Autonomous System Partitioning and Policy Routing in the Japanese Internet

Kazuhiko Yamamoto [1]
Akira Kato [2]
Masaki Hirabaru [3]
Yuko Murayama [4]
Motonori Nakamura [5]
Noritoshi Demizu [6]

Abstract

As the number of connected networks increase and multiple service providers exist in a single autonomous system, the networks face routing problems including dependency on service providers, the chance of the appearance of unreachable networks, and overhead in terms of bandwidth used by routing protocols. This paper describes a case study on implementing autonomous system partitioning to solve the routing problems with the Japanese Internet. The authors examine a criterion of autonomous system partitioning for policy routing. It is shown that the policy that gives preference to links can be expressed by dividing autonomous systems which holds less preferred links into multiple autonomous systems.

I. Introduction

The first wide area IP network in Japan(the Japanese Internet) was established by the WIDE project in 1987. JAIN and TISN followed and were interconnected in 1989. The Japanese Internet initially adopted RIP[1] as its routing protocol because no other candidate was available at that time. This resulted in the formation of the Japanese Internet into a single AS. As the early network projects – WIDE, JAIN, and TISN – were administrated cooperatively, there had been no motivation to divide the Japanese Internet into multiple ASes. Since that time, the Japanese Internet has steadily grown and the number of connected IP networks presently exceeds 200. Moreover, the number of service providers has increased, and their policies have been diversified. Since all service providers make use only of RIP, the Japanese Internet still forms a single AS.

The growth of the Japanese Internet utilizing the traditional routing technology has presented routing problems. To solve these problems, JEPG/IP(Japan Engineering and Planning Group) and the authors are planning to divide the Japanese Internet into multiple ASes and combine current reliable routing technology, specifically BGP[2] and appropriate intra-domain routing protocols. Moreover, we are looking into methods of AS partitioning that will reflect the service providers' policies.

First, this paper specifies the routing problems with the Japanese Internet. Next, we propose an AS partitioning plan and explain how it solves these routing problems. Moreover, we examine a method for division of an AS, and show that policy that gives preference to links can be expressed by certain AS partitioning. Lastly, we consider policy which cannot be implemented by any AS partitioning or current routing technology.

II. Routing Problems with the Japanese Internet

This section describes urgent routing problems with the Japanese Internet that must be resolved.

II.A. Dependency between Service Providers

As all service providers in Japan utilize a single inter-domain routing protocol, RIP, the obtained paths are highly dependent on the network topology of related service providers. It is necessary for service providers to contact one another so that each will have knowledge of the whole topology, and therefore be able to coordinate its connection and metric. Thus, it is required to introduce technology that can hide the topology of each service provider, allowing service providers to independently design new paths for additions or changes within their own jurisdictions.

[1] Mr. Yamamoto belongs to Kyushu University. He may be reached at kazu@csce.kyushu-u.ac.jp

[2] Mr. Kato is with Keio University. He may be reached at kato@wide.ad.jp

[3] Dr. Hirabaru is with University of Tokyo. He may be reached at hi@nc.u-tokyo.ac.jp

[4] Dr. Murayama is with WIDE project. She may be reached at murayama@wide.ad.jp

[5] Mr. Nakamura belongs to Kyoto University. He may be reached at motonori@kuis.kyoto-u.ac.jp

[6] Mr. Demizu is with OMRON. He may be reached at demizu@nff.ncl.omron.co.jp

II.B. Limitation of RIP Metric

The diameter of the Japanese Internet exceeds 16 hops, which means that some networks may be unreachable from certain sites. The default routes are generated at each international gateway and announced to the Japanese Internet. They summarize all networks outside of Japan and all domestic networks with distances of over 16 hops from any given site. Thus, any two networks separated by 16 hops can communicate by the default route. However, there are two problems caused by limitations of the RIP metric:

(1) When a certain link is down and a backup path is selected, the distance between a network and an international gateway may exceed 16 hops.

(2) The limitation of 16 hops does not allow for further significant growth of the Japanese Internet.

Thus, in order to stabilize the Japanese Internet, the size limitation must be eliminated.

II.C. Overhead in Terms of Bandwidth

RIP announces routing information every 30 seconds even if there has been no topology change in the network during that interval. Related overhead in terms of bandwidth may be negligible in small networks, but the Japanese Internet now exceeds 200 integrated networks. The bandwidth consumed by routing information exchanged via RIP may not be insubstantial on slower links, particularly those of 64Kbps or lower. This is a more serious problem for service providers who depend on IP over X.25.

Moreover, multiple Class C addresses will be assigned to organizations according to CIDR[3] based address allocation. This will accelerate the growth of the bandwidth required for routing information exchanged via RIP. Thus, we expect that routing information will expand. For these reasons, a mechanism to reduce the overhead of routing information exchange is required.

II.D. Difficulty of a Path Design

Each service provider is interconnected with others at one or more connecting points. Some networks belong to multiple service providers. This results in a complex topology including multiple internet exchanges(Figure 1). It is difficult to design a backup path amongst such a large set of connections using a metric lower than 16. Therefore, it is essential to make backup path designs simple.

II.E. Inadequate Metric

Almost all links were established at 64Kbps in the early days of the Japanese Internet. Since 1992, however, the link speeds have begun to vary. Presently, hop counts do not reflect the differences in link speed. However, it is desirable to evaluate the metric based on link speed.

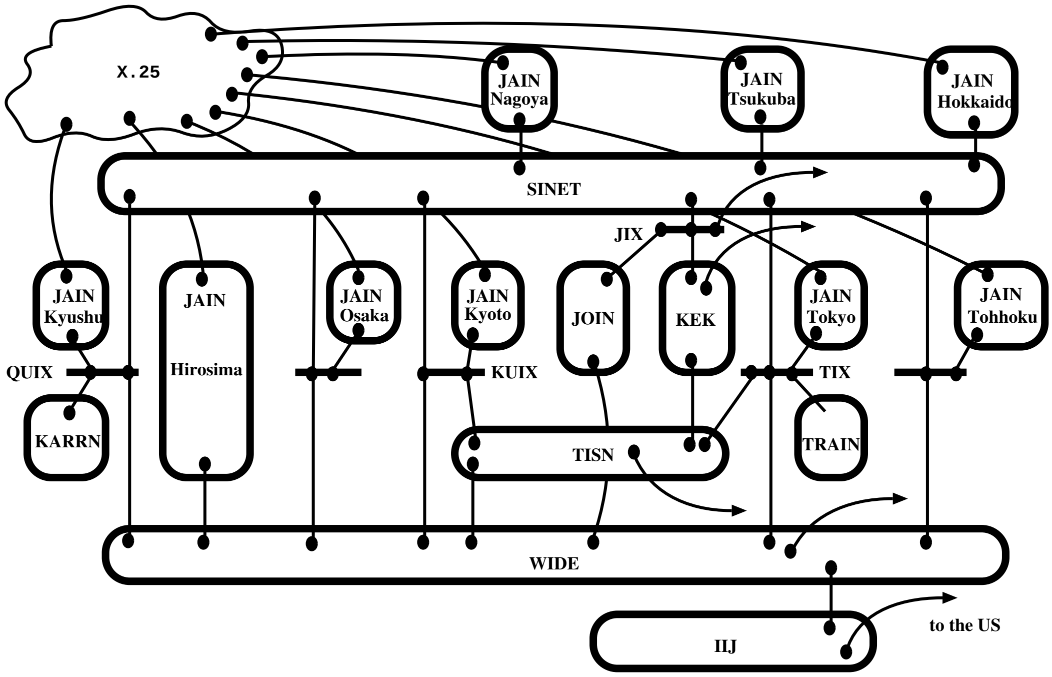

Figure 1: The abstracted topology of the Japanese Internet in Feb. 1993

II.F. Service Provider's Policies

The number of service providers continues to increase, and their policies continue to diversify. Rudimentary policy routing is executed by trusting or refusing neighbor gateways at internet exchanges which are implemented on an Ethernet. Examples of related concrete policy are described in section IV.

III. AS Partitioning of the Japanese Internet

To solve the urgent problems using currently available routing technology, the following plan is proposed:

(a) We will divide the Japanese Internet into multiple ASes. That is, a service provider forms one or more ASes by itself.

(b) BGP will be used as inter-domain routing protocol.

(c) The IX model[4] will be adapted to internet exchanges.

(d) Transit ASes will use appropriate intra-domain routing protocol(s) such as OSPF[5].

(e) Multi-homed ASes or stub ASes may use any intra-domain routing protocol(s).

The AS partitioning of the Japanese Internet serves to scale down each AS. This guarantees the independence of each service provider, and diminishes the diameter of each AS. BGP is a loop free inter-domain routing protocol which allows arbitrary connections of ASes. Thus, BGP is applicable to a large set of connections of service providers. Because BGP maintains the AS paths to destinations, path design becomes easier. We can reduce the bandwidth required to exchange routing information via BGP and a sophisticated intra-domain routing protocol such as OSPF, as they both utilize incremental updates. As each border gateway on an internet exchange can explicitly configure its connectivity by enumerating the neighbors with which it wished to form a peer connection, the current rudimentary policy routing can be maintained. The gateways to be replaced or reconfigured in this plan are border gateways of each service provider and some gateways in transit ASes. Since modification is not necessary for the large majority of gateways, it can be carried out.

IV. Policy Routing and Limitation of Current Routing Technology

This section provides examples of service providers' policies in Japan. Some of them can be implemented by AS partitioning, whereas others can not be represented by current routing technology.

IV.A. A Link Speed Problem

Consider two service providers which are operated cooperatively but their link speeds are different. In figure 2, the speed of a service provider (A)'s links L4 and L5 is much lower than that of a service provider (B)'s links L6 and L7. Links L1, L2, and L3 are shared by these service providers. Hosts H11, H12, and H13 and gateways G11, G12, and G13 belong to (A). Similarly, Hosts H21, H22, and H23 and gateways G21, G22, and G23 belong to (B). In this case, both service providers may want to implement the following policy:

(i) Traffic from (B) to (A) is required to traverse (B)'s link as far as possible.

(ii) Traffic from (A) to (B) is required to leave (A)'s link as soon as possible.

For example, traffic from H11 to H23 is required to traverse L1, L6, and L7 rather than traverse L4, L5, and L3. Traffic from H23 to H11 should traverse L7, L6, and L1 rather than L3, L5, and L4. If (A) and (B) form a single AS, and use BGP as an inter-domain routing protocol, requirement (i) is satisfied, as BGP can choose one link for one destination using an Inter-AS metric(or MULTI-EXIT-DISC). However, requirement (ii) is not satisfied as it is impossible for a network in one AS to use different links to another destination within a given AS. Therefore, it is better to divide (A) into multiple ASes in order to satisfy requirement (ii). In this case, if (A) is divided into three ASes, each including one gateway and one host, this requirement is satisfied. The reference[6] describes similar policy between NSFNet and DDN that were interconnected on the east and west coasts, however, EGP[7] could not express policy (ii).

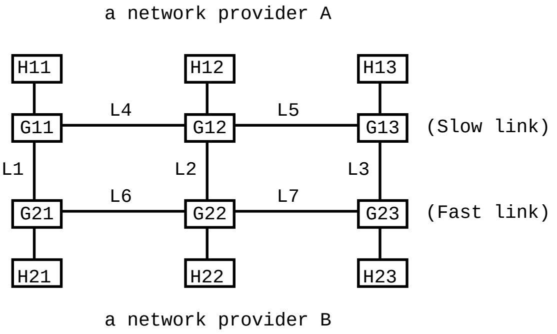

Figure 2: Two service providers which holds different speed links

IV.B. Overseas links

Some companies which have branches in the US have their own internal overseas links(Figure 3). From a funding perspective, it is natural for these companies to form single overseas ASes. Moreover, there is a technical motivation to make them single ASes. Traffic from the Japan side to the US side ought to traverse the company's internal link rather than a general link. Thus, the optimal traffic flow is as follows:

```
the Japanese Internet to Japan Side:
    the Japanese Internet
    -> the Japan side

the US Internet to the US side:
    the US Internet
    -> the US side

the Japanese Internet to the US side:
    the Japanese Internet
    -> the Japan side
    -> internal link
    -> the US side

US side to the Japanese Internet:
    the US side
    -> internal link
    -> the Japan side
    -> the Japanese Internet
```

If the Japan side and the US side are included in the Japanese Internet and the US Internet respectively, this policy can not be expressed as it is impossible to use a different link to one destination in a specific AS using BGP. Thus, the Japan side and the US side should form a single AS rather than being separated into the Japanese Internet and the US Internet respectively.

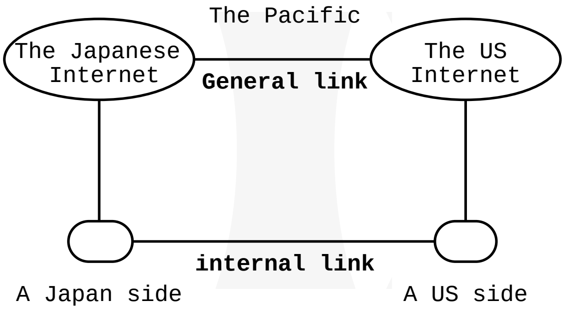

Figure 3: Overseas links

IV.C. Link Preference

Generalizing the two policies expressed in section IV.A and IV.B, preference will be given to each link that runs in parallel. A potential solution is the division of the domain holding the less preferred link into multiple ASes and making the domain that holds a preferred link into a sigle AS. Then, BGP is used as the inter-domain routing protocol. With this solution, there is one open question: Is it appropriate to divide a service provider into several ASes only to express its policy?

Another considerable solution is to develop an inter-domain routing protocol which can make different links to one destination within a specific AS. We consider that such a protocol can be implemented based on the current forwarding technology. Again, there comes another open question: Is it worth developing an inter-domain routing protocol to express such a domain's policy?

IV.D. Policy Limited by Packet Forwarding Technology

Consider a case in which two service providers invest in one link. In figure 4, a service provider (A) contains a network (a) and a gateway (e) and sponsors links L1, L4, and L6. Similarly, a service provider (B) contains a network (b) and a gateway (f) and sponsors links L2, L5, and L7. Gateways (c), (d), and network (g) belong to both service providers, and link L3 is sponsored by them. In this case, (a) wants to send its packets to (g) via (c), (d), and (e). (b) wants to send its packets to (g) via (c), (d), and (f). However, (c) cannot express this policy because present forwarding technology is based only on the destination. (A possible solution is to make use of tunneling technology.)

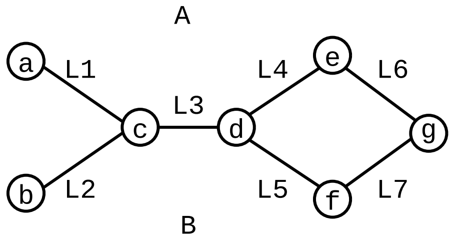

Figure 4: A link funded by two service providers

IV.E. Policy Limited by AS Technology

In figure 5, a transit AS (A) funded by the government carries non-profit traffic. That is, traffic from company to company cannot traverse (A). A university (a) is a member of (A). A multi-homed AS (B) consists of a university (b) and a company (c). (B) wants to use (A), which holds high speed link, to communicate with other ASes as much as possible. A multi-homed AS (C) which has a similar policy as (B) consists of a university (d) and a company (e).

Given these policies, traffic from (b) to (d), from (b) to (e), and from (c) to (d) can traverse (A). However, traffic from (c) to (e) should be carried by a direct link between (B) and (C). In order to implement these policies, we consider the following strategy. To make the example simple, we pick up communication from (C)'s member to (B)'s member. Thus, routing information follows from (B) to (C).

(B) announces routing information of (b) and (c) to (A). Then, it announces less preferred routing information of (b) and (c) to (C). (A) reannounces the routing information from (B) to (C). Although this strategy seems to express each service provider's policy precisely, it does not work as well as we might expect. That is, it is impossible to announce the routing information of (c) directly from (B) to only (e) in order to specify different paths to (c) for each source network. This is caused by the limitation of forwarding technology described in section IV.D. Moreover, this means one AS is not transparent to other ASes although the AS works well by itself. Thus, technology that treats all members of one AS similarly(e.g. AS-Path) is not enough to express this policy.

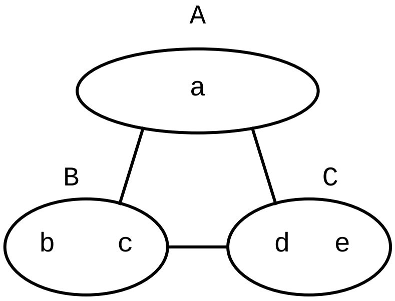

Figure 5: Transit AS which carries only non-profit traffic

V. Conclusion

This paper has enumerated the routing problems caused by the growth of the Japanese Internet – dependency between service providers, limitation of RIP metric, overhead in terms of bandwidth, difficulty of path design, and an inadequate metric. We proposed the AS partitioning plan utilizing currently available technology – (a) Each service provider forms one or more AS, (b) BGP is to be used as the inter-domain routing protocol, (c) The IX model will be adapted to internet exchanges, (d) Transit ASes will use appropriate intra-domain routing protocols, (e) Multi-homed ASes or stub ASes may use any intra-domain routing protocol(s). This plan can solve the routing problems in the Japanese Internet and ensure rudimentary policy routing.

We showed a criterion of AS partitioning for certain policy routing. The policy that gives links preference can be implemented by AS partitioning of service providers that hold less preferred links into multiple ASes. However, this left an open question as to whether or not it is appropriate to divide a service provider into several ASes only to express its policy. We examined policies that can not be expressed because of the limitations of forwarding and AS technology.

Acknowledgement

Authors are deeply grateful to members of WIDE policy routing working group for their discussion. Authors thank to Geoff Hustion for his discussion that is contained section IV.E. Authors would also like to express thanks to Kejiro Araki, Zengo Furukawa, and Keizo Saisho(Kyushu University) for their advice.

References

[1] C. Hedrick, "Routing Information Protocol", *RFC 1058*, 1988.

[2] K. Lougheed and Y. Rekhter, "A Border Gateway Protocol 3 (BGP-3)", *RFC 1267*, October 1991.

[3] V. Fuller, T. Li, J. Yu, and K. Varadhan, "Suppernetting: an Address Assignment and Aggregation Strategy", *RFC 1338*, June 1992.

[4] Geoff Huston, Elise Gerich, and Bernhard Stockman, "Connectivity within the Internet - A Commentary", *In Proceedings of INET '92*, pp. 133–142, Internet Society, 1992.

[5] J. Moy, "OSPF Version 2", *RFC 1247*, June 1991.

[6] J. Yu and H-W. Braun, "Routing between the NSFNET and the DDN", *RFC 1133*, November 1989.

[7] D. Mills, "Exterior Gateway Protocol Formal Specification", *RFC 904*, April 1984.

Author Information

Kazuhiko Yamamoto is a graduate student at Kyushu University's Department of Computer Science and Communication Engineering. He graduated in Kyushu University at 1992. He is interested in policy routing, network security, and education for beginners of internet. He is a member of the Internet Society.

Akira Kato is a research associate of Keio University since April 1990 and has joined a development group of their campus networking system at Shonan Fujisawa campus. He is also working for WIDE Project before its establishment. He received his bachelor and master of engineering from Tokyo Institute of Technology in 1984 and 1986 respectively. He is interested in the internet routing technologies and is a chair of routing working group of Japanese Internet Engineering Planning Group for IP. Mr. Kato is a member of IEICE, IPSJ, ACM, Usenix, Japan Unix Society and a pioneer member of the Internet Society.

Masaki Hirabaru is an associate professor of Kyushu University, Japan. He is in charge of many Japanese network activities, mainly, JAIN, WIDE, regional networks, and JPNIC. He received the B.E., M.E. and D.E. degrees in computer science from Kyushu University in 1983, 1985 and 1989, respectively. He is a member of IEEE, ACM, Internet Society, and IPSJ.

Yuko Murayama has been a research member of the WIDE Project of Japan since March 1992. Within WIDE, she is the leader of the Policy Routing Working Group, and her current interest is in routing by preference. Dr. Murayama received her B.S. in mathematics from Tsuda College in Japan, and her M.S. and Ph.D. from the University of London, where she was enrolled in the University College of London's (UCL) Department of Computer Science. Her research topic at UCL was configuration detection in computer networks.

Motonori Nakamura is a doctoral student of Kyoto University. He graduated from Kyoto University with a major in Information Science in 1989, and went on to earn his masters degree, also in Information Science, from the same institution in 1991. His research areas are vector/parallel processing and computer networking. He is a member of the Internet Society, the Information Processing Society of Japan, and the Japan Society for Software Science and Technology.

Noritoshi Demizu is with OMRON Corporation for five years, where he developed windowing software and has been maintaining network environment there. He also joined Graduate School of Information Science, Advanced Institute of Science and Technology, Nara, Japan in 1993, where he has been conducting research on "ddt" as versatile IP over IP technology. Mr. Demizu received his B.S. from Kyoto University in 1988. He has also been a member of WIDE project since 1991.