Identifying IPv6 Network Problems in the Dual-Stack World [*†]

Kenjiro Cho
IIJ/WIDE Project
kjc@iijlab.net

Matthew Luckie
U.Waikato/NLANR/CAIDA
mjl@wand.net.nz

Bradley Huffaker
CAIDA/SDSC/UCSD
bhuffake@caida.org

1 Introduction

The IPv6 Internet is in a state of transition from a collection of experimental research networks, such as the 6bone [4], toward a collection of production networks. One of the major hurdles limiting IPv6 adoption is the existence of poorly managed experimental IPv6 sites that negatively affect the perceived quality of the IPv6 Internet. To promote the use of IPv6, many operating system IP stacks prefer IPv6 to IPv4 when both protocols are available to be used in communicating with another system. A *dual-stacked* system is a system with both IPv4 and IPv6 protocol stacks available and configured. When an IPv6 user encounters an IPv6 system across a relatively poor IPv6 network, the user-perceived performance is considerably degraded. When this occurs, a frustrated user may hastily conclude that the problem lies with IPv6.

As IPv6 connectivity becomes available, some advanced users start experimenting with IPv6, possibly by using IPv6-in-IPv4 tunnels. Typically, they find IPv4 connections performing better than IPv6 connections. As initial experimental interest fades away, some users stop using IPv6 altogether, and may unintentionally leave poorly managed IPv6 networks behind. If use of IPv6 fails, communication automatically falls back to IPv4. Many users are not aware of their use of IPv6, nor problems in the IPv6 network. IPv6 network problems are often overlooked because of the transparent design of IPv6 systems.

Making the IPv6 Internet fully functional will require a major change. No simple solution appears, other than fixing each individual path problem as it is identified. Although traditional tools such as `ping` and `traceroute` are useful for investigating IPv4 and IPv6 independently, we can gain a better understanding of IPv6 problems with tools specifically designed to compare IPv6 and IPv4 measurements. By comparing IPv6 and IPv4 paths, we can focus on problems that are present only in the IPv6 path.

We are exploring methods to illustrate IPv6 network problems that provide insight to network operators and system administrators. Our approach makes use of the availability of both of the IPv4 and IPv6 protocol stacks to compare the two types of paths. Our results appear promising for understanding the status of IPv6 deployment and for improving the quality of the IPv6 Internet.

One can measure and diagnose problems in the IPv6 Internet using similar techniques to those used in the IPv4 Internet. Most network management tools available for IPv6 are simple replacements of the tools developed for IPv4. Because the IPv6 Internet is being deployed via tunnels over as well as in parallel (native) with the existing IPv4 Internet, we can develop new techniques specifically designed to manage both networks. Our focus is on dual-stack tools that measure and compare IPv4 and IPv6 paths to provide insight to network operators and system administrators.

2 Methodology

Our methodology is simple. First, by monitoring DNS messages, we create a list of systems with IPv6 and IPv4 addresses in actual use. Second, we measure delay with `ping` to each address in order to select a few nodes per site based on the IPv6:IPv4 round-trip time (RTT) ratios. Finally, we run `traceroute` with Path MTU (PMTU) discovery [10] to the selected sites, and visualize the results for comparative path analysis.

2.1 Dual-Stack Node Discovery

It is challenging to produce an address list that provides a reasonable coverage of dual-stacked sites and systems in the world. Existing studies often select targets semi-manually from a larger set such as a client list obtained

[*] wide-draft-mawi-dualstack-00.pdf. 2005/01/26.

[†] This paper appeared in SIGCOMM'04 Workshops, Aug 2004, Portland, Oregon, USA.

from server access logs. Our approach is to monitor DNS responses and record those with AAAA Resource Records (RRs). A AAAA (quad-A) record maps an IPv6 address to a hostname in a similar way to how an A record maps an IPv4 address to a hostname. We assume that a DNS response with AAAA records indicates that an IPv6 address is likely to be in actual use without prejudging the service it offers.

We define a dual-stack node as a system that has both IPv4 and IPv6 protocol stacks implemented and configured for operation. Our measurement targets are only those nodes with both IPv4 and IPv6 addresses registered in the DNS. Since an IPv6 address can be automatically configured, having an IPv6 address configured does not necessarily indicate an intention to use IPv6. When a system has an IPv6 address registered in the DNS, we assume it is intended to provide some service over IPv6. Although there are cases where a hostname points to topologically different IPv4 and IPv6 addresses, we do not distinguish them, since they are the same service from a user's point of view.

To find dual-stack nodes in real use, we passively monitor for DNS responses and record hostname and IPv6 address pairs appearing in the answer, authority, and additional sections. Many IPv6-capable clients first search for IPv6 addresses of a hostname, and then for IPv4 addresses of the same name. The answer section contains the Resource Records that answer a query, for example, a AAAA record containing an IPv6 address for a given hostname. The authority and additional sections provide auxiliary information about the authoritative name servers for the hostname and/or address. We extract any name server information from the authority and additional sections because we prefer DNS servers as measurement targets, since they are generally well-maintained and robust to occasional measurement.

From the list obtained, we extract nodes that have legitimate global unicast IPv6 addresses, and perform DNS lookups for both IPv4 and IPv6 addresses for the hostname. This is to confirm that the nodes actually have both IPv4 and IPv6 addresses for the given hostnames. This process provides a list of target dual-stack nodes for use in the dual-stack ping measurement.

2.2 Dual-Stack Ping

Our dual-stack ping is a script that obtains the IPv4 and IPv6 RTT delays for a set of target nodes by running `ping` and `ping6`.

ICMP-based RTT measurement bears well-known limitations: many firewalls filter ICMP packets, and some routers process ICMP packets in the slow forwarding path, rendering measured RTT artificially larger than that of other packets. Nonetheless, `ping` provides an estimation of the comparative difference between IPv6 and IPv4 that is close enough for our purposes.

From the dual-stack ping results, we can identify the percentage of dual-stack nodes reachable only by IPv4 even though they have AAAA records. When a node is unreachable by both IPv4 and IPv6, the target node may be off-line, or there may be a network problem not specific to IPv6. The number of nodes reachable only by IPv6 is not reliable since many sites filter ICMP for IPv4 but not for IPv6. Conversely, it is unlikely that ICMP is filtered only for IPv6. Therefore, we assume that when a node is reachable only by IPv4, it is an indication of IPv6 network problems that need further investigation.

From this set of nodes we select a few representative nodes per site. By the current IPv6 address assignment rules, we assume an organization has a fixed prefix length of 48 bits, which is a Site-Level Aggregation or SLA [5], where a site is loosely defined as an organizational unit in a single geographical location.

We select up to two representative nodes for each /48 using the following rules. Nodes reachable by both IPv4 and IPv6 are classified by their IPv6:IPv4 RTT ratios into 3 groups; Large ($ratio > 1.25$), Small ($ratio < 0.8$), and Equal ($0.8 \leq ratio \leq 1.25$). One node is selected from each group, except when both the Large and Small groups are not empty then the Equal group is omitted.

A node with the largest (smallest) RTT ratio is selected as a representative node for the Large (Small) group. For the Equal group, we select one with the shortest string length for the concatenated numeric-address and hostname. This works well for selecting suitable targets since important servers tend to have a manually assigned shorter address form (e.g., prefix::1) and a shorter hostname (e.g., ns.example.com).

If the /48 has no node reachable by both IPv4 and IPv6 but there is a node reachable only by IPv4, we select the representative node using the same heuristics as used for the Equal group.

Often only one node is selected for a site because all nodes in the site share the same network path. If a specific node has a large RTT, we select it along with another representative node, to facilitate comparative analysis in distinguishing a node problem from a site problem.

We also take the distribution of the IPv6:IPv4 RTT ratio among the nodes reachable by both IPv4 and IPv6. We categorize the distribution into different geographical regions to observe regional differences. We base our classification on the publicly available IP address assignment database provided by the Regional Internet Registries (RIR). The resulting statistics provide an estimation of the quality of the IPv6 network relative to that of IPv4.

2.3 Dual-Stack Traceroute and Visualization

The third step is to identify specific problems and their causes through discovering and visualizing the forward topology. Most problems lie in routing, e.g., routing loops, vanishing routes, and roundabout routes. A roundabout route is not always caused by a routing problem *per se*, but by a lack of peering or IPv6-capable paths. Because IPv6 exchange points and paths are still fairly limited, a packet could travel much further with IPv6 than the same packet might travel with IPv4. One of our goals is to identify a lack of peering or paths for IPv6. Another related problem is poorly configured tunnels that disregard the underlying topologies. Tunnels are useful during the early stages of IPv6 deployment, but poorly configured tunnels, especially in the backbone, present performance problems and other issues when left untended after infrastructural changes.

It is difficult to identify path problems by simply running the traditional `traceroute` program, since it often requires comparative analysis of multiple paths using knowledge of the underlying topology. Our method employs visual comparison of IPv4/IPv6 path pairs to intuitively recognize path anomalies. If necessary, we can use traditional `traceroute` to further investigate details of a path in question.

Our dual-stack traceroute tool is `scamper` [14], successor of `skitter` [6]. Both `skitter` and `scamper` are designed for large scale topology measurement, run multiple traceroutes in parallel at a specified packet-per-second rate, and terminate a trace as soon as the destination is detected to be unreachable. In addition, `scamper` can probe both IPv4 and IPv6 addresses, and has the ability to perform PMTU discovery.

We use PMTU discovery to identify IPv6-in-IP4 tunnels, since a drop in MTU at an intermediate router indicates a possible tunnel entry point. It is useful to identify tunnels, especially those ignoring the underlying IPv4 topology. The tunnel discovery is also useful for troubleshooting since problems in tunnels are often caused by the underlying IPv4 networks and hard to debug with IPv6 tools alone. Colitti *et al.* use PMTU discovery for tunnel detection in [1] and propose several techniques to infer and confirm the existence of IPv6-in-IPv4 tunnels. We use only PMTU discovery because tunnel detection is not the goal and is only used as auxiliary information for path analysis.

The visualization script reads the output of `scamper`, and creates graphs comparing IPv4 and IPv6 path pairs. The graph juxtaposes IPv4 and IPv6 path pairs for neighboring destinations, and plots intermediate hops according to their RTTs.

3 Results

Data was collected by measurement from three locations in June, 2004. The three locations are 1) WIDE [15], a research network in Tokyo, Japan; 2) IIJ [8], an ISP providing commercial IPv6 services in Tokyo, Japan; and 3) Consulintel [2], in Madrid, Spain, directly connected to MAD6IX that is part of Euro6IX [3]. These three measurement points are arguably among the best connected IPv6 sites in the world, and are referred to as the WIDE, IIJ, and ES sites in this paper.

3.1 Dual-Stack Node Discovery Results

We set up several DNS monitors within the WIDE network from April to June in 2004. By monitoring AAAA records, we obtained 11,834 unique hostname and IPv6 address pairs. Table 1 shows a breakdown of the obtained IPv6 address prefixes.

Table 1: IPv6 address prefixes captured within the WIDE network

prefix	prefix use	pairs
2001::/16	Aggregatable Global Unicast for sub-TLA	6,585
3ffe::/16	6bone	1,762
2002::/16	6to4	241
::ffff/96	IPv4-mapped IPv6 address	97
::/96	IPv4 compatible IPv6 address	31
fe80::/10	Link-local Unicast	6
fec0::/10	Site-local Unicast	2
other	reserved or unassigned address	3,110

We extracted a total of 8,347 pairs for ‘2001::/16’ and ‘3ffe::/16’ since only global unicast IPv6 addresses are of interest. We then performed DNS lookups by hostname for A and AAAA records, and found that 4,711 pairs actually have both A and the matching AAAA. After removing invalid IPv4 addresses (e.g., RFC1918 addresses and local addresses) and duplicates with identical IPv6 and IPv4 address pairs but with different host names, we obtained 4,086 target dual-stack nodes.

Table 2 shows the distribution of the target dual-stack nodes by their country code, representing 47 countries. We obtain the country code for each pair by matching the IPv6 addresses against the allocated IPv6 prefix in the RIR’s database. We use the country code of the address block assignee in the database entry. Limiting this approach is that the real location of a node may be different from the registered country. In addition, we do not consider the associated IPv4 addresses at all.

3.2 Dual-Stack Ping Results

We performed the dual-stack ping from the three locations, from the WIDE and IIJ sites on June 10 and from the ES site on June 23, using the same list obtained by the dual-stack node discovery within the WIDE network. Table 3 lists the numbers of unreachable and reachable

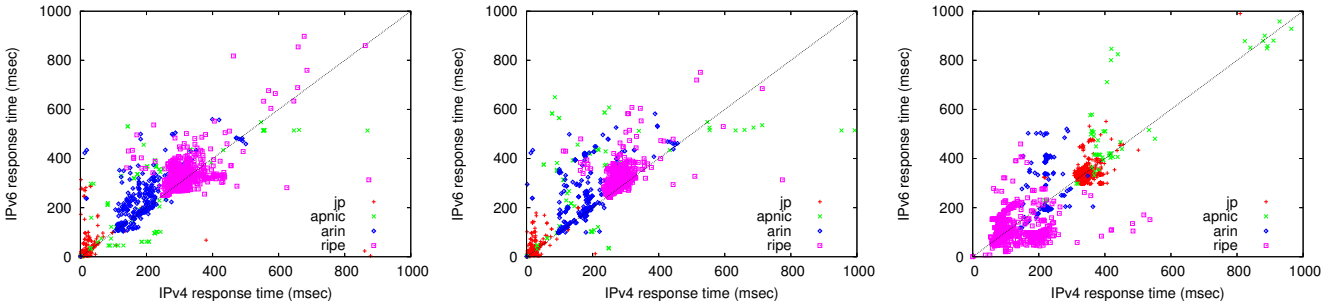

Figure 1: Distribution of IPv6/IPv4 RTT from WIDE (left), IIJ (middle), and ES (right)

Table 2: Number of dual-stack targets by country code based on their IPv6 address

JP:1155	ID:79	NO:34	KR:17	LU:9	PH:4
NL:497	FI:68	CZ:30	MY:17	RU:8	TN:4
US:464	IT:68	DK:29	BR:16	TH:8	YU:4
DE:431	SK:68	TW:27	HU:13	ZA:8	AR:2
FR:251	CH:59	AT:25	LT:13	BE:6	RO:2
UK:186	PL:57	EU:21	CN:10	SG:6	CL:1
CA:144	AU:41	EE:18	MX:10	GR:4	IL:1
SE:93	IE:39	PT:18	ES:9	HK:4	

nodes by IPv4 and IPv6 from the WIDE site. The results from IIJ are almost identical, and the results from ES are similar to WIDE's. About 66% are reachable by both IPv4 and IPv6. However, about 16% are reachable by IPv4 but not by IPv6 even though they have AAAA records. These sites would force communicating peers to timeout with IPv6 before falling back to IPv4. In Table 3, the nodes are classified into four regions by matching their IPv6 address prefixes to the RIR database; 'jp' for Japanese nodes, 'apnic' for non-jp APNIC nodes, 'arin' for ARIN and LACNIC nodes, 'ripe' for RIPE NCC nodes. Japanese nodes are separated from other APNIC nodes since their node number is large and most of the Japanese nodes are in Tokyo, so that their RTT is usually less than 10 msec from the WIDE and IIJ sites. Since the number of LACNIC nodes is so small, we merge them with the ARIN nodes.

When we examined the two middle groups, those that had only IPv4 or IPv6 responding addresses, there was an unusual difference in the ratio of addresses in these two groups across different RIRs. The ratios in Japan and RIPE NCC were around 0.6, 0.57 and 0.67 respectively. In contrast, ARIN was about half that with 0.23 and APNIC was almost four times with 2.43. The low level of IPv6 responding in ARIN could be the result of the low level of commitment to IPv6 in the US, which makes up a large portion of ARIN's membership. The surprisingly strong ratio of responding IPv6 addresses in APNIC might be the result of stronger support for their relatively large IPv6 blocks in comparison to their small IPv4 allocations.

Table 3: Number of unreachable and reachable nodes by dual-stack ping from WIDE.

IPv6		unreach	unreach	OK	OK
IPv4		unreach	OK	unreach	OK
total	4086 (100%)	370 (9.0%)	634 (15.5%)	384 (9.4%)	2698 (66.0%)
jp	1155 (100%)	83 (7.2%)	126 (10.9%)	72 (6.2%)	874 (75.7%)
apnic	213 (100%)	37 (17.4%)	28 (13.2%)	68 (31.9%)	80 (37.6%)
arin	645 (100%)	80 (12.4%)	168 (26.1%)	38 (5.9%)	359 (55.7%)
ripe	2042 (100%)	162 (7.9%)	306 (15.0%)	204 (10.0%)	1370 (67.1%)

Figure 1 shows the scatter graphs of the observed RTTs. We plot a node's IPv4 RTT on the X-axis and its IPv6 RTT on the Y-axis. Each graph plots about 2,700 nodes that were reachable by both IPv4 and IPv6. When the response time of IPv6 is equal to that of IPv4, we plot the node on the unity line, $y = x$. For nodes above this line, IPv4 outperforms IPv6, and for nodes below this line, IPv6 outperforms IPv4. We again categorize the nodes into four regions.

These results indicate that the majority of the nodes have similar RTT for both IPv4 and IPv6. A number of individual nodes far above the unity line have IPv6 performance issues specific to the node or the site. The clusters above the unity line indicate the existence of roundabout paths within the backbone network.

Compared to WIDE, IIJ has fewer nodes below the unity line, probably due to Acceptable Use Policies (AUPs) of academic IPv6 networks. The ES site has a large cluster of RIPE nodes below the unity line, likely connected through Euro6IX [3]. The majority of nodes are around the unity line; the percentage of nodes whose IPv6:IPv4 RTT ratio is less than 1.25 is 80.1% for WIDE, 74.3% for IIJ, and 82.5% for ES.

APNIC nodes have large variance in RTT ratios due to its topological diversity; many APNIC countries are connected to Japan through the US or Europe, and many

satellite links connect islands. Also, some networks are funded to promote IPv6, such that there are nodes with a direct IPv6 path but with an IPv4 path that must go through the US.

Next, we select representative nodes for each /48 using the rules described in Section 2.2. For the WIDE site, we selected 1,334 nodes out of 4,086 nodes for 1,469 /48s. For the IIJ and ES sites, we selected 1,320 and 1,310 nodes, respectively. We selected fewer nodes than the number of 48-bit prefixes since we selected no nodes in sites not reachable by IPv4. The reduction rate is about $1/3$ for these results, but it improves if more nodes are available per site.

3.3 Dual-Stack Traceroute Results

We ran `scamper` to the representative nodes with PMTU discovery on June 11, 2004 from the WIDE and IIJ sites, and on June 16 from the ES site. To visualize the results, a set of scripts divide the `scamper` output into smaller target groups and create a graph for each group. Each graph contains 10 target nodes, yielding about 130 graphs for each measurement; the script also creates a web page for each graph along with scrollable `scamper` text output. While there have been a number of attempts to visualize traceroute-derived topology [11, 6], we are not aware of published work that extensively compares IPv4 and IPv6 paths. In the graph we map IP addresses into Autonomous System (AS) numbers to simplify the presentation [7, 9, 12].

Figure 2 shows example outputs of the `scamper` visualization towards 2001:468:X::/48, the nodes within the ABILENE address block (selected simply because it is less controversial for publishing results). The top graph is from the WIDE site, the middle graph is from the IIJ site, and the bottom graph is from the ES site. The target nodes are slightly different for each measurement site since they are selected based on the dual-stack ping results of each site.

For each target node in the graph, two lines are drawn from the source to the destination, the upper line for IPv4 and the lower line for IPv6. A missing line indicates the destination is unreachable. To the left of the line, the graph shows the total hop number and destination RTT.

The graphs plot intermediate hops at their RTTs from the source. We map IP address of a hop to its AS number by finding the best matching prefix and origin AS in the publicly available Routeviews BGP table [13]. There were 165,289 prefixes for IPv4 and 520 prefixes for IPv6 in the BGP table at the time of measurement.

When a drop in MTU is detected, the graph marks the MTU on the right side of the hop; it suggests a likely tunnel between this hop and the previous hop. If `traceroute` terminated with an error, the graph marks the error code at the hop using the `traceroute` notations (e.g., '!X' for communication administratively prohibited).

In the WIDE and IIJ graphs, most destinations have similar RTTs for IPv4 and IPv6. In the WIDE graph, the IPv6 paths are similar to the IPv4 paths, therefore they appear to be IPv6-native dual-stack paths. In contrast, the IIJ graph shows IPv6 paths going through ASes different from IPv4 paths, which is more common in the current IPv6 Internet. AS-level comparison yields insight into path differences since many IPv6 paths do not follow their IPv4 counterparts.

In the ES graph, the IPv6 paths are much longer in time than the IPv4 paths. All the IPv6 paths share the long hop after the hop with 1280 MTU, consistent with a tunnel that makes a detour to the destinations. Note that this is not a typical path to the US from the ES site; we observed better paths to other US destinations in other graphs but did not include them in this paper.

Table 4 shows the distribution of the Path MTU size detected by `scamper`. Since the target nodes include nodes reachable by `ping` but not by `ping6`, the number of nodes unreachable by `ping6` is shown at the bottom. A 1454-byte MTU is common for PPPoE, and 1280 and 1480 bytes are default MTU sizes for popular tunnel implementations. WIDE stands out with a high number of 1500-byte IPv6 PMTUs, likely a result of their efforts to promote native IPv6 connections.

Table 4: distribution of Path MTU size

	IPv4			IPv6		
PMTU	WIDE	IIJ	ES	WIDE	IIJ	ES
1500	761	751	732	575	76	33
1492	6	6	8	-	-	-
1480	2	2	1	64	96	15
1476	2	1	1	8	2	-
1472	1	1	1	2	-	-
1456	-	-	1	-	-	-
1454	90	95	82	-	-	-
1450	-	-	-	-	2	-
1448	2	2	2	-	-	-
1446	1	1	1	-	-	-
1400	-	-	1	-	-	-
1280	1	1	1	184	622	500
1258	1	-	1	-	-	-
unknown	46	43	44	47	21	83
unreach	421	417	433	454	501	679
unreach by ping6	-	-	-	249	276	273

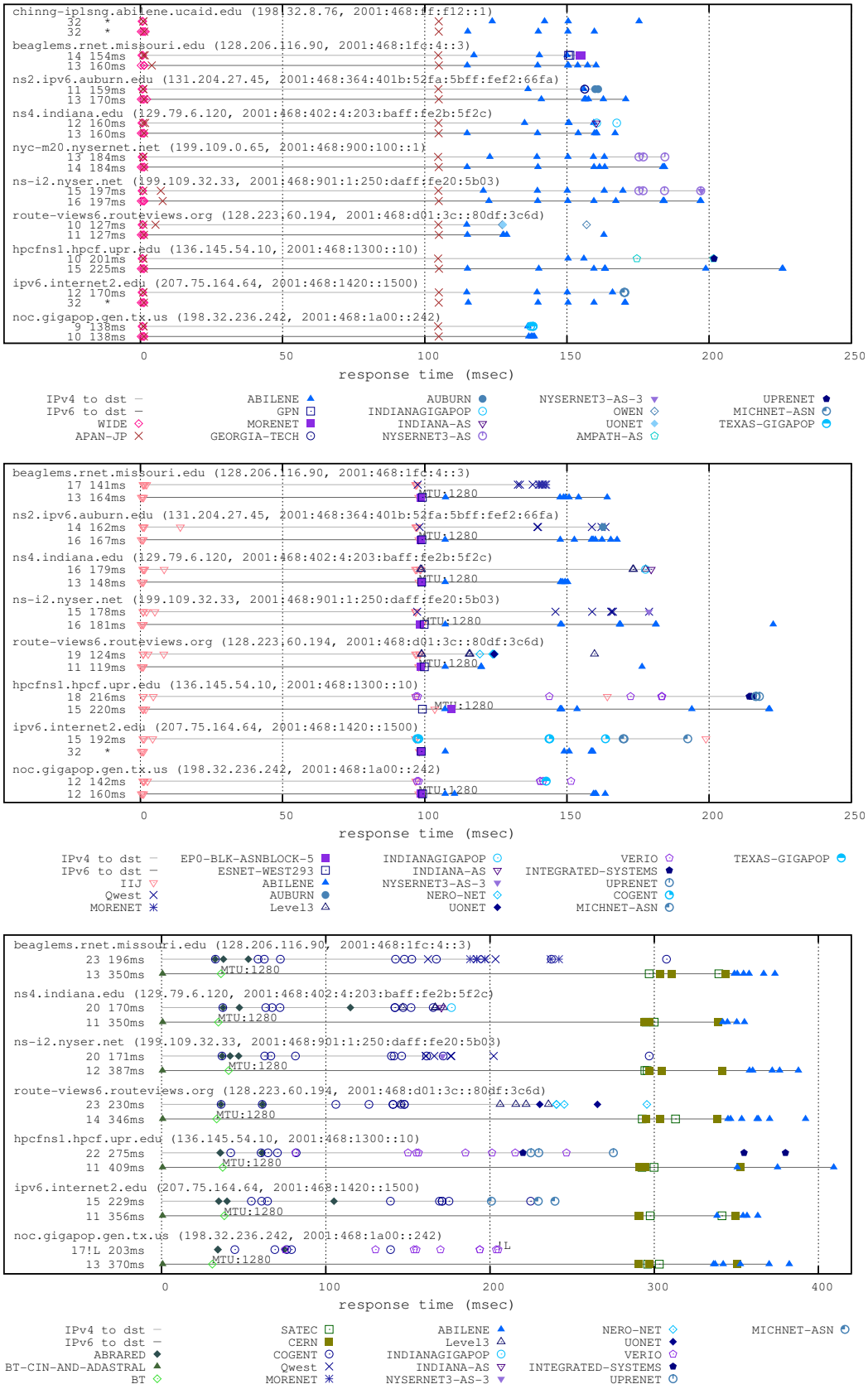

Figure 2: Path visualization towards 2001:468::/16 from WIDE (top), IIJ (middle) and ES (bottom)

4 Conclusion

It is essential to IPv6 deployment to improve the quality and performance of the IPv6 Internet. In order to illustrate IPv6 network problems for network operators, we are developing tools to compare IPv6 measurements with corresponding IPv4 measurements. Our techniques include the dual-stack node discovery for finding dual-stack nodes, the dual-stack ping for selecting representative nodes, and `scamper` for detailed path analysis. Our test results indicate that we can improve IPv6 network quality by identifying and fixing a limited amount of erroneous settings.

Our tools are still under development and need improvements. We plan to fully automate the measurement procedure in order to perform regular measurements and archive results. This long-term measurement strategy will provide a way to evaluate the progress of IPv6 deployment. We would also like to increase the coverage of measurement points including developing countries. Another important step is to establish procedures to notify responsible parties of problems we find using our tools.

The results of our measurements, along with our tools, are available from http://mawi.wide.ad.jp/mawi/dualstack/.

References

[1] L. Colitti, G. D. Battista, and M. Patrignani. IPv6-in-IPv4 tunnel discovery: methods and experimental results. *IEEE Transactions on Network and Service Management*, 1(1), Apr. 2004.

[2] Consultores Integrales en Telecomunicaciones. http://www.consulintel.es.

[3] Euro6IX project. http://www.euro6ix.net.

[4] R. Fink. 6BONE pTLA and pNLA Formats (pTLA). RFC 2921, IETF, Sept. 2000.

[5] R. Hinden. Proposed TLA and NLA Assignment Rules. RFC 2450, IETF, Dec. 1998.

[6] B. Huffaker, D. Plummer, D. Moore, and k claffy. Topology discovery by active probing. In *SAINT2002*, Jan. 2002.

[7] Y. Hyun, A. Broido, and kc claffy. On third-party addresses in traceroute paths. In *Passive and Active Measurement Workshop 2003*, La Jolla, CA, Apr 2003.

[8] Internet Initiative Japan. http://www.iij.ad.jp.

[9] Z. M. Mao, J. Rexford, J. Wang, and R. H. Katz. Towards an accurate AS-level traceroute tool. In *SIGCOMM2003*, Aug. 2003.

[10] J. Mogul and S. Deering. Path MTU discovery. RFC 1191, IETF, Nov. 1990.

[11] R. Periakaruppan and E. Nemeth. GTrace: A graphical traceroute tool. In *USENIX LISA*, pages 69–78, Nov. 1999.

[12] Ripe NCC RISwhois. http://www.ripe.net/ris/riswhois.html.

[13] Route Views Project at the University of Oregon. http://www.routeviews.org.

[14] Macroscopic IPv6 Topology Measurements at CAIDA. http://www.caida.org/analysis/topology/macroscopic/.

[15] WIDE Project. http://www.wide.ad.jp.