IP address twice (TCP and UDP) and is limited to 1024 file descriptors
- zone files are perfectly consistent
  - no lame delegations, non-answering name servers, or other errors.
- each hostname has only one A record
- no CNAME records
- each SLD has only two NS records
- TTLs are larger than the time to run an experiment

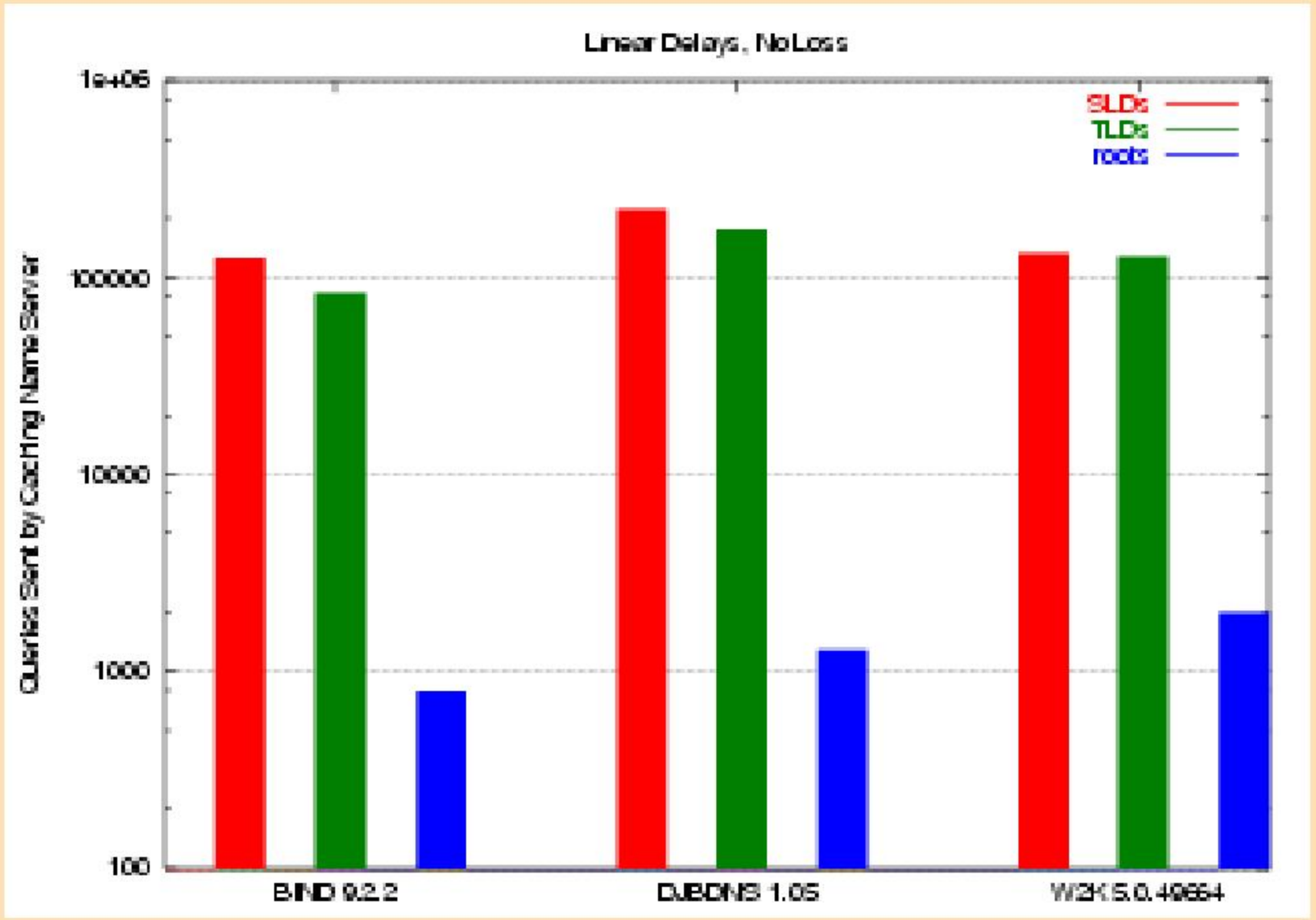
once a record is cached, it stays cached

overall comparison of caching name servers

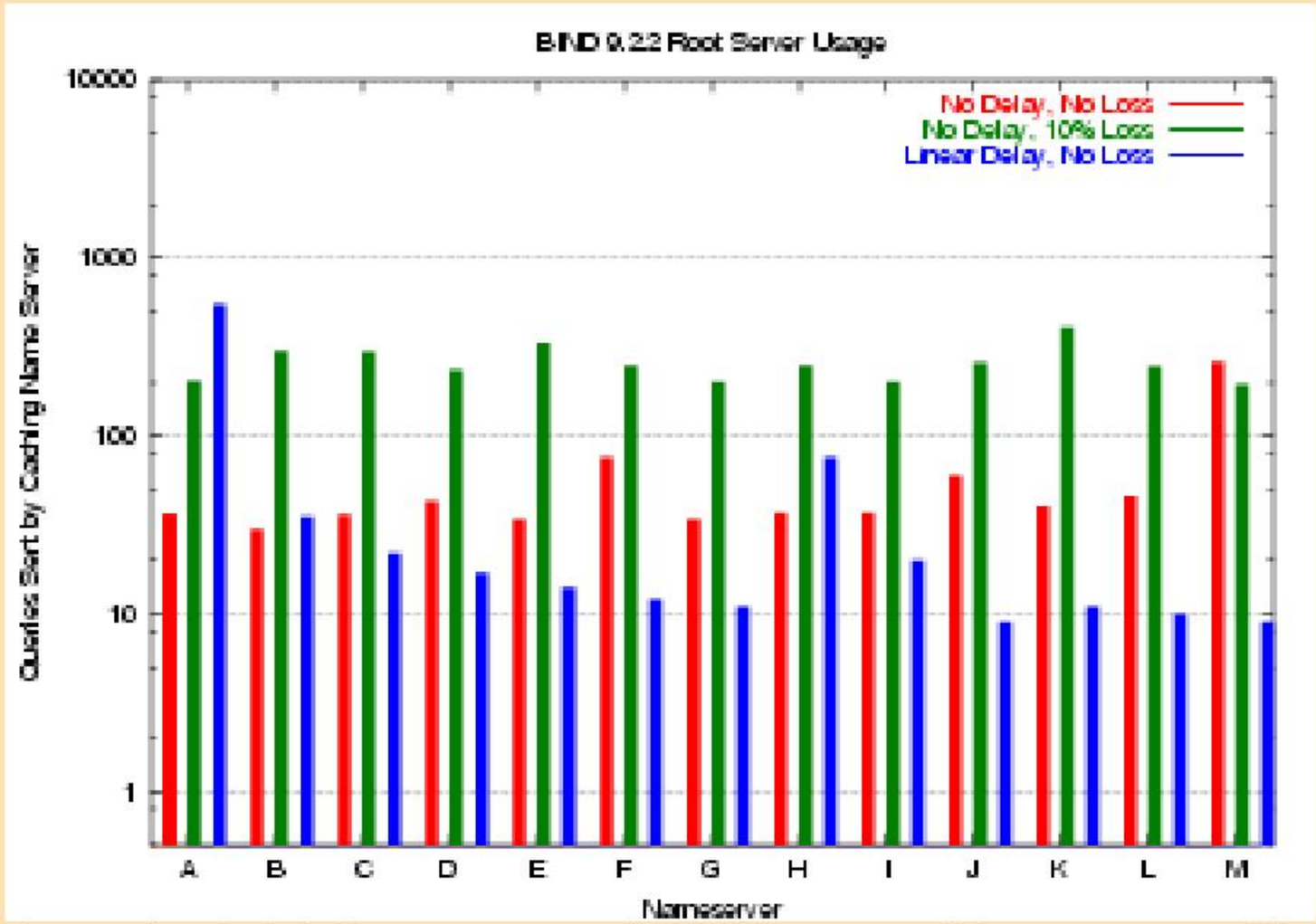


W2K 5.0.49664 exhibits more root queries because it appears to delete negative responses from the cache after only 15 minutes.

overall comparison of caching name servers

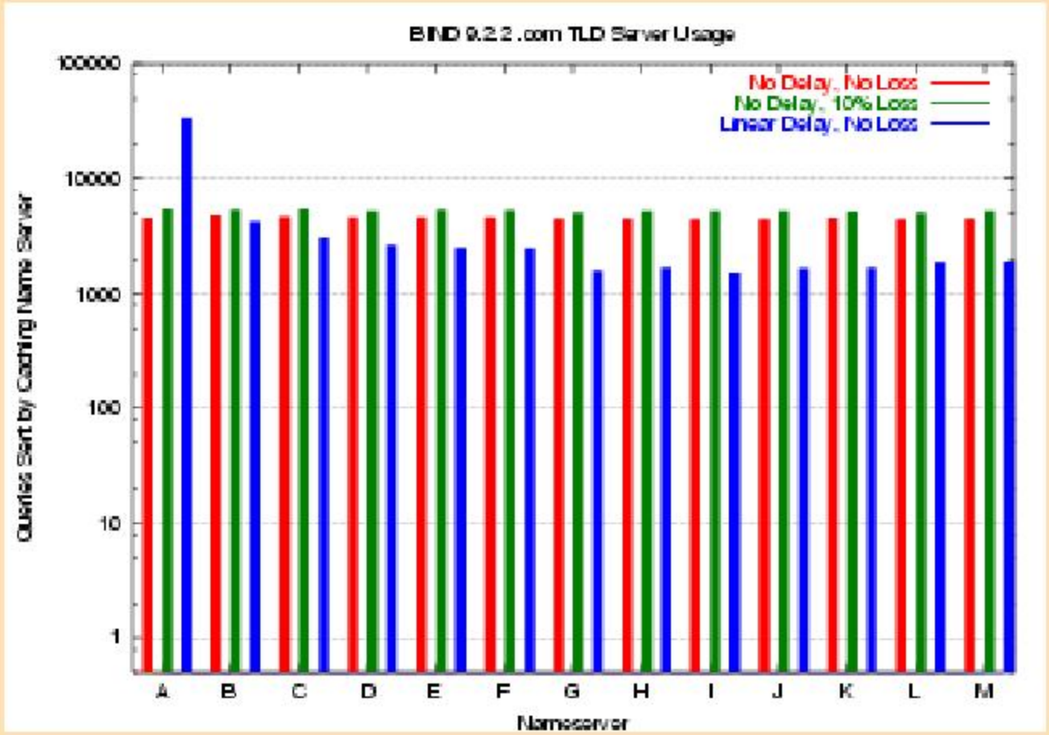


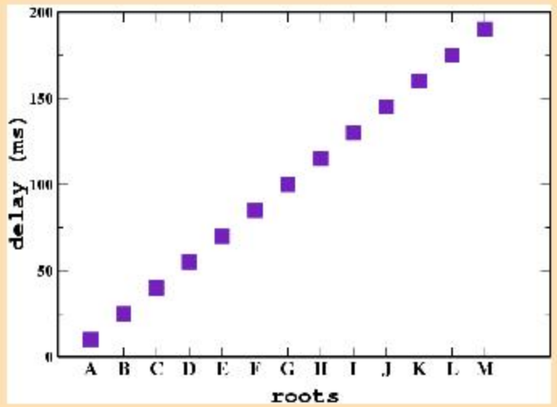
BIND 9.2.2 queries to roots



BIND 9.2.2 always selects the initial root at random, but transitions to a uniform distribution in all experiments with no delay (red and green). High peaks at M & F (red) and K & E (green) in the root usage graph result from BIND's random initial root selection. In Experiment 3 (blue) with linear delay, BIND initially randomly selected H to handle root queries. For both root and .com TLD queries, BIND detects delay by evaluating RTTs and then tends to select roots with the lowest delay (A-D).

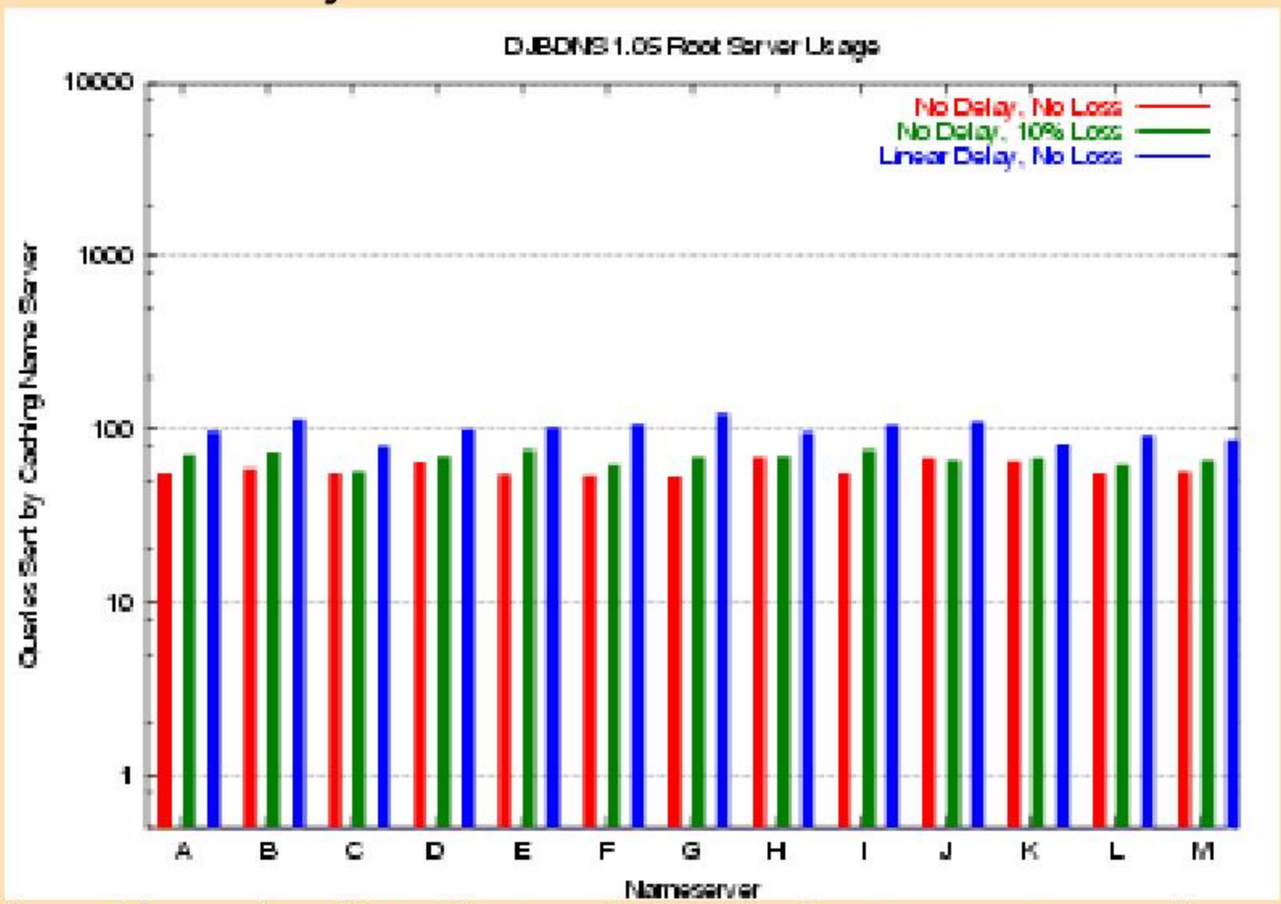
BIND 9.2.2 queries to .com TLDs





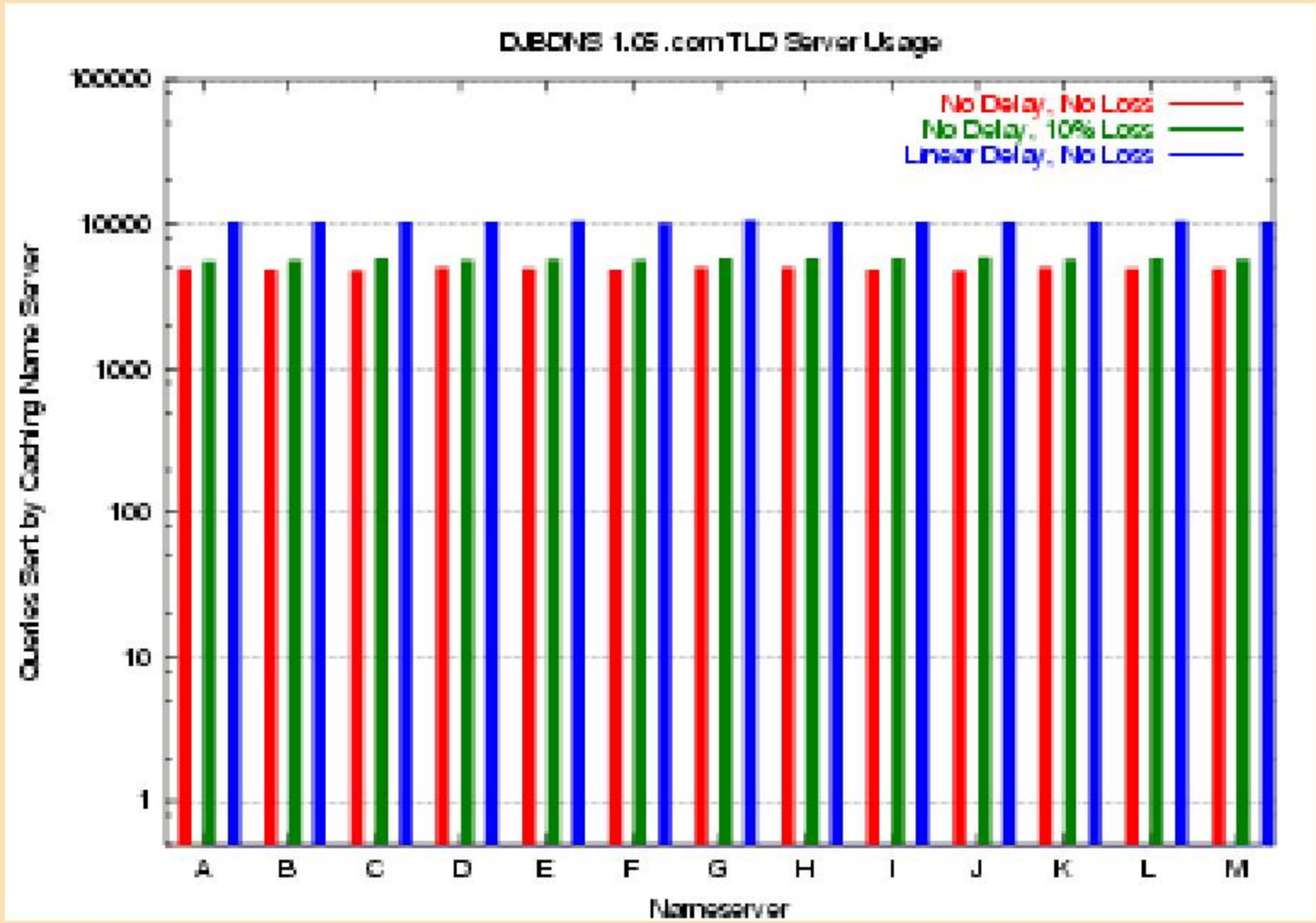
djbdns 1.05 queries to roots

While 1MB is the default cache size for djbdns, compared to a 100MB cache, the 1MB cache increases the load on the TLD and SLD nameservers by a factor of 4, and increases root server load by a factor of 10. To avoid this, we simulated with only the 100MB cache.

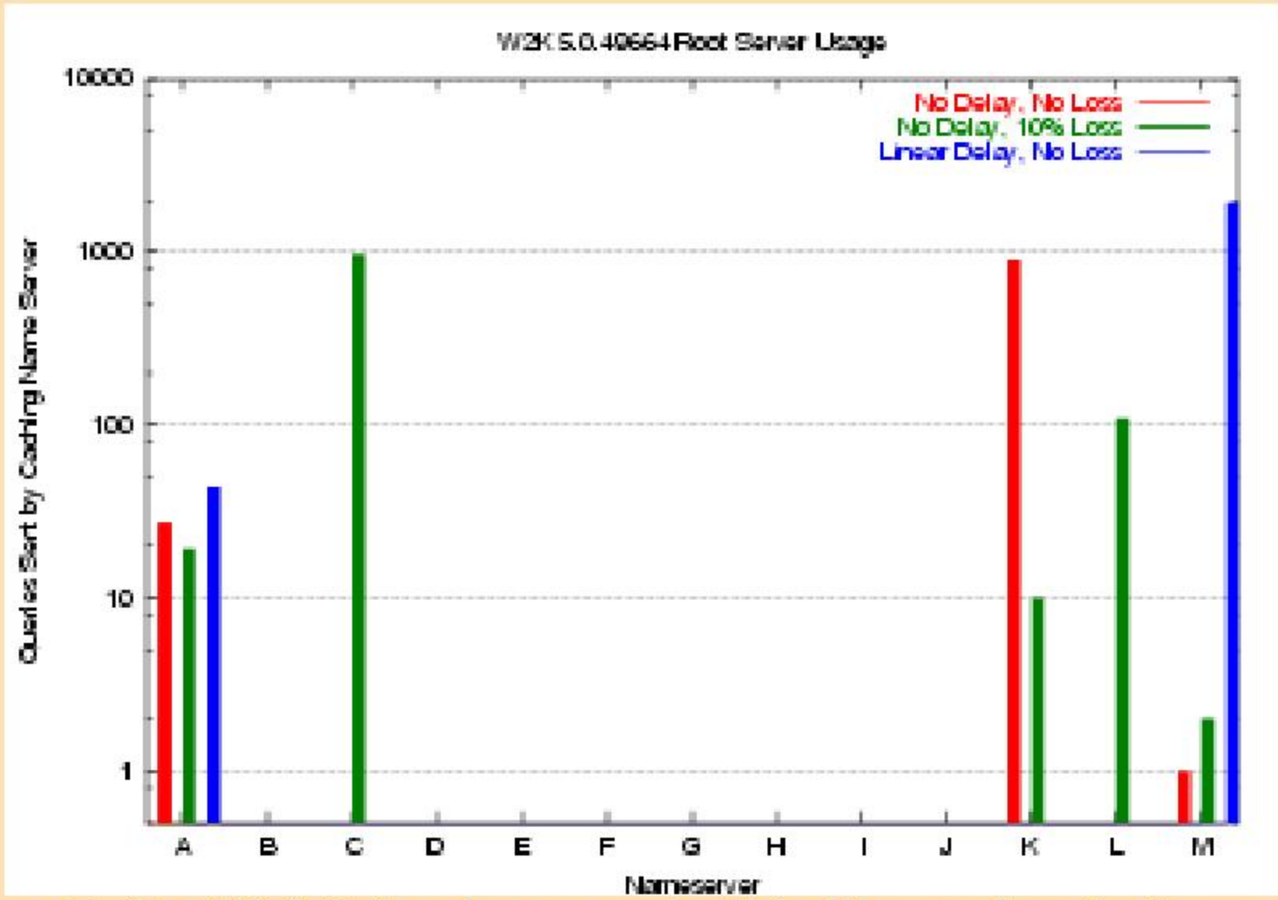


djbdns 1.05 (100M cache) uniformly distributes its queries to roots and .com TLDs in all three experiments. Increased query load in the linear delay experiment (blue) is an artifact of our simulation because the caching name server repeats queries before a response is cached. (Note that this effect does not occur in BIND9.)

djbdns 1.05 queries to .com TLDs

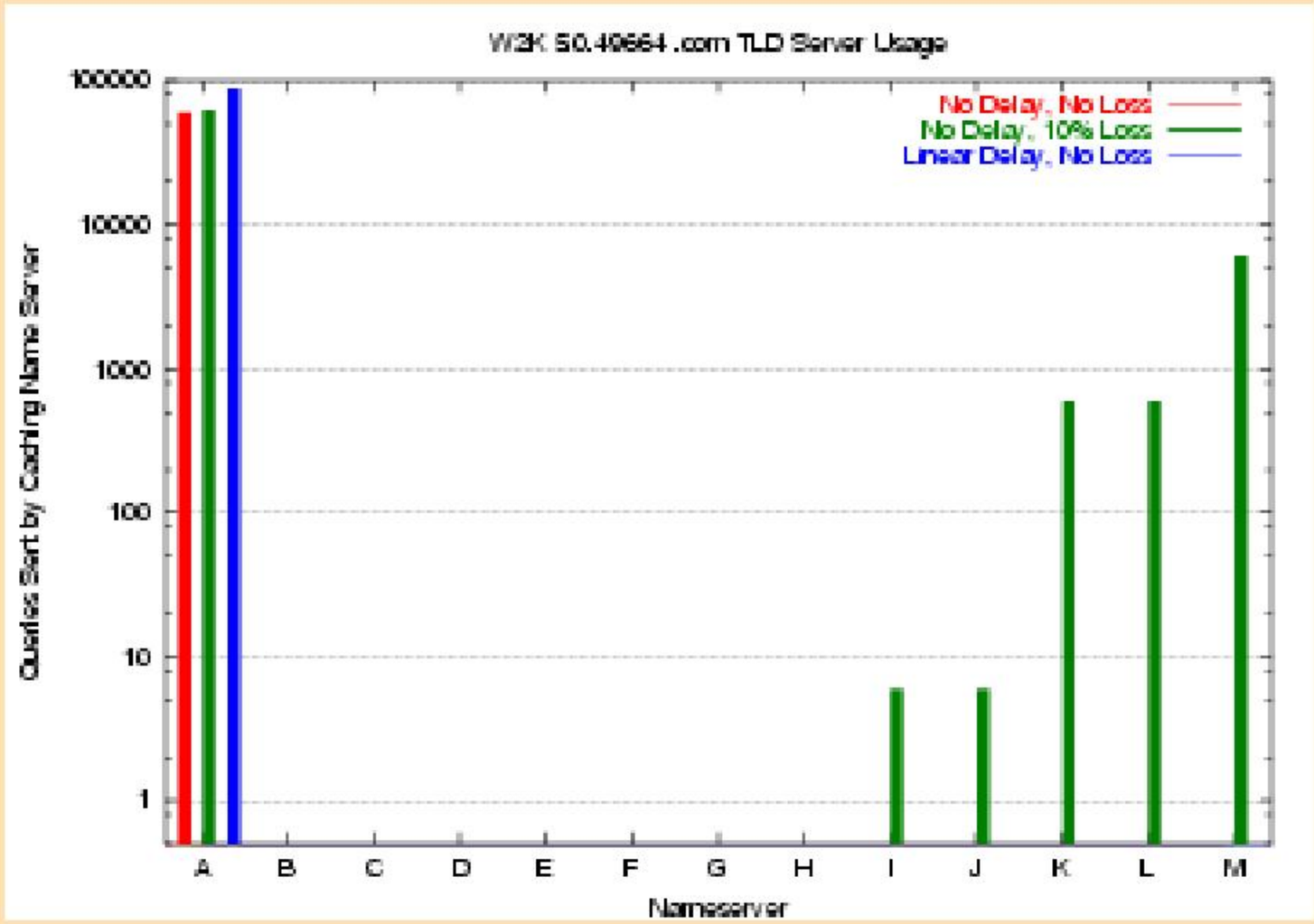


W2K v5.0.49664 queries to roots

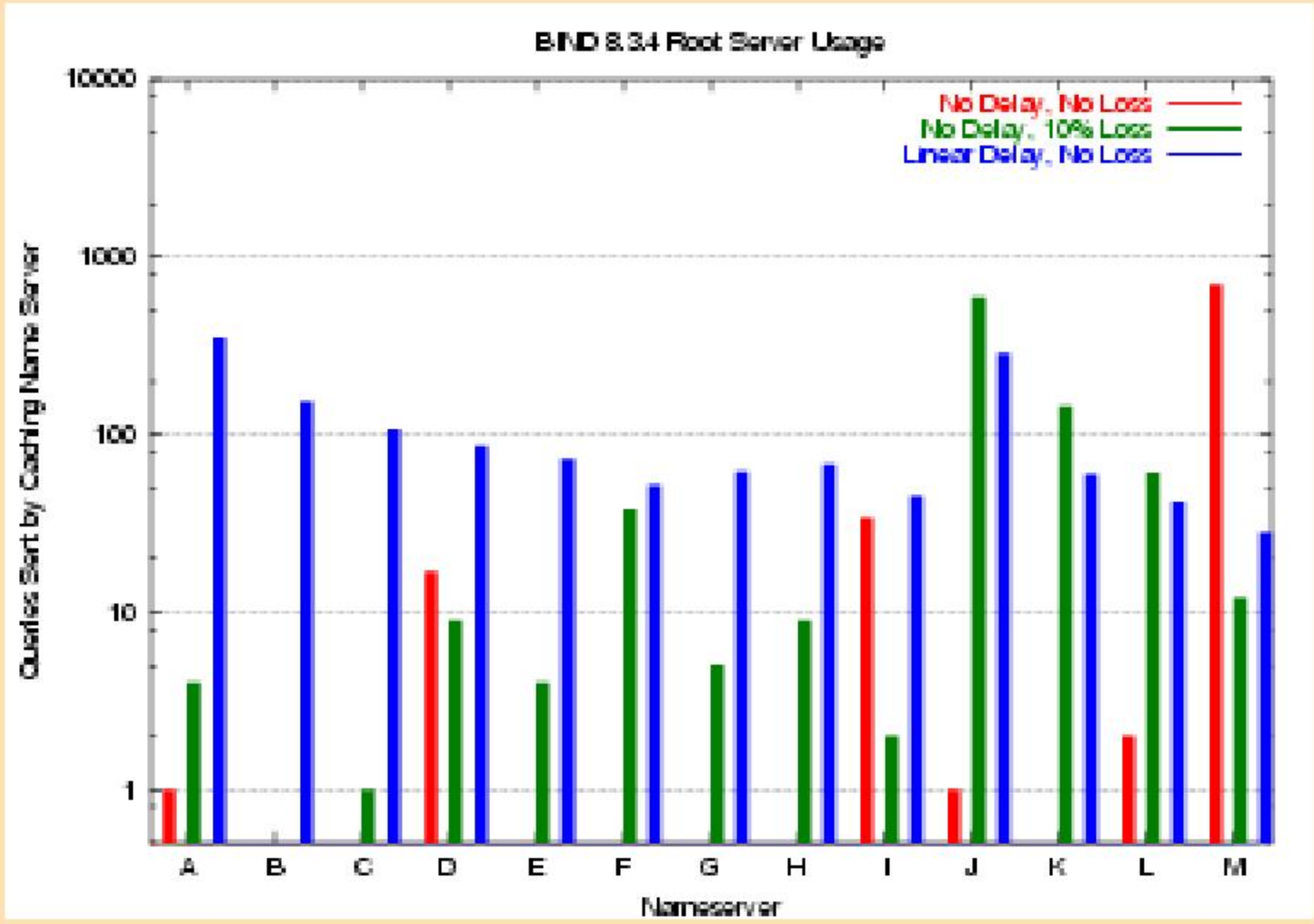


- windows W2K v5.0.49664 always initially select A root
- A few seconds later an "NS ." query is sent to M, which returns an ordered list of roots with a random starting point.
- First root in this list becomes the designated choice.
- Root selection temporarily changes only in the event of packet loss, but then returns to the designated root.

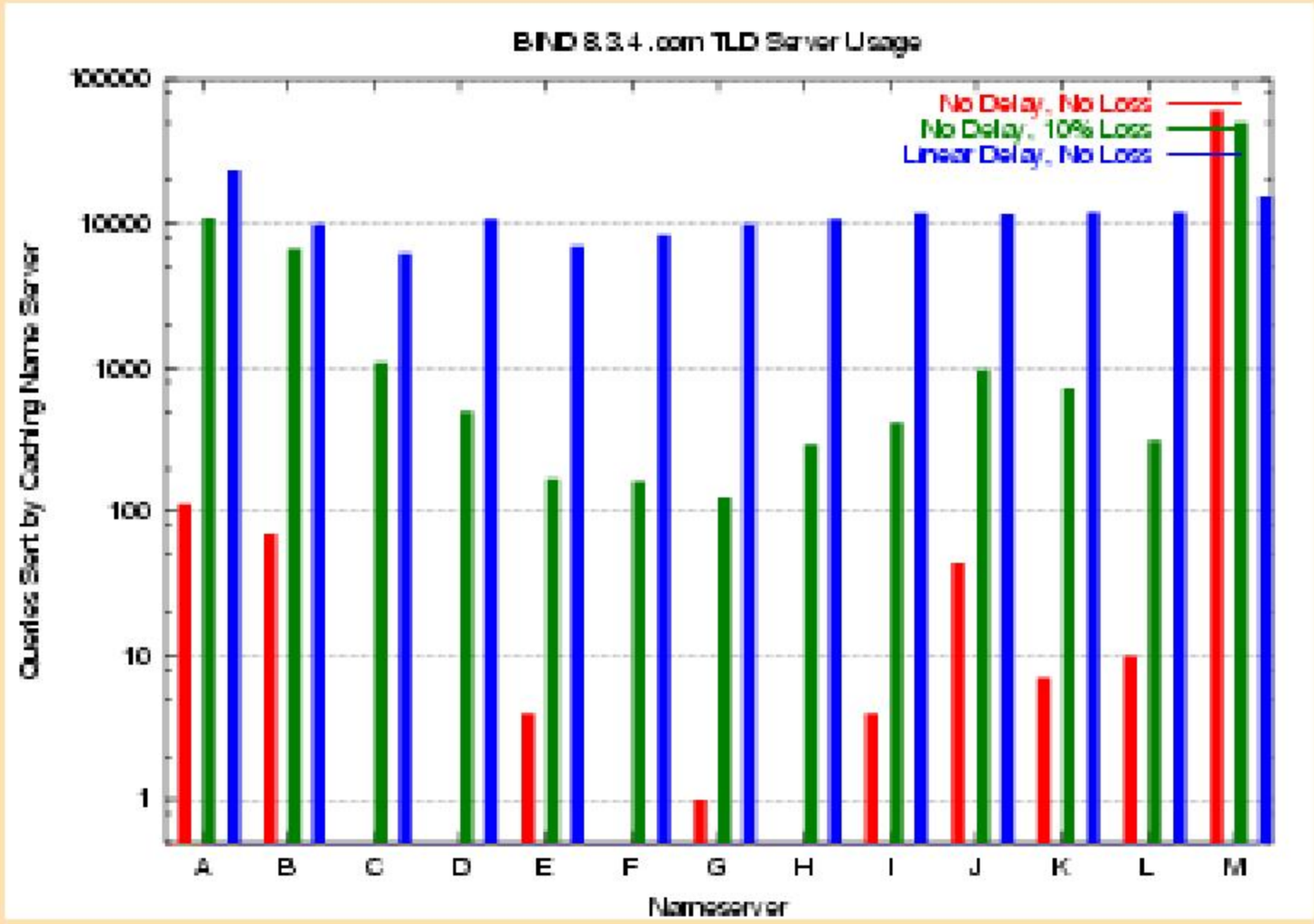
W2K v5.0.49664 queries to .com TLDs



BIND8 queries to roots



BIND8 queries to TLDs



all other plots:

- http://www.packet-pushers.net/dns/simulations/sample-plots/

contribution to national measurement priorities

challenges in Internet measurements `12-step program': step 3-5, 9-10, 12
NSF ANIR PI meeting Jan 2003
http://www.caida.org/outreach/presentations/nsfpi200301/

- (3) mathematical frameworks to find structure/patterns in traffic
  - a la scott's encouragement to `formalize some of what we (and providers) know'
  - macroscopic as well as microscopic
  - theory of joint spatial/temporal locality
  - spectroscopy, tomography

- (4) source modeling (for realistic inputs into simulations, models)
  - extract a set of source models from an aggregate trace
    - **feature extraction problem**
    - **10,000 gnutella port numbers are not 10,000 flows**
  - ultimate goal: augment libraries of source level models w generation of own
  - calibrate models by evaluating their power for prediction

- (5) empirically validated simulation of significant aspect of Internet
  - already much work in large-scale simulations, but no recognized empirically validated simulation of any signficant piece of the Internet.
  - requires cooperation from providers and vendors to get default and configured parameters of OSes and algorithms. NSF should shepard/foster this cooperation
    - **(note: large scale means in size as well as # of protocols)**

contribution to national measurement priorities

challenges in Internet measurements `12-step program': step 3-5, 9-10, 12
NSF ANIR PI meeting Jan 2003
http://www.caida.org/outreach/presentations/nsfpi200301/

- (9) discovering pervasive hidden bugs
  - any modeling or analylsis must handle impact of this huge component of traffic
- (10) how does measurement affect/support security goals
  - infer bgp, firewall, and virus spread behavior
  - how do you get networks to share security-related information
  - protection of measurement infrastructure from security compromises
- (12) encouragement of strategic measurement in new networks
  - based on what we learned from what we did wrong in old networks

contribution to national measurement priorities

payoffs of Internet measurement

- improve accuracy, validity, repeatability of network research
- provide reference points or baselines for simulation and model validation
  - other fields, e.g., architecture, have had this for years
- build a solid understanding of network behavior
  - including subtleties not otherwise detected
  - including damage not otherwise detected
- accelerate present and future modeling, simulation, and analysis efforts
  - avoid duplication of effort

// scientific apparatus offers a window to knowledge,
but as they grow more elaborate,
scientists spend ever more time washing the windows.
-- Isaac Asimov //

NMS collaborations

students visiting caida

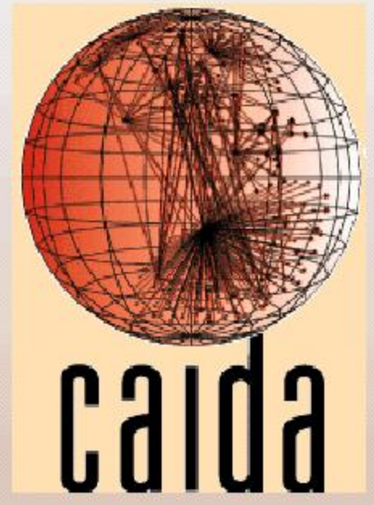
- Robert Nowak's student Ryan King (2002) (Rice)
- Michalis Faloutsos' student Thomas Karagiannis (UCR)
- Srikant Rayadurgam's student Srinivas Shakkottai (UIUC)
- Ellen Zegura's student Ruomei Gao (GaTech)
- Ken Calvert's student Aditya Namjoshi (U.KY)
- Edmond Jonckheere's student Khushboo Shah (USC)

other NMS PI users of caida data/infrastructure:

- George Riley (GaTech), John Heidemann (USC), Srikant (UIUC), Yuri Pryadkin (USC)

other relevant collaborations

- CAIDA testing of Reidi's (Rice) pathchirp for DOE bwest project
- SLAC (Les Cottrell) on LSN presentation on measurement priorities



// disorder increases with time
because we measure time in the
direction in which disorder increases //
-- stephen hawking

www.caida.org