Active Probing to Classify Internet Address Blocks

USC/ISI Technical Report ISI-TR-653, August 2008

(poster abstract)

Xue Cai John Heidemann
USC/Information Sciences Institute, {xuecai,johnh}@isi.edu

Categories and Subject Descriptors

C.2.1 [**Computer-Communication Networks**]: Network Architecture and Design—*Network topology*; C.2.3 [**Computer-Communication Networks**]: Network Operations—*Network management*

General Terms: Measurement

Keywords: Internet address allocation, survey, pattern analysis, clustering, classification, availability, volatility

1. INTRODUCTION

Previous Internet topology studies mostly focused on AS- and router-level topologies [1,2,4,8,9], providing insight into AS relationships and interdomain routing. However, relatively little is known today about the demographics of Internet edge hosts and the use of the IPv4 address space. Since the transition to classless routing [3], external observers have only limited view into how IP address blocks are allocated, managed, and used. Only recently have researchers tried to characterize address usage [11], with a methodology limited to addresses used by clients.

In this poster we begin to explore the potential of *active probing* and *external classification* of address block usage. Our work makes three assumptions: a significant number of Internet addresses will respond to external probes, some patterns of repeated probes correspond can indicate different address usage, and contiguous addresses are often used for similar purposes. While there are cases where these assumptions do not hold, in recent work we study probing effectiveness with ICMP and TCP, showing they detect the majority of used addresses [5]. In this poster we expand on the second hypothesis, using probe results to study address allocation strategies and to infer address usage (always be occupied or not in use, use by stable or frequently inaccessible hosts, etc.).

In ongoing work we are evaluating the accuracy of this approach and these assumptions; in this poster we suggest this as a promising new direction for research. We expand on the goals of this research, the applications of classification, and our preliminary results below.

2. GOALS AND APPLICATIONS

Our goal is to to classify Internet address blocks based how addresses respond to frequent probes over the source of one week. By developing and validating this technique, we hope to answer specific questions, including:

Copyright is held by the author/owner(s).
SIGCOMM'08, August 17–22, 2008, Seattle, Washington, USA.
ACM 978-1-60558-175-0/08/08.

Do groups of blocks show consistent patterns? How many vary? Groups of consistent patterns will allow clustering and classification of address usage with only external probes.

What are the sizes of blocks that show consistent usage? These patterns will suggest how address blocks are managed at fine granularities (smaller than /24s).

Can we map consistent, popular patterns to operational network usage? Can we identify groups of servers, cable or DSL customers, dial-up users, etc., by ping responses?

While these questions may seem academic, they enable answering important operational and policy questions. We expect that our approach will be approximate, yet even *approximate* answers to these new questions would be a huge step forward from today's understanding based primarily on anecdote or occasional surveys of providers.

A first important policy question is that understanding the address allocation and usage can give the academic community and the ISPs a better vision of current IPv4 address usage. Best estimates suggest ICANN will allocate its last free block of IPv4 addresses as soon as 2011 [6], so understanding address utilization is essential to Internet governance. For example, in each block, is addresses utilization high or low? Would new methods of address assignment improve efficiency? We expect that answering these questions and quantifying the costs of IPv4 address management will motivate IPv6 adoption and guide methods of IPv6 allocation.

Secondly, understanding dynamic usage of IP addresses will be helpful to the network research community concerning security. Xie et al. have begun to explore this question with a goal of identifying dynamic blocks to assist spam prevention [11]. Our method uses a completely different approach and so can extend and corroborate their findings.

Finally, understanding IP address usage patterns may help in locating the servers and clients, thus allows the service providers to better distribute their services on the Internet. Combining block usage with address geolocation can identify geographic regions of clients that are distant from services.

3. METHODOLOGY

Our methodology consist of four steps: surveying addresses with probes, pattern analysis of individual addresses, clustering to generalize these results, and broad classification.

Survey: We frequently probe a fraction of Internet address space [5]. We selected 24,000 /24 blocks (1% of the allocated address space), used ICMP Echo Request packets, and actively probed each IP addresses in these blocks every 11 minutes for around a week. While ICMP can be blocked, studies of a university and random addresses confirm that it is the most accurate method of active probing (more accurate than TCP) and solicits relatively few complaints. Our

results here are based on data from June 2007 [10].

Pattern Analysis: We analyze the usage of each IP address based on the survey replies. We define several metrics to characterize addresses. First, for each address, we consider it to be up when it responds to pings, and we define *up durations* (U_i, for each uptime i) as the time from the first response until the next non-response. (Precision of up durations estimates is limited by 11-minutes probe interval.)

We define *availability* as the sum of all up durations, normalized by total survey time. We define *volatility* as the number of up durations, normalized by the maximum number of possible state changes (half the number of probes). These metrics define an (A, V) plain of address blocks.

We also consider the distribution of lengths of up durations, including mean, median and maximum.

Clustering: We cluster adjacent IP addresses with similar usage (based on the above metrics) into blocks. We are currently experimenting with different clustering algorithms to explore all possible blocks with prefix lengths from /29 through /24 (groups of 8 to 256 addresses).

Classification: We classify the blocks we got from *clustering* into five categories: servers, and blocks that are stable, intermittent, underutilized, or unclassifiable. We trained our classification by manual examination of hostnames and services in several hundred blocks, and then we verify that it works by evaluating it on hundreds of other, randomly chosen blocks.

Server block: highly available and stable (high A, low V)

Stable block: stable, usually continuously up for more than 6 hours (low V, and high median U). Statically assigned addresses usually fall into this category.

Intermittent block: short up durations (low median U). For example, many DSL addresses are reassigned every few hours.

Underutilized block: low A values. Most wireless and some dial-up addresses are severely underutilized.

Unclassifiable block: unclassifiable due to too few responders. We frequently have difficulty classifying blocks of size /29 or smaller.

4. PRELIMINARY OBSERVATIONS

Figure 1 shows our preliminary results: a density plot of mean (A, V) values for all /24 block. Blocks that cluster to smaller block sizes. Darker dots indicate multiple blocks sharing the same availability and volatility. We have a preliminary classification of four categories, as indicated in the plane. The right rectangle around $(A, V) = (0.95, 0.0016)$ are blocks of servers with high availability and a low volatility. The next group, with a high A/V ratio (more than 16.4) are blocks of stable addresses. The A/V ratio indicates the typical duration of address occupation, with a higher ratio suggesting greater stability. The largest group, those with a low A/V ratio (less than 16.4), we call intermittent blocks. They show greater volatility than servers or stable blocks. Finally, the left side we call underutilized, where most addresses are sparsely occupied with low availability.

We developed these classifications by training on a subset of blocks, based on DNS names and occasional service discovery for many addresses. We get DNS data both from manual lookup and an ISC Internet Domain Survey [7]. As an example, we call out eight blocks in the figure. The hostnames in Block A contain "dns", a common type of server. Blocks B, C have hostnames with the word "static". Names

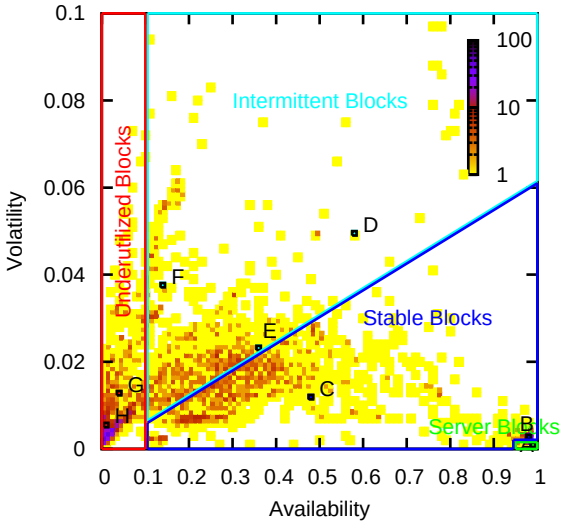


Figure 1: Volatility vs Availability of /24 blocks

of addresses in blocks D, E, and F contain the terms "dynamic" or "dsl". Finally, names of addresses in blocks G and H include the term "dialup".

Our accompanying poster provides more detail and analysis of this classification scheme, and validation of our classification on a random sample of blocks. As ongoing work we are evaluating clustering algorithms and block size accuracy.

5. REFERENCES

[1] X. Dimitropoulos, D. Krioukov, M. Fomenkov, B. Huffaker, Y. Hyun, kc claffy, and G. Riley. AS relationships: Inference and validation. *ACM Computer Communication Review*, 37(1):29–40, Jan. 2007.

[2] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. In *Proc. of the ACM SIGCOMM Conference*, pages 251–262, Cambridge, MA, USA, Sept. 1999. ACM.

[3] V. Fuller, T. Li, J. Yu, and K. Varadhan. Classless inter-domain routing (CIDR): an address assignment and aggregation strategy. RFC 1519, Internet Request For Comments, Sept. 1993.

[4] L. Gao. On inferring automonous system relationships in the internet. *ACM/IEEE Transactions on Networking*, 9(6):733–745, Dec. 2001.

[5] J. Heidemann, Y. Pradkin, R. Govindan, C. Papadopoulos, G. Bartlett, and J. Bannister. Census and survey of the visible internet (extended). Technical Report ISI-TR-2008-649, USC/Information Sciences Institute, February 2008.

[6] G. Huston. IPv4 address report. `http://bgp.potaroo.net/ipv4/`, June 2006.

[7] Internet Software Consortium. Internet Domain Survey, ISC_DS-2007JAN. web page `http://www.isc.org/ds`, Jan. 2007.

[8] W. Mühlbauer, O. Maennel, S. Uhlig, A. Feldmann, and M. Roughan. Building an AS-topology model that captures route diversity. In *Proc. of the ACM SIGCOMM Conference*, pages 195–204, Sept. 2006.

[9] L. Subramanian, S. Agarwal, J. Rexford, and R. H. Katz. Characterizing the Internet hierarchy from multiple vantage points. In *Proc. of the IEEE Infocom*, pages 618–627, June 2002.

[10] USC/LANDER project. Scrambled Internet Trace Measurement, PREDICT ID USC-LANDER/internet_address_survey_reprobing_it17w-20070601. web page `http://www.isi.edu/ant/lander`, June 2007.

[11] Y. Xie, F. Yu, K. Achan, E. Gillum, M. Goldszmidt, and T. Wobber. How dynamic are IP addresses? In *Proc. of the ACM SIGCOMM Conference*, Kyoto, Japan, Aug. 2007. ACM.

Active Probing to Classify Internet Address Blocks

Xue Cai, John Heidemann
USC/Information Sciences Institute
xuecai, johnh@isi.edu

USC Viterbi School of Engineering

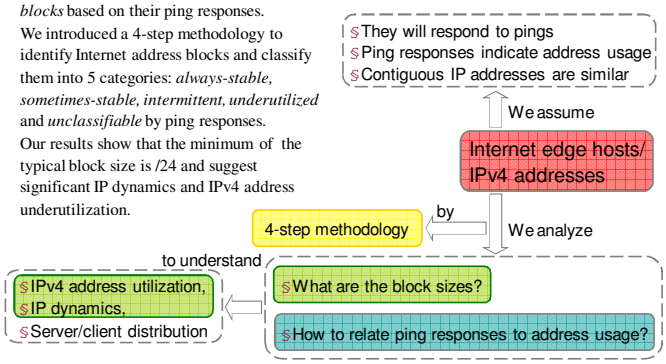
Information Sciences Institute

Introduction

Previous Internet topology studies focus on router or AS connectivity. In this poster we instead examine how *Internet edge hosts* are used in the IPv4 address space. In ongoing work we are exploring methodologies to actively probe all IPv4 addresses with ICMP echo requests (pings) [Heidemann08a]. Here we examine how to *classify Internet address blocks* based on their ping responses.

We introduced a 4-step methodology to identify Internet address blocks and classify them into 5 categories: *always-stable, sometimes-stable, intermittent, underutilized* and *unclassifiable* by ping responses.

Our results show that the minimum of the typical block size is /24 and suggest significant IP dynamics and IPv4 address underutilization.



Methodology

Internet edge hosts/IP addresses

Survey: ping each address in random /24 blocks every 11 minutes for a week.

Ping responses per IP address

Pattern Analysis: define *availability*, *volatility*, medianUp for addresses and blocks.

IP addresses with usage patterns

Block Identification: adjacent addresses with similar usage patterns are grouped to blocks.

Blocks with addresses with similar usage

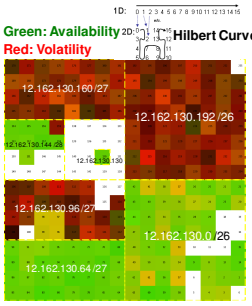
Classification: blocks are classified into five ping-observable categories.

Blocks classified to **5 *ping-observable categories***: always-stable, sometimes-stable, intermittent, underutilized, and unclassifiable.

For each **address**,
- D = probing duration (i.e., 1 week)
- I = probing interval (i.e., 11 min)
- N = number of pings = D/I
- r_i = ith ping response (positive/negative), i=1, …, N
- u_j = up durations, j=1, …, N_u
  = duration of the jth run of continuous positive r_is
- *Availability(host)* = $\sum r_i/N$
- *Volatility(host)* = $N_u/(N/2)$
- ***MedianUp(host)*** = median(u_j)

For each **block**,
- *Availability(block)* = median(Availability(host))
- *Volatility(block)* = median(Volatility(host))
- ***MedianUp(block)*** = median(MedianUp(host))

An example of classifying address blocks within one /24 block is shown on the right. There are 7 blocks indentified. 1 /26, 1 /27, 1 /28, 1 /32are classified as stable blocks, while 1 /26, 2 /27 blocks are classified as intermittent blocks.



Training: how to relate pings to usage? We define 5 ping-observable categories based on hostname-inferred usage categories

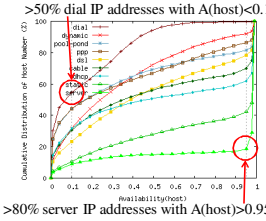
Hostname-inferred usage categories → Ping-observable categories

From ISC data: static, dynamic, dhcp, pool-pond, ppp, dial, dsl, cable,wireless, ded, biz, res, client, server, rtr-gw [Sullivan06].

From survey data:
- ***Always-stable*: A > 0.95, V < 0.0016 (N_u=1)**
- ***Sometimes-stable*: MedianUp >= 6hours**
- ***Intermittent*: MedianUp < 6hours**
- ***Underutilized*: A < 0.1**
- ***Unclassifiable*: < 20% responses**

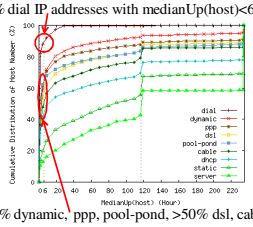
§We classified hostnames into usage categories using ISC data [ISC07].
§We relate usage to ping-observation by IP address.
§This process maps ping-observable categories to hostname-inferred usage categories.

>50% dial IP addresses with A(host)<0.1



>80% server IP addresses with A(host)>0.95

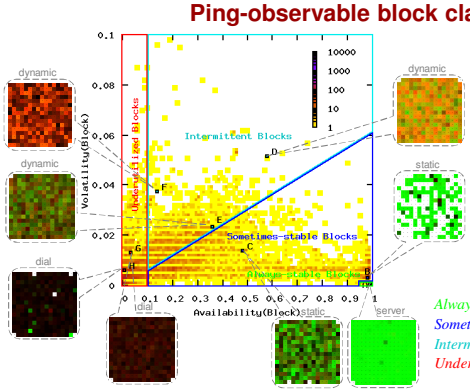
>90% dial IP addresses with medianUp(host)<6hours



>60% dynamic, ppp, pool-pond, >50% dsl, cable, >40% dhcp IP addresses with medianUp(host)<6hours

Results: Internet address blocks classified into 5 ping-observable categories

Ping-observable block classification example



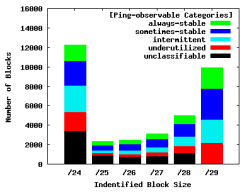
Density (A, V) plot at left show 12,198 indentified /24 blocks.
Eight examples around edge show (A, V) of each address in /24 blocks.

The density A/V plot shows 4 (without *unclassifiable*) ping-observable categories:

Always-stable: A > 0.95, V < 0.0016 (N_u=1)
Sometimes-stable: MedianUp >= 6hours
Intermittent: MedianUp < 6hours
Underutilized: A< 0.1

What are the sizes of blocks that show consistent usage?

Method: identify blocks by ping-observations.
Answer: many /24 blocks are used consistently (58.5% of all responded /24s), and most addresses are in a consistent /24 block. But, there are also many smaller consistent block sizes.
Validation: we see similar results in different survey [ITsurvey17w] and [ITsurvey16w]. The correlation coefficient is 0.9998.



Are there wasted blocks?

Method: look for underutilized blocks.
Answer: There are many underutilized blocks (22.7% of identified&classifiable /24s).
Warning: firewalls may lead to non-response, thus fail the results, see [Heidemann08].
Validation: we see similar results in different survey [ITsurvey17w] and [ITsurvey16w]. There are 22.5% of identified&classifiable /24s in [ITsurvey16w] are underutilized.

How many dynamic blocks? [Xie07]

Method: look for intermittent blocks.
Answer: There are many intermittent blocks (30.4% of identified&classifiable /24s).
Validation: to validate the ratio, we see similar results in different survey [ITsurvey17w] and [ITsurvey16w]. There are 28.3% of identified&classifiable /24s in [ITsurvey16w] are intermittent. To validate the correlation between *intermittent* and *dynamic*, we classified *338 hostname-inferred dynamic /24 blocks* into *ping-observable categories*. 48.2% of them are intermittent, 36.4% are underutilized. We see similar results in different surveys, the correlation coefficient is 0.9789.

Conclusion

§New 4-step methodology to map active probes(pings) to address usage.

§Ping-observation suggests
- the minimum of typical block size is /24,
- many wasted IPv4 blocks,
- many dynamic blocks, new method of discussing IP dynamics.

§Training with hostname-inferred usage data to categorize address blocks.

References

[Heidemann08] J. Heidemann, Y. Pradkin, R. Govindan, C. Papadopoulos, G. Bartlett, and J. Bannister. Census and Survey of the Visible Internet. *In Proceedings of the ACM Internet Measurement Conference*, p. to appear. Vouliagmeni, Greece, ACM. Oct. 2008.
[Xie07] Y. Xie, F. Yu, K. Achan, E. Gillum, M. Goldszmidt, and T. Wobber. How dynamic are IP addresses? In Proc. of the ACM SIGCOMM Conference, Kyoto, Japan, Aug. 2007. ACM.
[Sullivan06] M. Sullivan and L. Munoz. Suggested Generic DNS Naming Schemes for Large Networks and Unassigned Hosts. RFC draft:http://tools.ietf.org/wg/dnsop/draft-msullivan-dnsop-generic-naming-schemes-00.txt, 2006.
[ISC07] Internet Software Consortium. Internet Domain Survey, ISC DS-2007JAN. web page http://www.isc.org/ds, Jan. 2007.
[ITsurvey17w] USC/LANDER project. Scrambled Internet Trace Measurement, PREDICT ID USCLANDER/ internet address survey_reprobing_it17w-20070601. web page http://www.isi.edu/ant/lander, Jun. 2007.
[ITsurvey16w] USC/LANDER project. Scrambled Internet Trace Measurement, PREDICT ID USCLANDER/ internet address survey_reprobing_it16w-20070216. web page http://www.isi.edu/ant/lander, Feb. 2007.

Support for this work is through DHS contract NBCHC040137 (the LANDER project) as part of the PREDICT program, and NSF contract CNS-0626606 (the MADCAT project). Conclusions of this work are those of the authors and do not necessarily reflect the views of sponsors.
For more details, see http://www.isi.edu/ant/
August 2008