VeriSign®

# Traffic Source Analysis of the J Root Anycast Instances

Piet Barber
Matt Larson
Mark Kosters

November 16, 2006

Where it all comes together.™

# History

- No analysis of root server anycast service had been done
  - Some questioned the deployment of anycast without testing
- "Life and Times of J-Root", NANOG 32 (October 2004)
  - Analyzed query data collected at some of the J root sites
  - Sliced and diced data in various ways; most results unsurprising
  - Two surprising findings:
    - Lots of non-priming queries to former (pre-11/2002) J root address coming from newer BIND versions
      - Interesting and still unexplained, but not this presentation's focus
    - 3.69% of all source IP addresses seen at two or more sites, **sometimes simultaneously or in quick succession**
      - We had no definitive explanation
- More presentations from others followed …

# NANOG 34 (May 05)

- "Anycast Measurements Used to Highlight Routing Instabilities" (Peter Boothe, Randy Bush)
  - Had volunteers run TCP and UDP probes
  - Conclusions:
    - Routing weirdness seems to affect anycast disproportionally
    - TCP and UDP probes had different failure rates
  - Most of the volunteers read NANOG and are presumed more clueful so the sample probes might not be representative
- "DNS Anycast Stability" (Daniel Karrenberg)
  - Looked at DNSMON data from 77 probes to various root server instances
  - Conclusions:
    - DNS availability higher with anycast
    - Saw inter-instance "switches" and linked them to routing instability
    - More research could be done on multiple paths with the same AS path length
  - Again, probes located at more clueful operators so might not be representative

# NANOG 37 (June 2006)

+ "Effects of Anycast on K-root Performance" (Lorenzo Coletti)
  - Analysis included stability and its effect on performance
  - Findings:
    - April 2005, 24 hours from two global nodes: 1.1% of unique source IPs switched sites
    - April 2006, 5 hours from five global nodes: 0.33% of unique source IPs switched sites
  - Is five hours enough data?
+ "Operational Experience with TCP and Anycast" (Matt Levine, Barrett Lyon and Todd Underwood)
  - Conclusion:
    - No problem with TCP and anycast
  - Attributes:
    - Large bandwidth provider
    - Carefully engineered their peering
    - Clients used high-bandwidth connections
  - Different environment than DNS root service?

# Why This Presentation?

+ Others' work on this topic has been tremendous: we are not throwing stones!
+ Our initial data was disturbing and we still didn't think we had seen an explanation for what we observed
+ Wanted to do our own follow-up analysis with more data

# The Data

+ *j.root-servers.net*: 19 active anycast instances for this analysis
  - All sourced from AS 26415
  - Mix of “global” and “local” instances
    - Local here means peering-only and upstream route filters
  - Mix of transit and peering
+ Two separate packet captures
  - UDP
    - Inbound only
  - TCP
    - Bi-directional
+ Each just over 24 hours from different, non-overlapping time periods
+ Every active anycast instance represented (!)

# Goals: Questions to Answer

- What kind of distribution of source IP addresses across all instances do we see?
- Can we explain what we see?
- Can we determine if what we see is causing a problem?

# UDP Packet Capture

- About 26 hours of inbound query data:
  - Start: around 1145 EDT 25 October 2006
  - Stop: around 1345 EDT 26 October 2006
  - Staggered start/stop for packet capture at each site
- High-level statistics:
  - **335,259,322** total queries
    - Around **3,500** queries per second
    - UDP only, 24-hour average
  - **859,784** unique source IP addresses
    - Ratio of **390** queries per unique IP address
  - **5,561** source IP addresses appear at more than one site
    - That's **0.646%** of total unique source IP addresses (1 out of every 154)
    - Down significantly from 3.69% in mid-2004 ("Life and Times of J-Root")

# Table Columns Explained

- **Source IP:** An IP address that sent a UDP DNS query to one of the *j.root-servers.net* instances
- **Transitions:** Number of times this source IP address switched instances
  - All measurements have one-second resolution
- **Sites:** Number of sites/instances this source IP address appeared at/sent queries to
- **Simultaneous Seconds:** The number of seconds in the packet capture interval when multiple instances received simultaneous queries from this IP address
- **Total Seconds:** The total number of seconds that any of the instances received traffic from this IP address
- **Percent Simultaneous:** (Simultaneous Seconds / Total Seconds) * 100
- **Specific Sites:** Which anycast instances received traffic and for how many seconds each
  - Instance abbreviation magic decoder ring at right

| | |
|---|---|
| **A** | Dulles, VA |
| **C** | Dulles, VA |
| **E** | Los Angeles |
| **F** | Seattle |
| **I** | Stockholm |
| **J** | Tokyo |
| **L** | Atlanta |
| **M** | Singapore |
| **N** | Amsterdam |
| **O** | Miami |
| **P** | Seoul |
| **Q** | Brasilia |
| **R** | Cairo |
| **S** | Dublin |
| **T** | Dulles, VA |
| **U** | Dulles, VA |
| **V** | Sao Paulo |
| **W** | Mountain View |
| **X** | Sydney |

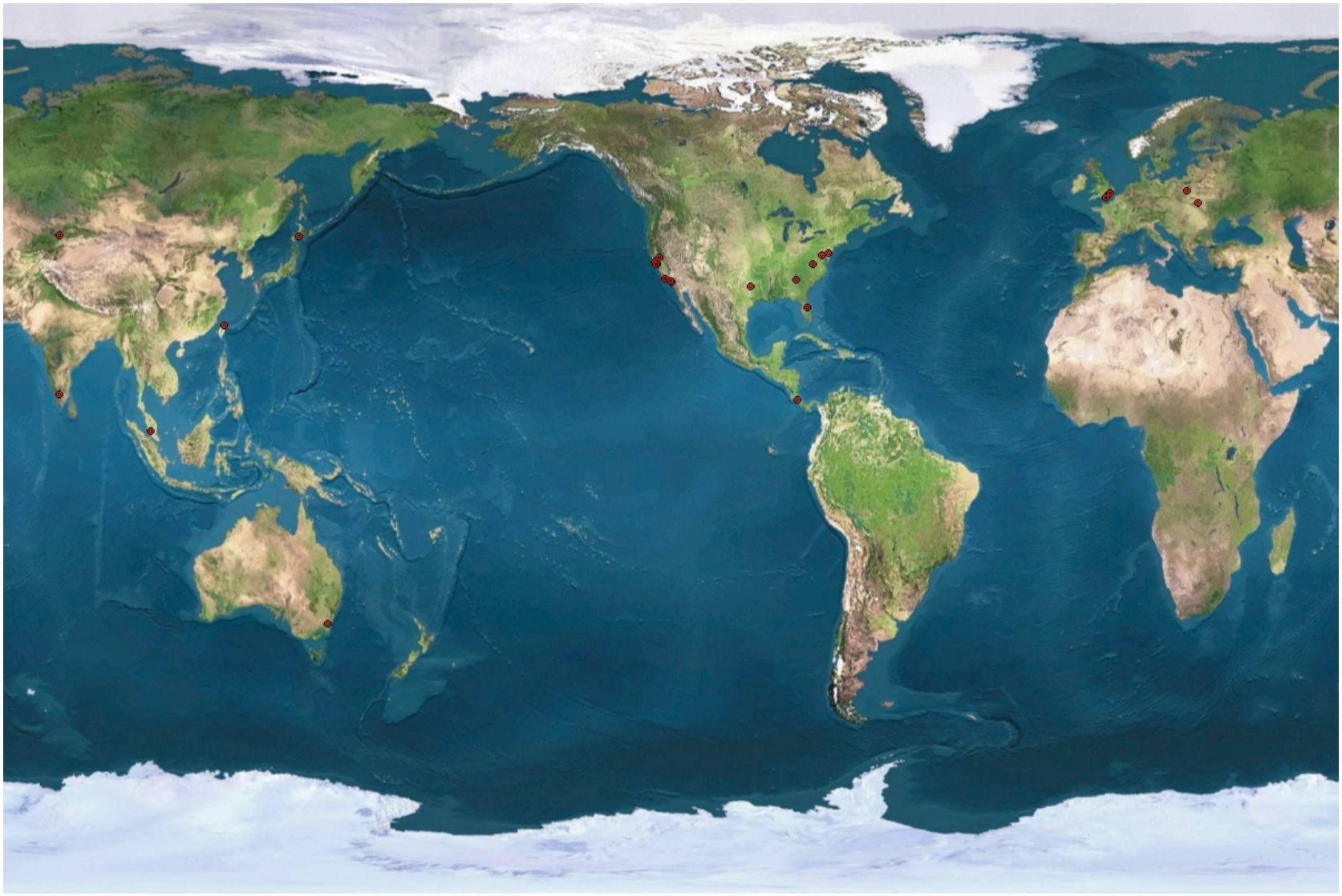
# A Difficult Decision

- "The IP addresses you are about to see are true. The addresses have not been changed to protect the innocent."
- We decided not to anonymize source IP addresses
- We are **not** trying to point out anyone's problems, shame anyone, etc.
- Rather, we are asking for the community's help to get to the bottom of the behavior we see
- Do you see an IP you recognize and know what's up?

# Top 30 Source IPs, Descending % Simul. Sites

| | Source IP | Trans. | Sites | Simul. Secs. | Total Secs. | **Percent Simul.** | Specific Sites |
|---|---|---|---|---|---|---|---|
| 1 | 202.45.133.9 | 1 | 2 | 1 | 2 | 50.00% | M=1 P=1 |
| 2 | 136.235.14.3 | 1 | 2 | 1 | 4 | 25.00% | I=1 J=3 |
| 3 | 207.243.192.15 | 406 | 2 | 158 | 678 | 23.30% | J=375 W=303 |
| 4 | 87.247.18.58 | 3 | 2 | 1 | 5 | 20.00% | P=3 T=2 |
| 5 | 147.62.42.119 | 2 | 2 | 1 | 6 | 16.67% | O=5 P=1 |
| 6 | 207.154.101.48 | 3 | 2 | 2 | 14 | 14.29% | T=5 W=9 |
| 7 | 24.248.93.71 | 2 | 2 | 1 | 7 | 14.29% | P=4 T=3 |
| 8 | 202.125.10.40 | 28 | 2 | 14 | 109 | 12.84% | J=72 P=37 |
| 9 | 202.45.133.3 | 10 | 2 | 5 | 39 | 12.82% | M=29 P=10 |
| 10 | 204.80.216.69 | 3752 | 2 | 1248 | 9739 | 12.81% | N=5763 T=3976 |
| 11 | 220.128.207.66 | 2 | 2 | 1 | 10 | 10.00% | J=9 M=1 |
| 12 | 66.180.96.10 | 13492 | 2 | 5350 | 54790 | 9.77% | L=37849 P=16941 |
| 13 | 207.154.75.57 | 34 | 2 | 16 | 172 | 9.30% | T=30 W=142 |
| 14 | 66.125.97.2 | 2 | 2 | 1 | 12 | 8.33% | P=10 T=2 |
| 15 | 209.4.229.202 | 6 | 2 | 2 | 31 | 6.45% | L=26 T=5 |
| 16 | 204.130.244.134 | 4 | 2 | 2 | 34 | 5.88% | N=17 W=17 |
| 17 | 220.225.140.98 | 33 | 2 | 4 | 70 | 5.71% | P=31 T=39 |
| 18 | 64.60.0.17 | 7276 | 2 | 2740 | 56345 | 4.86% | N=13317 T=43028 |
| 19 | 193.115.14.67 | 4 | 2 | 1 | 21 | 4.76% | J=3 N=18 |
| 20 | 210.156.102.90 | 23 | 2 | 4 | 86 | 4.65% | J=64 P=22 |
| 21 | 66.184.164.100 | 19 | 2 | 3 | 65 | 4.62% | L=53 P=12 |
| 22 | 217.9.0.250 | 74 | 2 | 8 | 220 | 3.64% | N=139 P=81 |
| 23 | 136.235.12.27 | 5 | 2 | 3 | 87 | 3.45% | I=5 J=82 |
| 24 | 209.4.229.201 | 4 | 2 | 1 | 30 | 3.33% | L=26 T=4 |
| 25 | 63.103.50.15 | 699 | 2 | 54 | 1915 | 2.82% | P=1188 T=727 |
| 26 | 136.235.12.17 | 4 | 2 | 2 | 77 | 2.60% | I=5 J=72 |
| 27 | 193.28.226.2 | 37 | 2 | 3 | 116 | 2.59% | I=63 N=53 |
| 28 | 198.6.248.11 | 6 | 2 | 1 | 44 | 2.27% | N=27 T=17 |
| 29 | 201.206.0.18 | 10 | 2 | 1 | 49 | 2.04% | Q=39 T=10 |
| 30 | 193.122.27.34 | 10 | 2 | 5 | 246 | 2.03% | J=14 N=232 |

# Top 30 Most % Simultaneous Geographic Distribution





VeriSign®

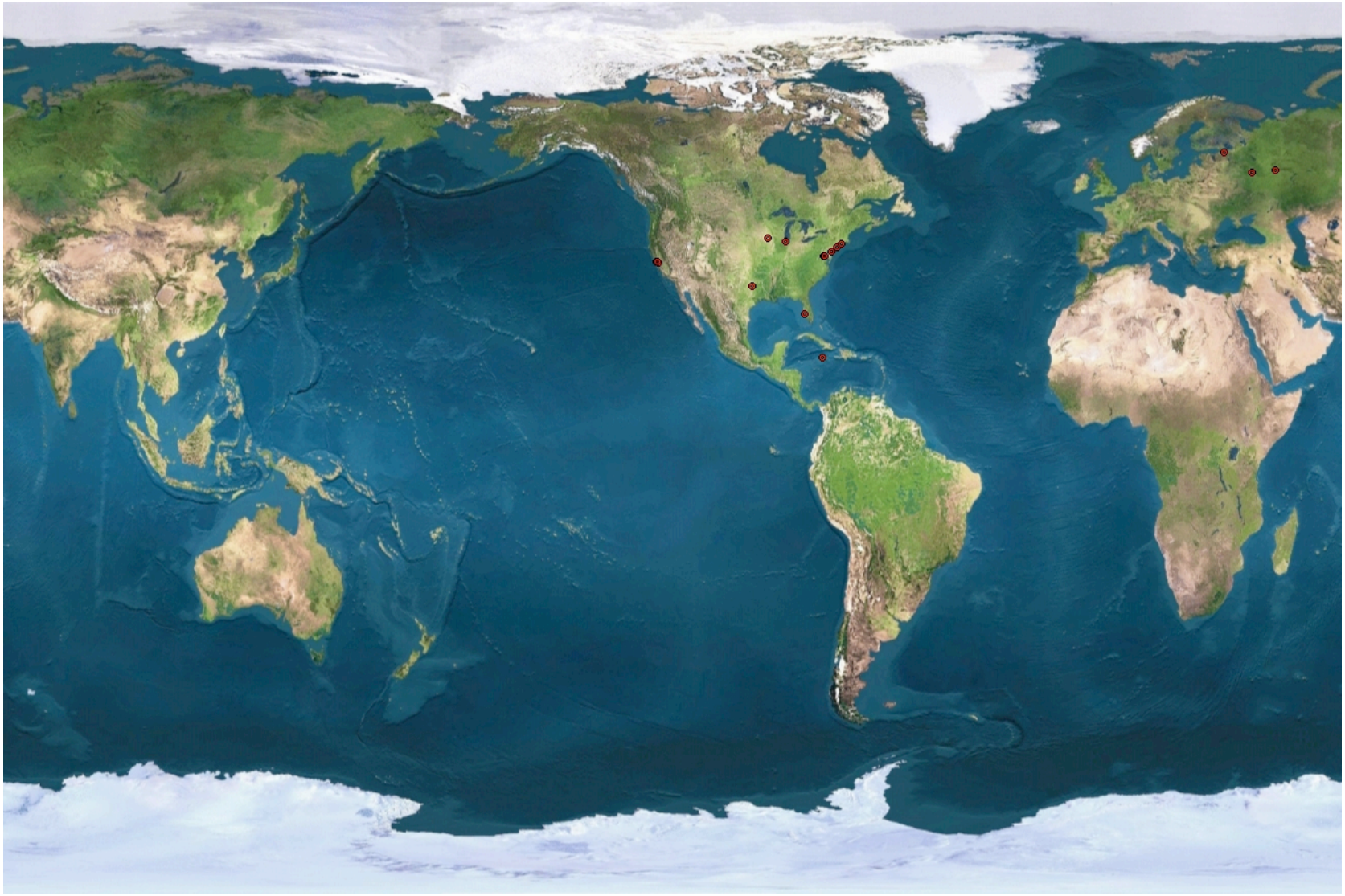
# Top 30 Source IPs, Descending Site Order

| | Source IP | Trans. | **Sites** | Simul. Secs. | Total Secs. | Percent Simul. | Specific Sites |
|---|---|---|---|---|---|---|---|
| 1 | 216.175.203.50 | 93 | 4 | 20 | 10277 | 0.20% | J=15 N=1 P=3305 T=6956 |
| 2 | 216.218.195.242 | 8 | 4 | 0 | 1065 | 0.00% | J=219 M=2 N=1 P=843 |
| 3 | 65.19.182.2 | 8 | 4 | 0 | 1012 | 0.00% | J=309 M=25 N=33 P=645 |
| 4 | 65.19.183.170 | 8 | 4 | 0 | 3455 | 0.00% | J=658 M=14 N=60 P=2723 |
| 5 | 209.51.187.132 | 7 | 4 | 0 | 4380 | 0.00% | J=1883 M=1 N=20 P=2476 |
| 6 | 64.38.5.242 | 7 | 4 | 0 | 6715 | 0.00% | J=1316 M=13 N=32 P=5354 |
| 7 | 209.8.109.153 | 6 | 4 | 0 | 850 | 0.00% | J=305 M=2 N=6 P=537 |
| 8 | 192.104.109.190 | 5 | 4 | 0 | 658 | 0.00% | J=160 M=1 N=1 P=496 |
| 9 | 216.218.253.190 | 5 | 4 | 0 | 732 | 0.00% | J=128 M=2 N=2 P=600 |
| 10 | 66.160.172.128 | 5 | 4 | 0 | 505 | 0.00% | J=129 M=2 N=1 P=373 |
| 11 | 66.228.118.71 | 5 | 4 | 0 | 29333 | 0.00% | J=4 P=24 Q=29304 T=1 |
| 12 | 194.186.252.34 | 4 | 4 | 0 | 36824 | 0.00% | I=2 N=13 O=246 T=36563 |
| 13 | 194.186.88.2 | 4 | 4 | 0 | 436 | 0.00% | I=1 N=1 O=10 T=424 |
| 14 | 195.239.16.228 | 4 | 4 | 0 | 1460 | 0.00% | I=1 N=8 O=77 T=1374 |
| 15 | 195.68.135.5 | 4 | 4 | 0 | 2952 | 0.00% | I=29 N=8 O=61 T=2854 |
| 16 | 212.44.130.6 | 4 | 4 | 0 | 12090 | 0.00% | I=54 N=47 O=284 T=11705 |
| 17 | 64.62.197.115 | 4 | 4 | 0 | 14 | 0.00% | J=4 M=2 N=6 P=2 |
| 18 | 194.67.21.177 | 3 | 4 | 0 | 7746 | 0.00% | I=37 N=27 O=156 T=7526 |
| 19 | 194.67.21.178 | 3 | 4 | 0 | 7748 | 0.00% | I=42 N=36 O=167 T=7503 |
| 20 | 194.67.21.179 | 3 | 4 | 0 | 7896 | 0.00% | I=45 N=40 O=195 T=7616 |
| 21 | 194.85.128.10 | 3 | 4 | 0 | 15766 | 0.00% | I=13 N=22 O=234 T=15497 |
| 22 | 195.28.32.3 | 3 | 4 | 0 | 396 | 0.00% | I=5 N=2 O=9 T=380 |
| 23 | 195.98.32.193 | 3 | 4 | 0 | 582 | 0.00% | I=3 N=3 O=7 T=569 |
| 24 | 213.33.206.146 | 3 | 4 | 0 | 1116 | 0.00% | I=1 N=15 O=88 T=1012 |
| 25 | 217.78.177.250 | 3 | 4 | 0 | 6706 | 0.00% | I=36 N=16 O=279 T=6375 |
| 26 | 64.51.133.2 | 491 | 3 | 190 | 17720 | 1.07% | J=45 P=13340 T=4335 |
| 27 | 208.163.52.115 | 82 | 3 | 10 | 18464 | 0.05% | O=38 Q=18199 T=227 |
| 28 | 217.112.37.10 | 46 | 3 | 21 | 530391 | 0.00% | I=189 N=168 O=530034 |
| 29 | 217.112.42.15 | 46 | 3 | 21 | 311599 | 0.01% | I=1524 N=159 O=309916 |
| 30 | 209.87.79.232 | 45 | 3 | 10 | 8276 | 0.12% | J=2 P=2402 T=5872 |



VeriSign®

# Top 30 Most Sites Geographic Distribution



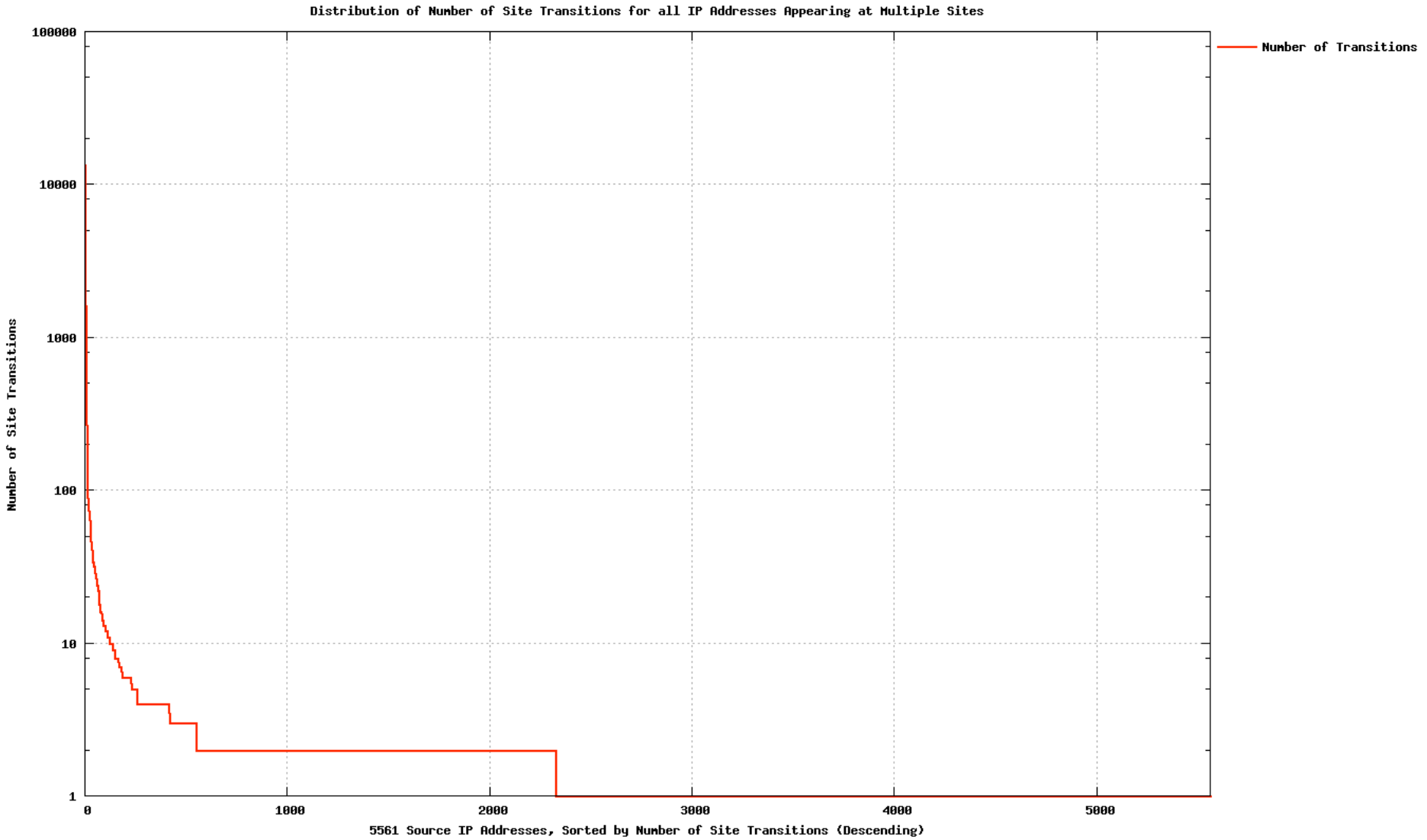
# Top 30 Source IPs, Descending Transition Order

| | Source IP | **Trans.** | Sites | Simul. Secs. | Total Secs. | Percent Simul. | Specific Sites |
|---|---|---|---|---|---|---|---|
| 1 | 66.180.96.10 | 13492 | 2 | 5350 | 54790 | 9.77% | L=37849 P=16941 |
| 2 | 64.60.0.17 | 7276 | 2 | 2740 | 56345 | 4.86% | N=13317 T=43028 |
| 3 | 204.80.216.69 | 3752 | 2 | 1248 | 9739 | 12.81% | N=5763 T=3976 |
| 4 | 63.103.50.15 | 699 | 2 | 54 | 1915 | 2.82% | P=1188 T=727 |
| 5 | 217.17.48.1 | 648 | 2 | 63 | 4544 | 1.39% | N=2932 P=1612 |
| 6 | 209.159.192.7 | 519 | 2 | 24 | 5437 | 0.44% | L=4845 T=592 |
| 7 | 64.51.133.2 | 491 | 3 | 190 | 17720 | 1.07% | J=45 P=13340 T=4335 |
| 8 | 207.243.192.15 | 406 | 2 | 158 | 678 | 23.30% | J=375 W=303 |
| 9 | 216.47.210.6 | 172 | 2 | 40 | 62350 | 0.06% | L=59198 T=3152 |
| 10 | 199.125.12.1 | 154 | 2 | 0 | 1537 | 0.00% | L=1152 T=385 |
| 11 | 64.18.100.11 | 140 | 2 | 5 | 1944 | 0.26% | L=1566 T=378 |
| 12 | 64.18.100.17 | 108 | 2 | 3 | 1389 | 0.22% | L=1107 T=282 |
| 13 | 206.108.60.11 | 96 | 2 | 0 | 317 | 0.00% | P=95 Q=222 |
| 14 | 216.175.203.50 | 93 | 4 | 20 | 10277 | 0.20% | J=15 N=1 P=3305 T=6956 |
| 15 | 216.183.68.111 | 84 | 2 | 32 | 90099 | 0.04% | P=85839 T=4260 |
| 16 | 208.163.52.115 | 82 | 3 | 10 | 18464 | 0.05% | O=38 Q=18199 T=227 |
| 17 | 62.123.17.67 | 82 | 2 | 0 | 370 | 0.00% | I=286 T=84 |
| 18 | 199.223.36.10 | 75 | 2 | 0 | 391 | 0.00% | P=247 T=144 |
| 19 | 217.9.0.250 | 74 | 2 | 8 | 220 | 3.64% | N=139 P=81 |
| 20 | 199.125.13.1 | 72 | 2 | 0 | 731 | 0.00% | L=603 T=128 |
| 21 | 200.38.100.210 | 66 | 2 | 0 | 407 | 0.00% | L=320 T=87 |
| 22 | 204.86.34.1 | 65 | 2 | 0 | 46758 | 0.00% | L=3842 W=42916 |
| 23 | 32.97.110.142 | 65 | 2 | 1 | 77043 | 0.00% | L=4802 W=72241 |
| 24 | 200.38.96.10 | 64 | 2 | 2 | 562 | 0.36% | L=426 T=136 |
| 25 | 204.96.181.75 | 62 | 2 | 0 | 870 | 0.00% | L=627 Q=243 |
| 26 | 207.154.65.10 | 53 | 2 | 0 | 243 | 0.00% | T=62 W=181 |
| 27 | 64.113.160.162 | 52 | 2 | 0 | 1053 | 0.00% | T=93 W=960 |
| 28 | 129.42.4.117 | 49 | 2 | 0 | 1516 | 0.00% | L=102 W=1414 |
| 29 | 216.171.238.66 | 47 | 2 | 0 | 220 | 0.00% | P=57 T=163 |
| 30 | 217.112.37.10 | 46 | 3 | 21 | 530391 | 0.00% | I=189 N=168 O=530034 |

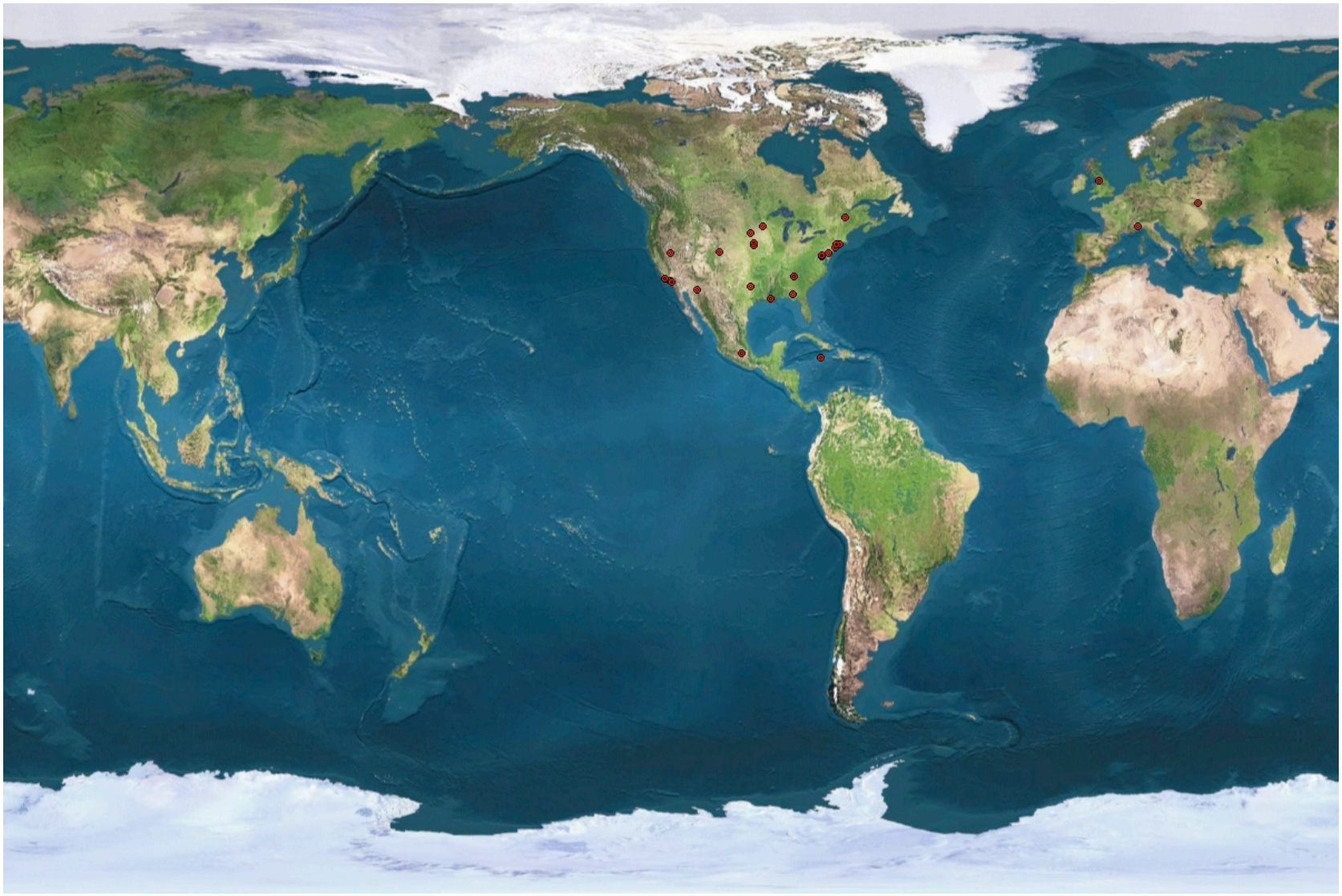


VeriSign®

# Transitions



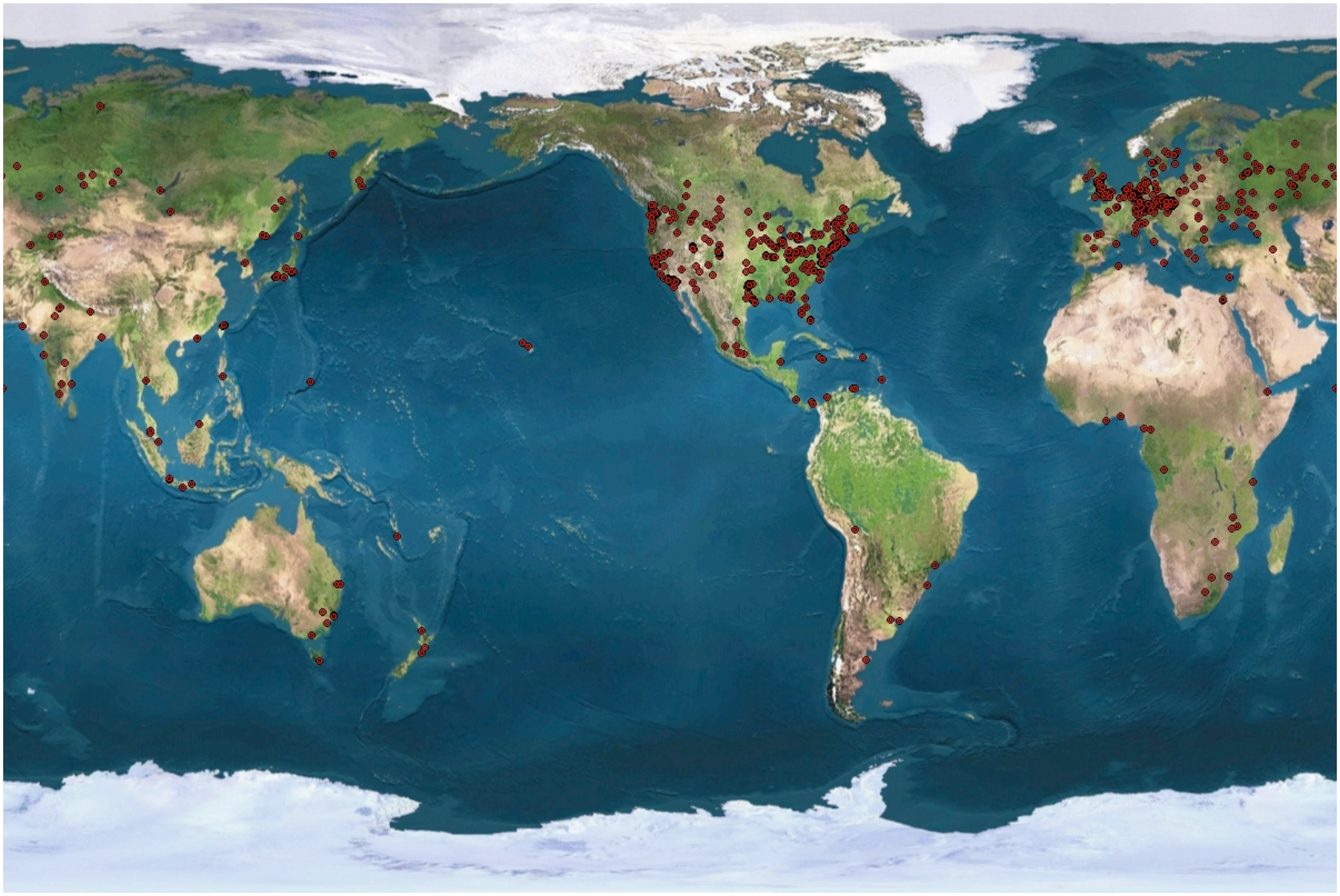
# Top 30 Most Transitions Geographic Distribution





VeriSign®

# Geographic Distribution of All Multi-site IPs





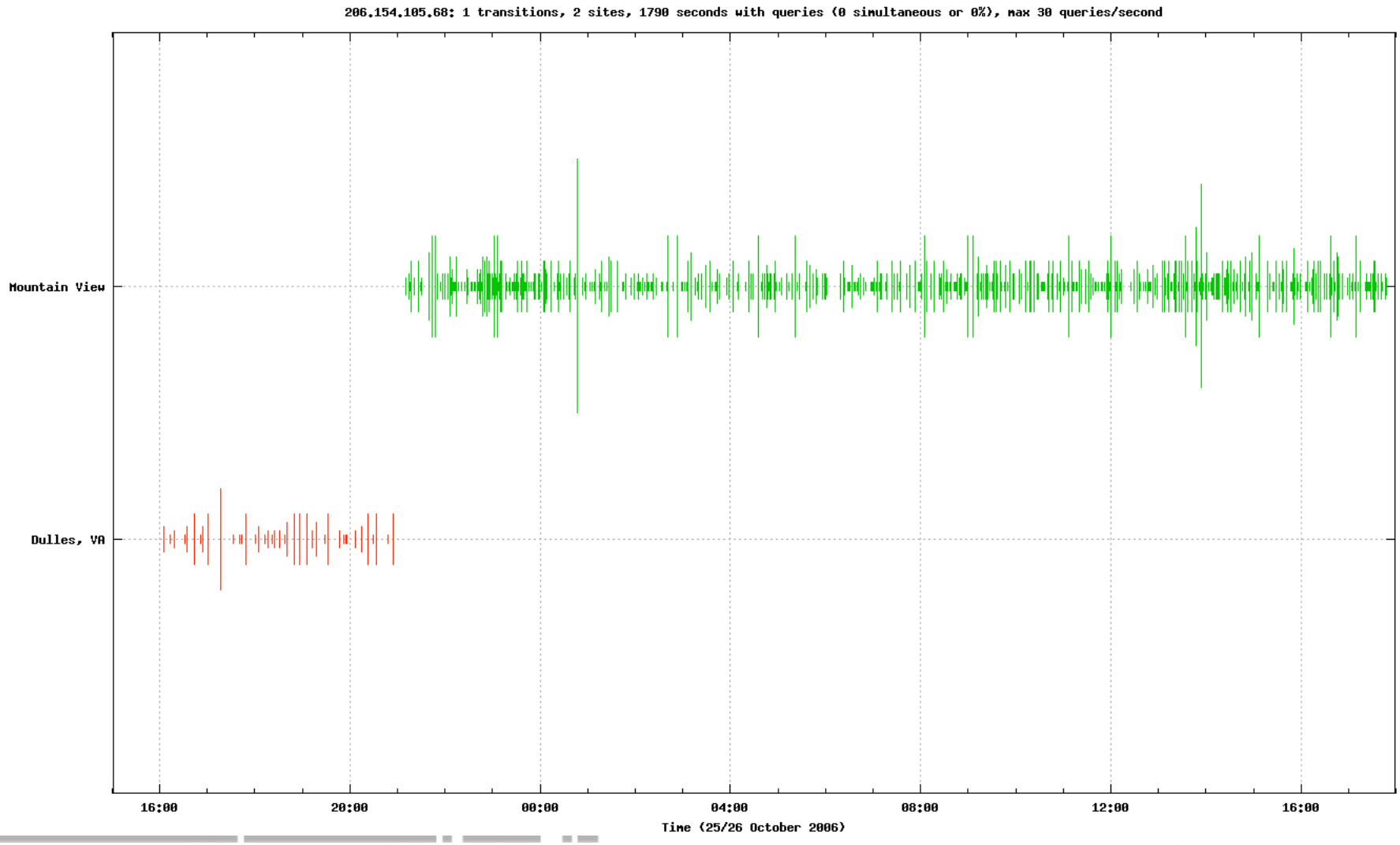
VeriSign®

# Graphs of “Interesting” Query Source IP Addresses

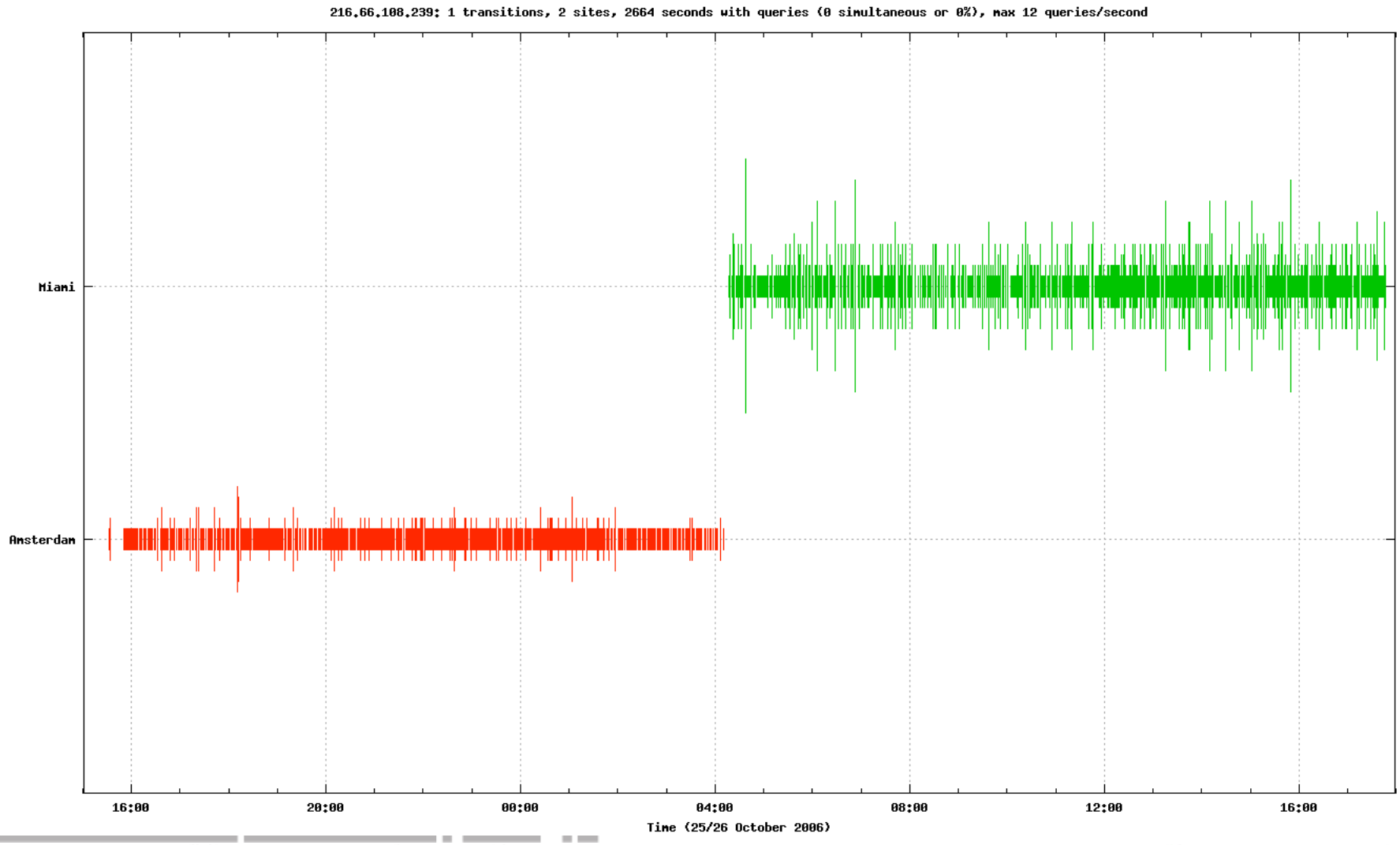
+ Plotted queries over time for “interesting” source IP addresses
+ Only sites/instances receiving queries from a given IP address are shown on its graph
+ Amplitude indicates query volume at that instance

+ Two graphs showing expected switching behavior, followed by several interesting ones

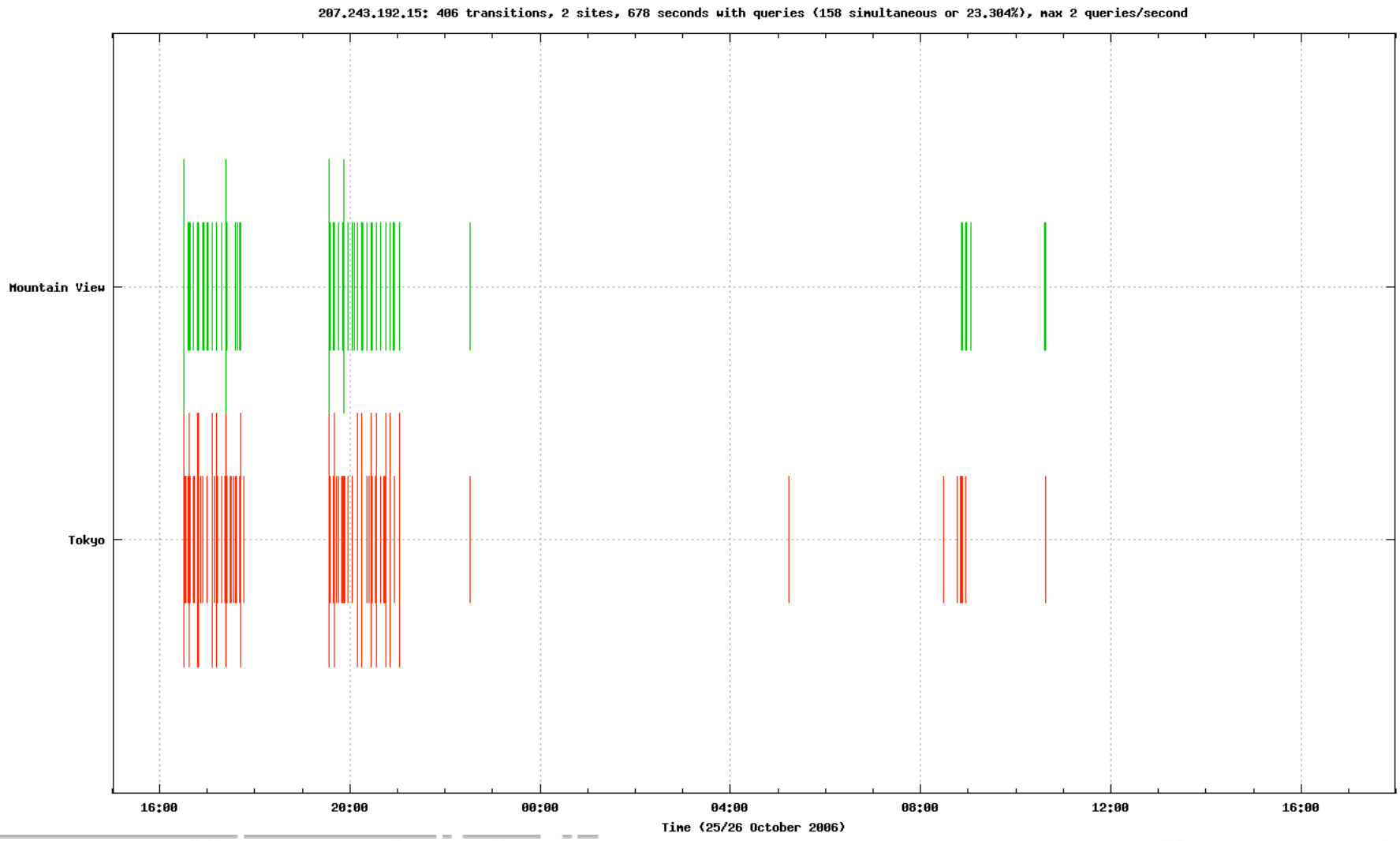
# An Expected Transition, Graph #1



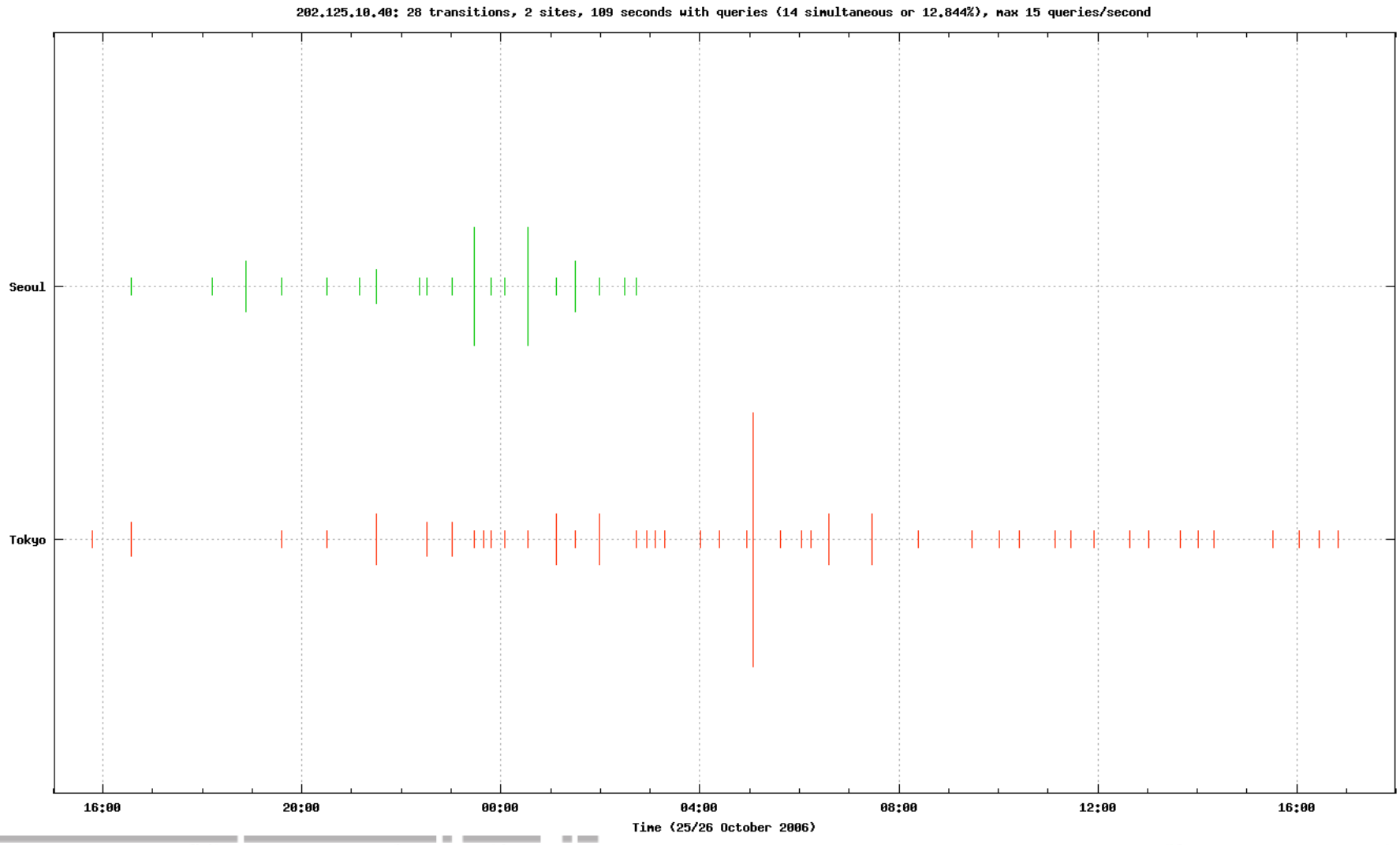
# An Expected Transition, Graph #2



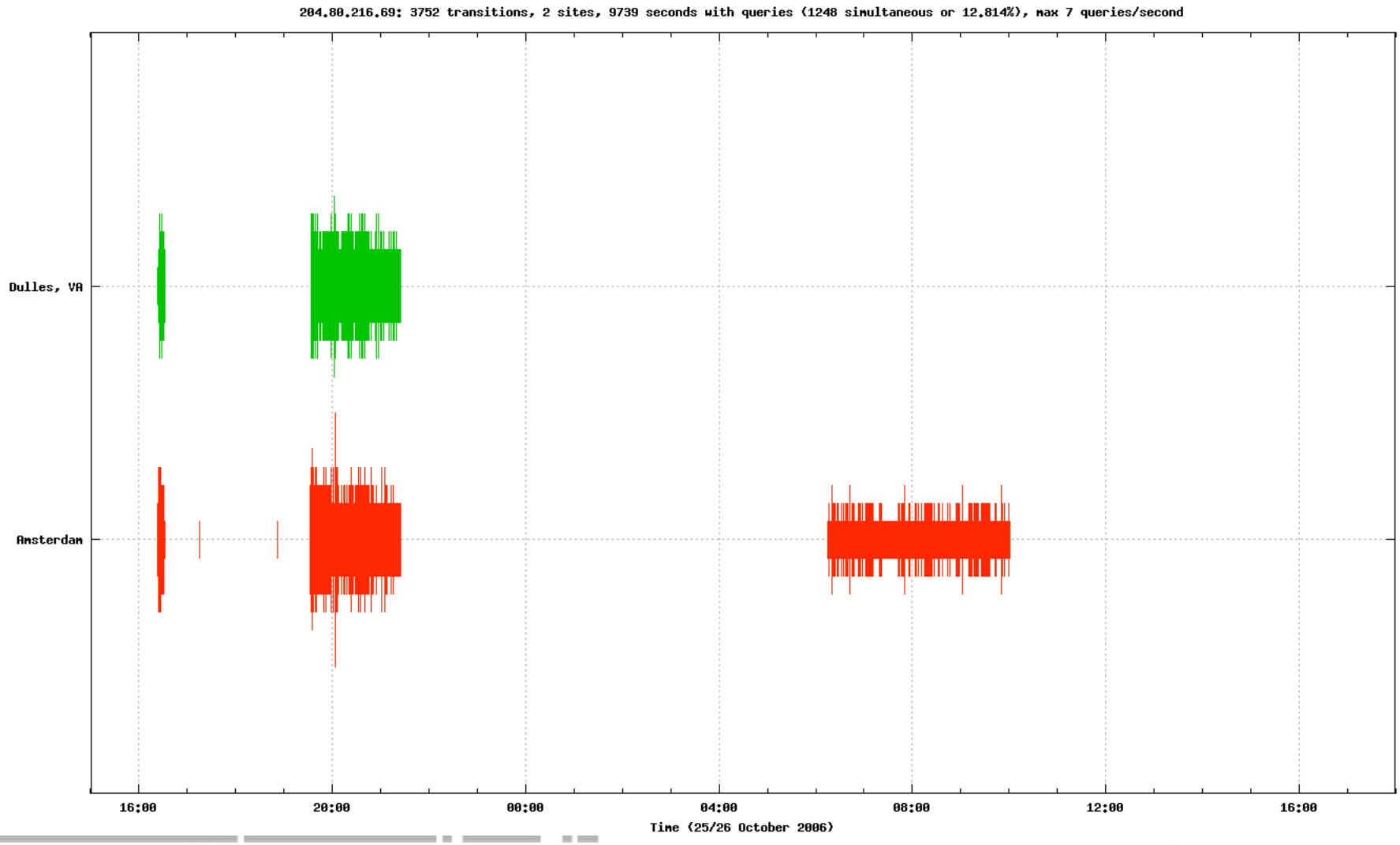
# Simultaneous Sites, Graph #1 (3rd on Top 30 List)



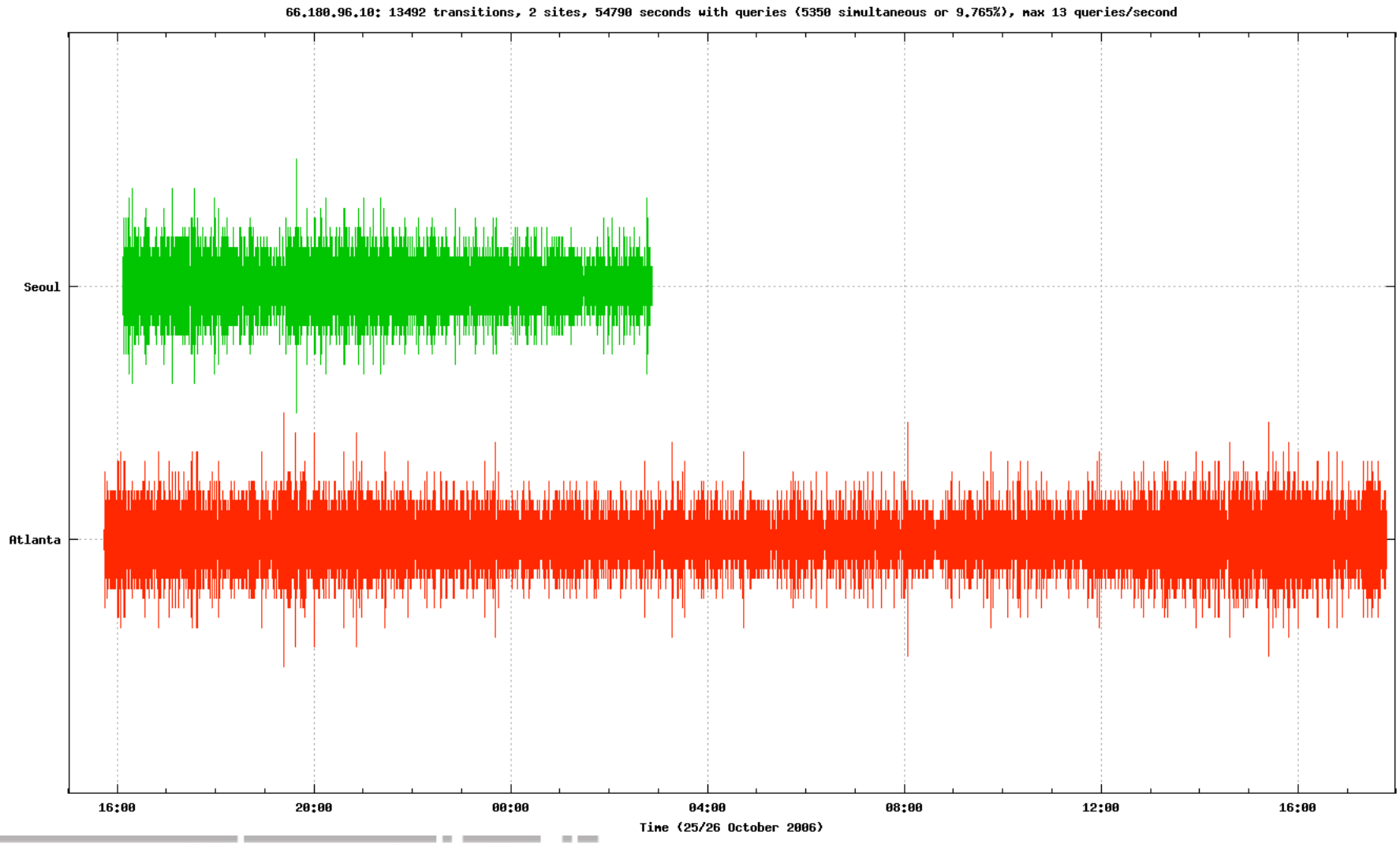
# Simultaneous Sites, Graph #2 (8th on Top 30 List)



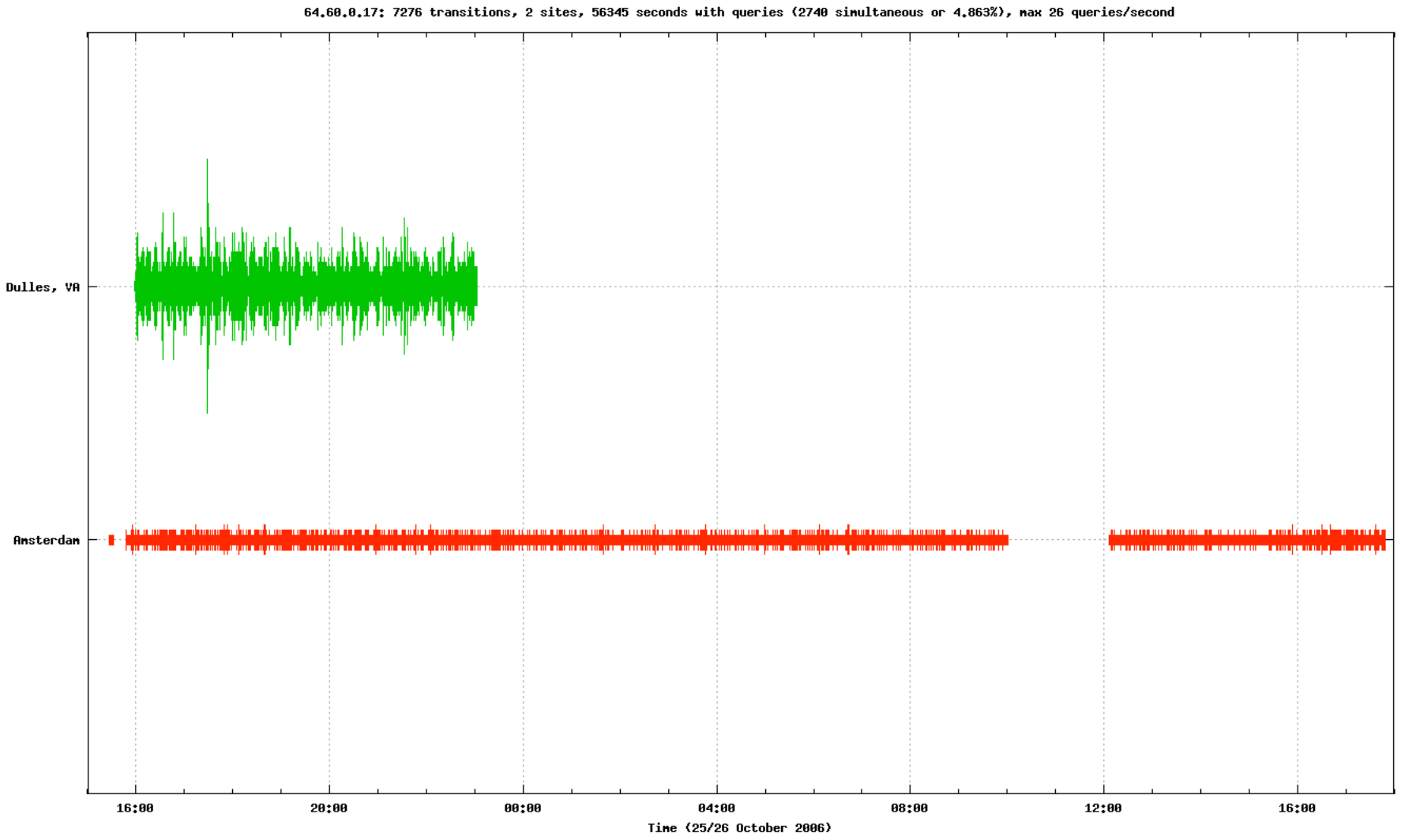
# Simultaneous Sites, Graph #3 (10th on Top 30 List)



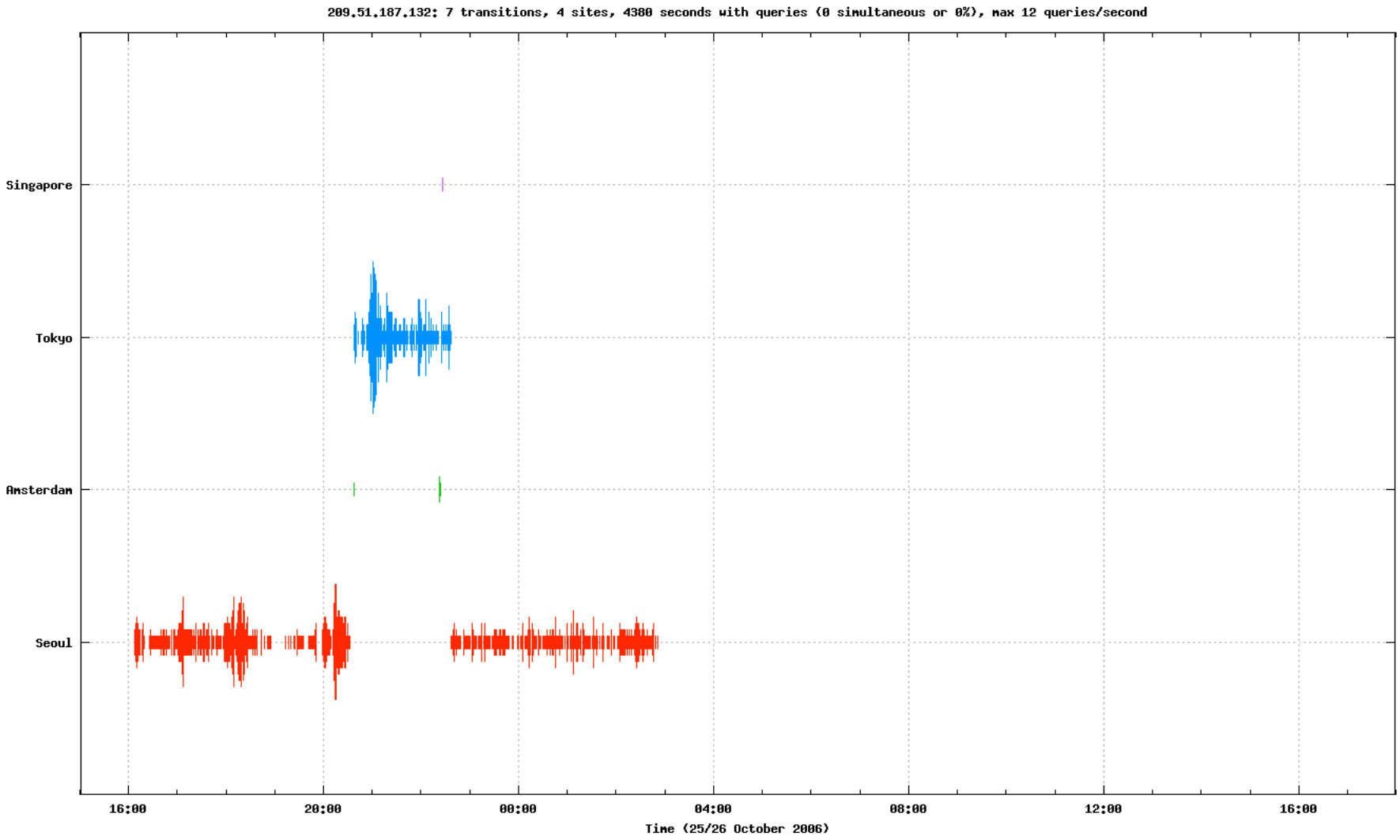
# Simultaneous Sites, Graph #4 (12th on Top 30 List)



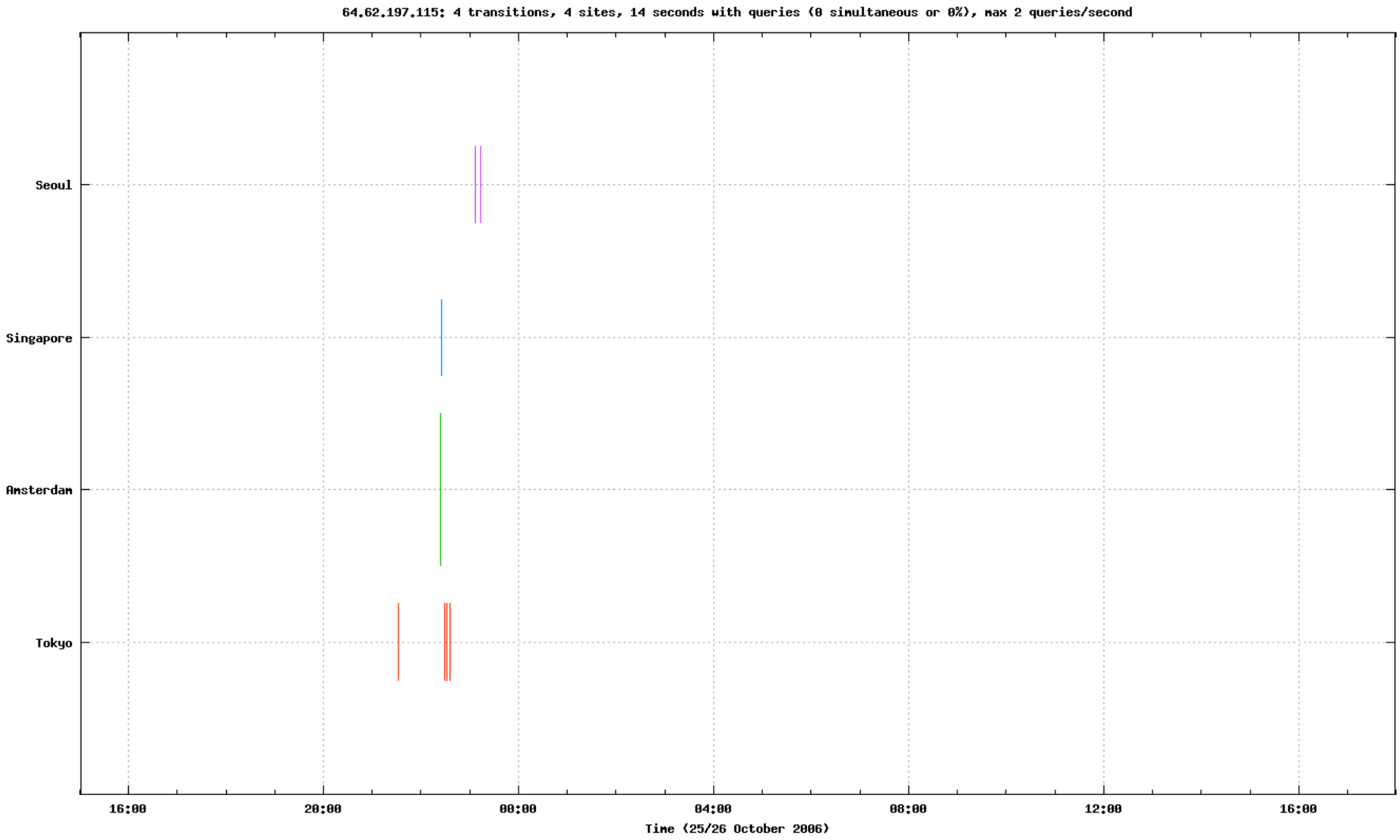
# Simultaneous Sites, Graph #5 (18th on Top 30 List)



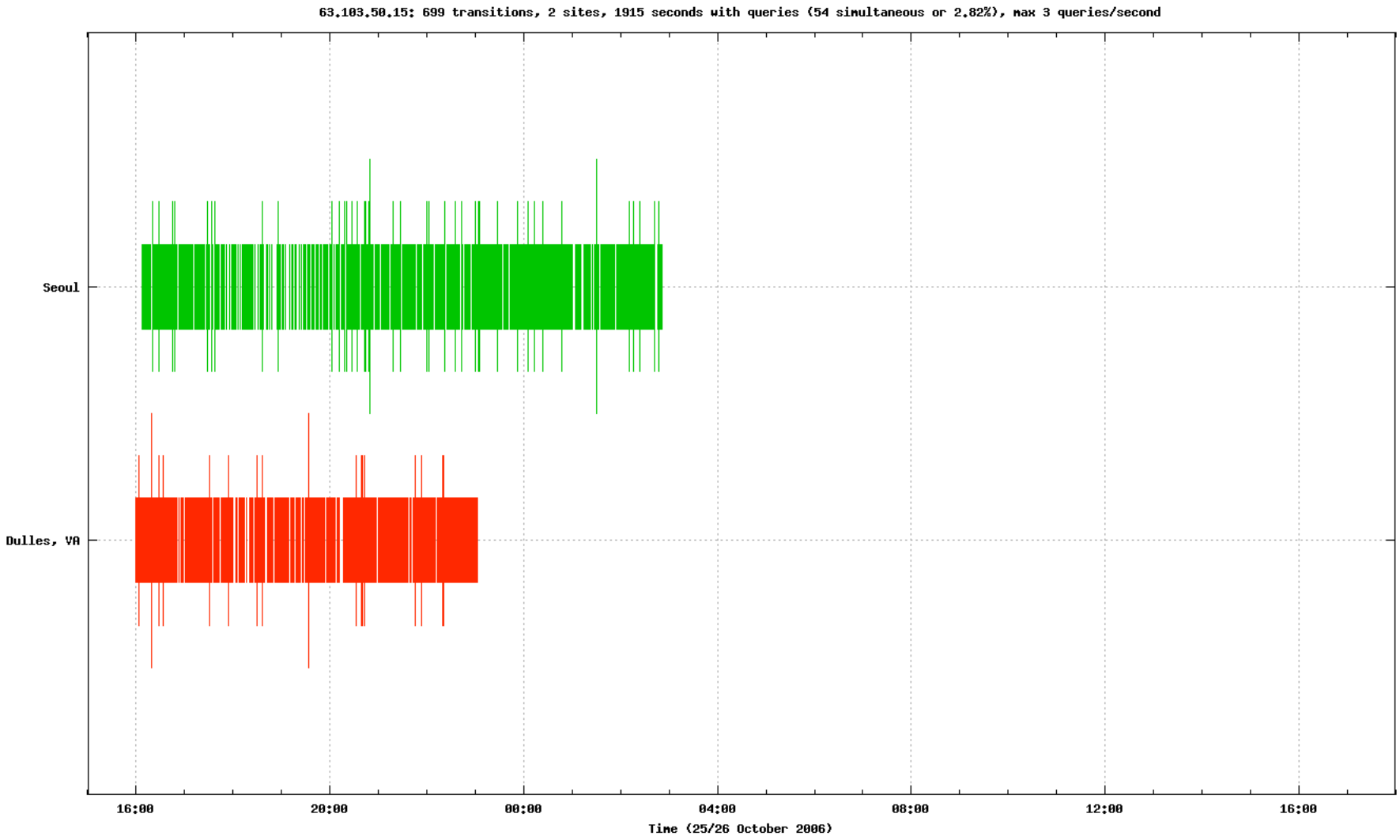
# Many Sites, Graph #1 (5th on Top 30 List)



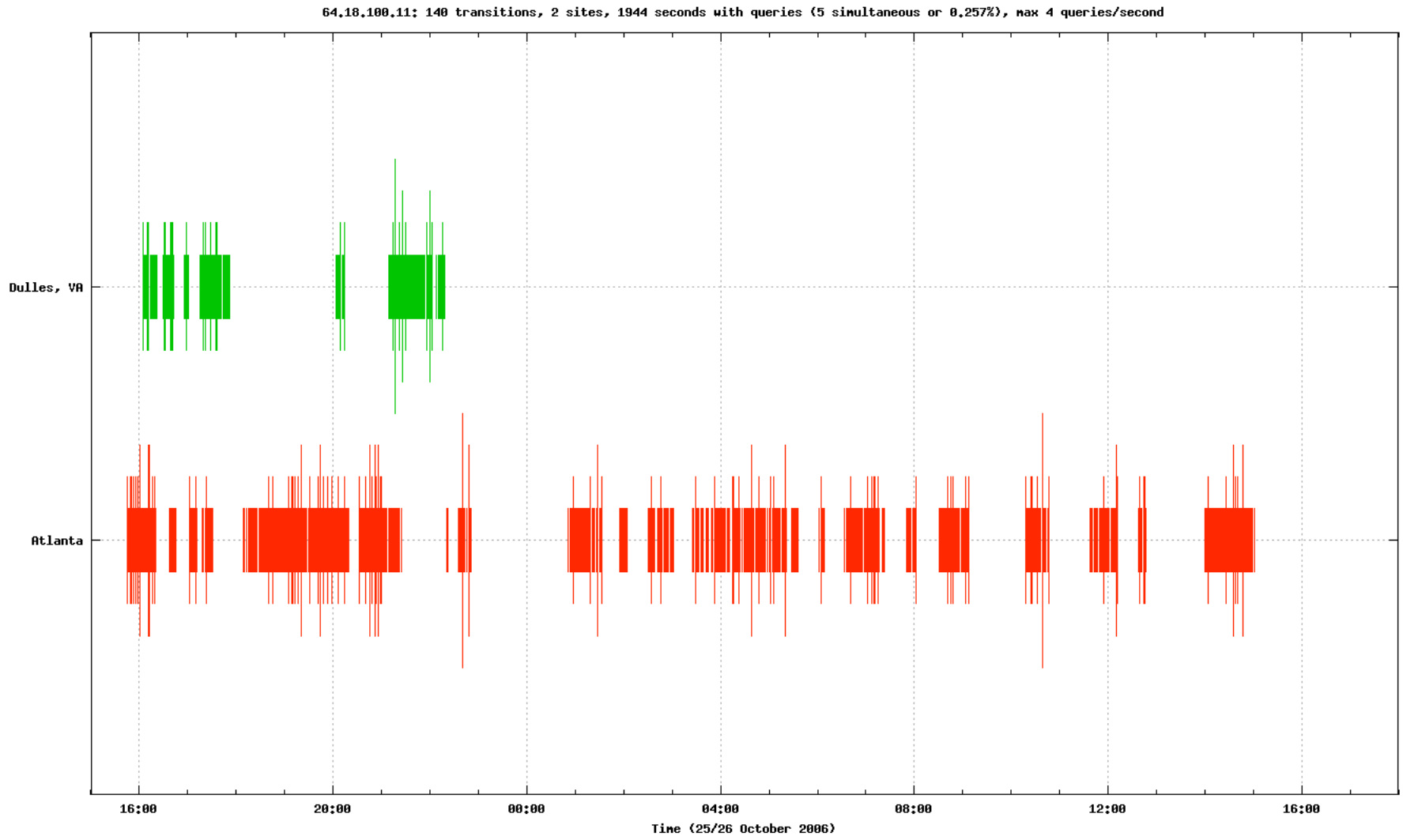
# Many Sites, Graph #2 (17th on Top 30 List)



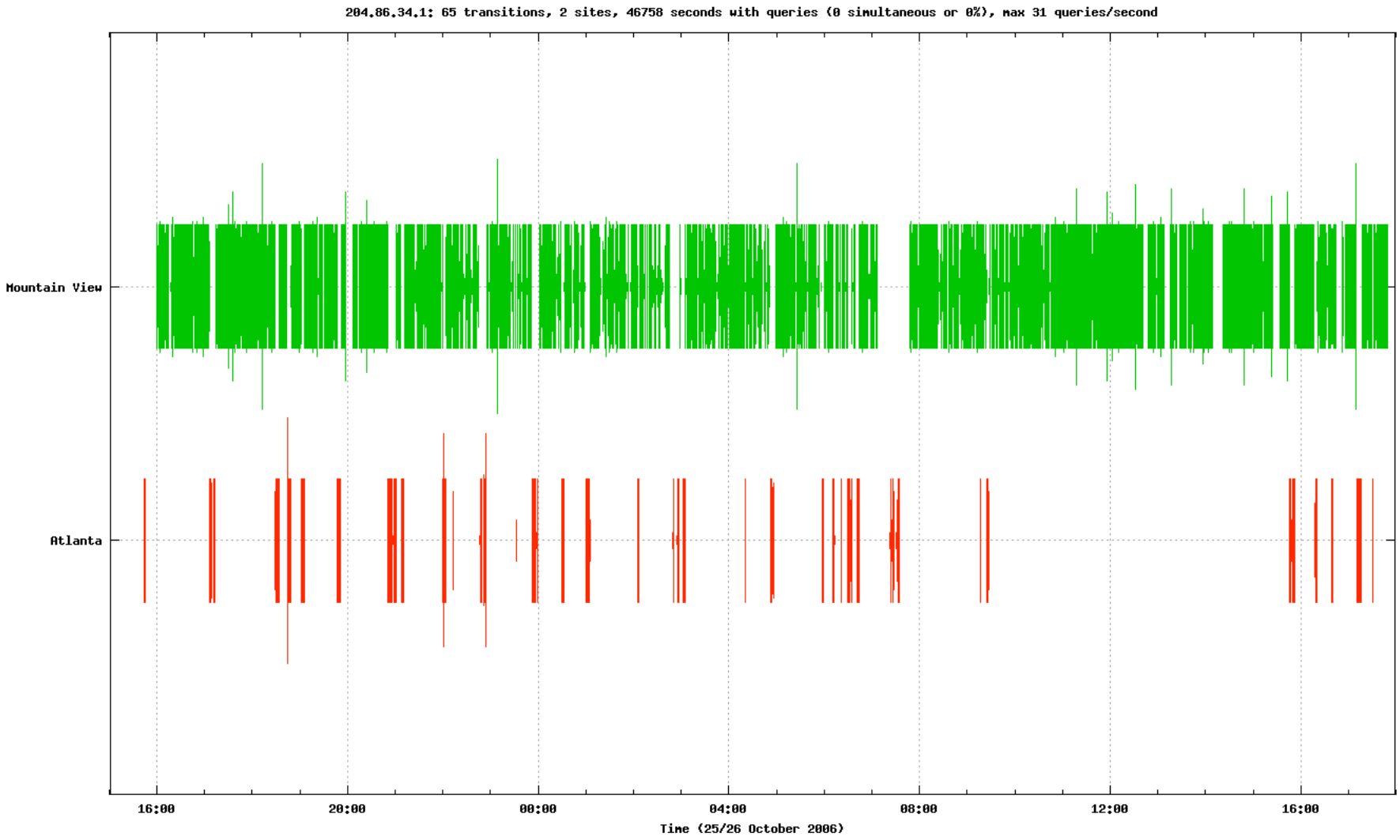
# Many Transitions, Graph #1 (4th on Top 30)



# Many Transitions, Graph #2 (11th on Top 30)



# Many Transitions, Graph #3 (22nd on Top 30)



# TCP Packet Capture

- About 25.5 hours of inbound query data:
  - Start: around 1100 EST 31 October 2006
  - Stop: around 1330 EST 1 November 2006
  - Staggered start/stop for packet capture at each site
- High-level statistics:
  - **606,822** total DNS responses sent
    - TCP queries harder to count and analyze
    - DNS query message can and did span multiple segments/packets
  - **238,932** TCP connections established
    - Counted SYN+ACK sent by server
    - Ratio of **2.5** queries per TCP connection
  - **22,854** unique source IP addresses
    - Ratio of **26.5** queries per unique IP address

# TCP Resets Sent

- Some TCP resets sent by J root servers would indicate a potential problem
  - E.g., route change during connection, new server instance sends reset in response to unrecognized TCP segment from client
- But some resets legitimate
  - E.g., server closes idle connection
- Counted unmatched resets sent
  - RST sent to <IP,port> not preceded by SYN received from <IP,port>
- **9320** unmatched resets sent to **481** unique IP addresses
  - One unmatched reset sent every **10 seconds**
  - Ratio of one unmatched reset sent for every **25** successful TCP connections
- Resets sent to **5** source IP addresses from two different sites
- Why so many resets? More analysis needed!

# TCP Queries to J Root

- But why is *j.root-servers.net* receiving **any** TCP queries?
- Either:
  - Root servers are sending truncated (TC bit set) DNS replies over UDP to prompt requery over TCP, or
  - Resolvers are querying over TCP first before UDP
- 26-hour UDP capture is inbound only, so couldn't check for TC bit on replies
  - But only 4 out of 495,714 responses had TC bit set in quick capture from Dulles, VA instance
- TCP queries are overwhelmingly QTYPE MX
  - See chart at right for QTYPE breakdown
- In actual UDP traffic:
  - Truncated replies from the root are hard to provoke
  - QTYPE A predominates (> 50%)
- Suspect TCP querying without UDP first: Who and why? More research needed!

| TCP query QTYPEs | | |
|---|---|---|
| **QTYPE** | **Count** | **% Total** |
| WKS | 3 | 0.0005% |
| A6 | 4 | 0.0007% |
| IXFR | 13 | 0.0021% |
| CNAME | 126 | 0.0208% |
| AAAA | 258 | 0.0426% |
| SRV | 302 | 0.0498% |
| AXFR | 389 | 0.0642% |
| PTR | 1,842 | 0.3040% |
| ANY | 1,942 | 0.3206% |
| SOA | 2,602 | 0.4295% |
| NS | 3,609 | 0.5957% |
| TXT | 11,353 | 1.8740% |
| A | 16,194 | 2.6731% |
| **MX** | **567,185** | **93.6224%** |

# Conclusions / Findings

+ Much of the “interesting” IP multi-site traffic looks explainable by per-packet load balancing
+ A puzzling and as-yet-unexplained large number of unmatched TCP resets sent by J root servers
+ Some iterative resolvers appear to send TCP queries without trying UDP first, particularly for MX records
+ Looks like TCP DNS over anycast is not exactly the same as unicast
  - **We do not think there is a problem here**
  - But more research would be good

# For The Record…

- "Life and Times of J-ROOT" (NANOG 32, October 2004) has this statement:
  - "DO NOT RUN Anycast with Stateful Transport"
- What it did not mean / what we never said:
  - "Don't run DNS over anycast"
- DNS over anycast works great!
  - VeriSign is a big proponent of anycasting critical DNS infrastructure
  - VeriSign anycasts root and *.com*/*.net* name servers
- TCP DNS is not broken by anycast
  - It clearly works (in most cases)
- But longer-running TCP sessions may have problems
  - We see that over transitions
  - Per-packet load balancing sites may have issues
  - Want to warn those who don't engineer these anycast solutions carefully

# Future Research

- Reasons for and circumstances surrounding unmatched TCP resets sent
- Source of (which implementation) initial TCP queries not preceded by UDP queries
- Investigate transitions in “global node” vs. “local node” context

VeriSign®

# Thank You