采用泛播路由构建高效中继路由系统*

郑健平[1,2+], 李克勤[3], 孙利民[1], 吴志美[1]

[1](中国科学院 软件研究所 多媒体通信与网络工程研究中心,北京 100080)

[2](中国科学院 研究生院,北京 100080)

[3](朗讯科技 贝尔实验室中国基础科学研究院,北京 100080)

The Use of Anycast Routing for Building Efficient Relay Routing System

ZHENG Jian-Ping[1,2+], LI Ke-Qin[3], SUN Li-Min[1], WU Zhi-Mei[1]

[1](Multimedia Communication Center, Institute of Software, The Chinese Academy of Sciences, Beijing 100080, China)

[2](Graduate School of the Chinese Academy of Sciences, Beijing 100080, China)

[3](Bell Labs Research China, Lucent Technologies, Beijing 100080, China)

+ Corresponding author: Phn: +86-10-62645408, Fax: +86-10-62645410, E-mail: jianping@ios.cn, http://www.iscas.ac.cn

Received 2003-11-06; Accepted 2004-10-09

Zheng JP, Li KQ, Sun LM, Wu ZM. The use of anycast routing for building efficient relay routing system. *Journal of Software*, 2005,16(5):1021–1027. DOI: 10.1360/jos161021

Abstract: Relay routing system consists of a set of relay routers, providing relay routing for routing domains that cannot exchange routing information. The key point of the system is to configure appropriate relay router for routing domains. In this paper, all the relay routers are taken as a virtual node by assigning them an anycast address which is accessed along the shortest path by anycast routing. In addition, source routing is used to route the data packets to relay router. The anycast-based relay routing system enables auto-configuration of relay routing, improves the performance and the reliability of relay routing, and it is compatible with existing network infrastructure with quite low overhead.

Key words: anycast; relay routing; routing system

摘 要: 中继路由系统由一组中继路由器组成,为不能交换路由信息的路由域提供中继路由.该系统的关键是为路由域配置恰当的中继路由器.为所有中继路由器分配一个泛播地址,将它们当作一个逻辑节点,借助泛播路由以最短路径到达该逻辑节点.此外,采用源路由的方法将数据报文路由至中继路由器.基于泛播的中继路由系统实现了中继路由的自动配置,提高了中继路由的性能和可靠性,并且与现有网络系统兼容,实施代价很小.

* Supported by the National Natural Science Foundation of China under Grant No.60272078 (国家自然科学基金)

ZHENG Jian-Ping was born in 1976. He is a Ph.D. candidate at the Institute of Software, the Chinese Academy of Sciences. His current research areas are Computer Network and Multimedia Communications. **LI Ke-Qin** was born in 1973. He is a Member of Technical Staff of Bell Labs Research China, Lucent Technologies. His current research areas are IPv6, Routing Protocols and Network Protocol Testing. **SUN Li-Min** was born in 1966. He is a professor at the Institute of Software, the Chinese Academy of Sciences. His research areas are Wireless Network and Broadband Access Network. **WU Zhi-Mei** was born in 1942. He is a professor and doctoral supervisor at the Institute of Software, the Chinese Academy of Sciences. His research areas are Computer Network and Multimedia Communications.

关键词: 泛播;中继路由;路由系统
中图法分类号: TP393 文献标识码: A

1 Introduction

Internet consists of thousands of routing domains, each of which having a set of routers that exchange internal routing information within an administrative domain. The exchange of external routing information between routing domains is via inter-domain routing protocol like BGP. However, there maybe exists such a case that no inter-domain routing protocol is provided for two different routing domains, or they are not willing to exchange routing information through inter-domain routing protocol due to the security or administration policy. In such cases, relay routing system is a reasonable solution to enable the communication between these two routing domains. A set of routers that can reach both of the two routing domains are deployed to provide relay routing for these two domains. The key point of the system is to configure appropriate relay router at the border routers of routing domains.

Statically binding a relay router at each border router of a routing domain seems to be straightforward for such configuration. However, when there are a large number of border routers in the routing domain, finding and configuring a default relay router manually is a significant work in practice. Moreover, static configuration may potentially cause trouble. For example, when the assigned relay router suddenly goes down, the border router will not be able to reach another routing domain any longer. In addition, due to lack of the available information of relay routers, the border router may configure a relay router that is non-optimal for itself, or optimal initially but becomes non-optimal later with the change of route to relay router or performance fall in the relay router, leading to poor performance of relay routing. For the above reasons, statically binding scheme is not good for relay routing system, which demands a more efficient scheme to be applied in relay routing system.

Anycast, which was first defined in RFC 1546[1], is a network service for a host to communicate with one member in a designated group. It can dynamically choose a “best” one from the group. In this paper, we take advantage of such feature of anycast communication, and present an anycast based relay routing system. All the relay routers are regarded as a virtual node by sharing the same anycast address, providing relay routing for two separated routing domains. The border routers of one routing domain take the anycast address as the route to reach another routing domain, and this route will be directed to a nearest relay router by anycast routing. Anycast based relay routing system avoids the problems in statically binding scheme and improves the efficiency and reliability of relay routing.

The rest of this paper is organized as follows. Section 2 provides necessary background information of anycast. Section 3 gives an overview of the system. Section 4 presents the design details of the system. Section 5 evaluates the performance of the system. Section 6 concludes this paper.

2 Background: Anycast Communication

Anycast is a network service for a host to communicate with one member in a designated group. Members in the group share an anycast address, which represents a virtual node in the network providing a certain kind of service. Host that wants to get such a service will send datagram to the anycast address and the internetwork is responsible for delivering the datagram to the nearest server, where ‘nearest’ is defined according to the routing system’s measure of distance[1]. Thus, accessing the nearest server enhances the performance perceived by the host, saves the network’s bandwidth and provides the desired service. Figure 1 illustrates an example of anycast

communication. Member1 and Member2 are in the same anycast group A. If Sender1 sends a packet to the anycast address, the network delivers it to Member1, while if Sender2 sends a packet to the anycast address, the network delivers it to Member2.

To enable anycast communication, anycast routing is necessary to build routing table for forwarding anycast packets to the nearest anycast server. Katabi[2] proposes a framework for scalable global IP-Anycast, and Jia[3] addresses the problem of routing table establishment and packet forwarding for anycast messages. There are two kinds of approaches to build anycast routing table. The first kind of approach extends the existing unicast routing protocol such as RIP (Routing Information Protocol) and OSPF (Open Shortest Path First) to support anycast routing. Discussion on how to modify RIP and OSPF to support anycast routing can be found in Ref.[4]. The second kind of approach builds a distribution tree for anycast group just as multicast protocol. The anycast packet is forwarded along the distribution tree. But different from multicast, the packet is only forwarded to one downstream at the branch node. In Ref.[3], Source-Based Tree (SBT) method and Core-Based Tree (CBT) method are proposed for building anycast distribution tree.

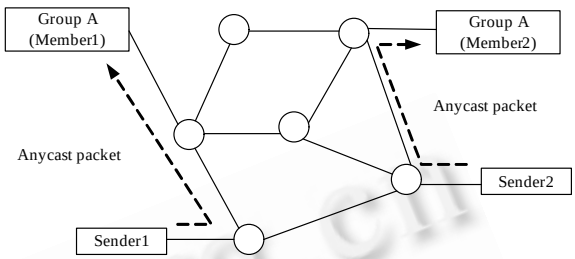


Fig.1 Illustration of anycast communication

Bhattacharjee[5] generalized the anycast communication paradigm and proposed to provide anycast service at the application layer. This approach attempts to build a directory system which, queried with a service name and a client address, returns the unicast address of the server that is nearest to the client.

Since more and more applications demand anycast services, in the latest version of IP specification, IPv6, anycast has been defined as a standard network service[6].

3 System Overview

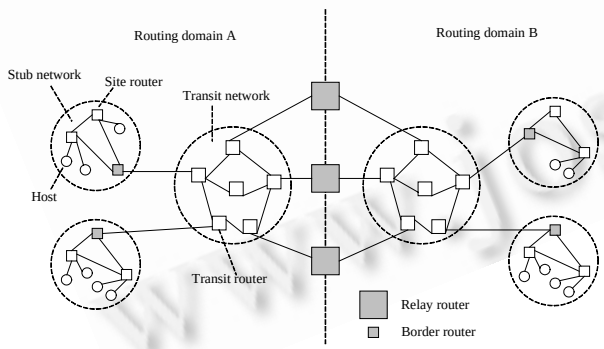


Fig.2 Relay routing system

Figure 2 illustrates an internetwork composed of two separated routing domains A and B. A and B don't exchange routing information with each other, but they are both connected with a set of relay routers. The relay routers participate in the routing system of both A and B respectively, but don't advertise the routing information of A to B and vice versa. Communication between A and B must be through one of the relay routers. The border router is the default gateway for its stub network, so a packet from inside the stub network to another routing domain will be routed to the border router. Then the border router routes this packet to one of the relay routers which relays it to the destination.

To route the packets destined to another routing domain, the border router should be configured with a default relay router to reach another routing domain. Provided that there are multiple relay routers available, how does the border router configure its default relay router? As discussed in Section 1, binding a fixed relay router for the border router is not a good solution. Instead, we use automatic configuration at border router with anycast mechanism to dynamically select a nearest relay router.

In this anycast based relay routing system, all the relay routers are regarded as a virtual node by assigning them an anycast address. The anycast route is advertised from all the relay routers and hence broadcasted to routing domain A and B respectively. Thus all the transit routers in the network keep one routing entry to reach the nearest relay router. All the border routers configure this anycast address as route to reach another routing domain, and the transit routers will direct this route to a nearest relay router.

4 Design Details

4.1 Anycast routing

To enable anycast routing, the relay routers should be assigned an anycast address. There are two ways to support anycast address; one uses the space of unicast address while another allocates a special class of IP address for anycast. In our system, we allocate unicast address *X* as anycast address. Such an allocating scheme can simplify the anycast routing since the anycast route can be handled as host route by the unicast routing protocol, so that no special anycast routing protocol is needed for the relay routing system.

Each relay router configures the anycast address *X* in its loopback interface, and advertises this anycast route to unicast routing system of routing domain A and B. The unicast routing system takes it as host route and broadcasts it all over its routing domain. When the transit router or border router receives this route for the first time, it will built a new anycast routing entry, and will update it upon receiving a preferred one. Thus the border and transit router will keep a route to the nearest relay router. Since all the relay routers periodically advertises such anycast route to the network, the border and transit router will always keep an up-to-date route to the nearest relay router.

The routing metric, i.e. how to define ‘nearest’ relay router, should be selected when designing an anycast system. There are a variety of routing metrics including hop count, round trip time (RTT), link cost, service response time and etc. The selection of routing metric should take the target of anycast system and the cost of providing the metric into account. The relay routing system is designed for relaying the communication between two separated routing domains, which is application-independent. This justifies selecting network layer metrics such as hop count and RTT. Here, we select hop count as routing metric since it can be provided by unicast routing protocol such as RIP and OSPF, and it is relatively stable. Thus, the anycast relay routing system adds no extra burden to unicast routing protocol.

Routing table at B

dest	hop count	next hop
A	2	D
C	1	C
D	1	D
E	2	C
F	2	D

Routing table at D

dest	hop count	next hop
A	1	A
B	1	B
C	2	E
E	1	E
F	1	F

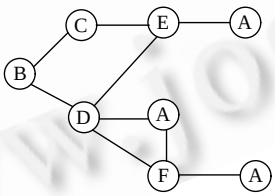


Fig.3 Example of anycast relay routing

Figure 3 gives an example of anycast relay routing. In this figure, A represents anycast relay routers, B is one of the border routers, and others are transit routers. After receiving anycast routing advertisement from all of the relay routers, the transit routers and border router will keep a routing entry to the nearest relay router in the routing table. In this example, B takes D as next hop to reach the nearest A, and D connects A directly. Thus the anycast routing automatically establish a shortest path from B to A.

4.2 Data transmission

Once the anycast routing for the relay routers has been established, communication between the two separated routing domains can be enabled. When the border router receives a packet destined to another routing domain, it shall first route this packet to the relay router which relays it to the destination. However, the destination field in the IP header is not the relay router. Thus special mechanism is needed for the border router to route this packet to the relay router. One such mechanism is IP tunnel, i.e. the border router encapsulates this packet in a new IP packet,

whose source is the border router and the destination is relay router, and when the relay router receives the encapsulating packet, it gets the original packet and forwards it to the destination. The disadvantage of IP tunnel is that the overhead of encapsulating and decapsulating in the border router and relay router is heavy, which will decrease the routing performance. So our system doesn't use IP tunnel. Instead, it uses a standard IP option – "Loose Source and Record Route (LSRR)" in IPv4[7] or "Routing Header" in IPv6[6].

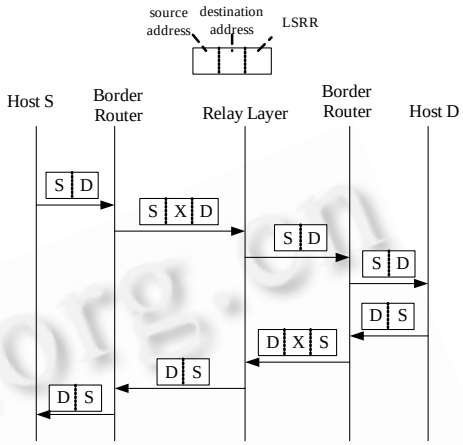


Fig.4 Data transmission in relay routing

The LSRR option or Routing Header can provides the intermediate routing information in forwarding an IP datagram to the destination. In our relay routing system, the routing information is not filled by the source but by the border router. This makes relay routing totally transparent to the end hosts so that the system requires no modification to host implementation. Supposed that a host *S* in one routing domain sends a packet to host *D* in another routing domain. When the border router of *S* receives this packet, it adds a LSRR option in the packet. In addition, it replaces the destination address *D* in the IP header with relay router anycast address *X*, and sets the address in the LSRR option with *D*. The modified packet is then routed by the transit routers to the nearest relay router according to the anycast routing. The transit routers don't need to handle the LSRR option, and it forwards this packet just as a general unicast packet. When the relay router receives this packet, it gets the address *D* from the LSRR option, sets the destination address in the IP header as *D*, deletes the LSRR option, and forwards it. The routing system in another domain will finally route this packet to host *D*. The packet returned from *D* to *S* is routed in the same way. Figure 4 illustrates the data transmission process between *S* and *D*.

5 Performance Evaluation

5.1 Efficiency

The current Internet with BGP as inter-domain routing protocol often experiences a low routing efficiency in the sense that many sub-optimal paths are used instead of optimal paths. For example, as reported in Ref.[8], for about 50% of the paths measured, there exists an alternative route with lower latency. In comparison, anycast based relay routing system selects the nearest relay router to relay the communication between two separated routing domains. It optimizes the route from source to relay router. In the following, we analyze the routing efficiency of anycast based scheme, in terms of hop count from source to relay router.

Supposed that there are N relay routers R_i (i=1,…,N), and we have the following definitions and hypotheses.

Definition 1. X_i: the hopcount of shortest path from source S to relay router R_i (i=1,…,N).

Definition 2. $f_{Xi}(h)$ is the probability that the hopcount from source to relay router R_i equals h, which can be denoted as $f_{Xi}(h)=P\{X_i=h\}$,(i=1,...,N).

Definition 3. $f_u(h,k)$ is the probability that the minimum value of X_i is h, and there are just k relay routers from source to each of the hopcounts, which can be denoted as

$f_u(h,k)=P\{\min\{X_i|\ i=1,\ldots,N\}=h\wedge\exists(X_{m1},X_{m2},\ldots,X_{mk}),(X_{m1}=X_{m2}=\ldots=X_{mk}=h);\ 1\le m_j\le N,\ j=1,\ldots,k\}$

Definition 4. $f_M(h)$ is the probability that the minimum hopcount from source to relay routers equals h, which can be denoted as $f_M(h)=P\{\min\{X_i|\ i=1,\ldots,N\}=h\}$

Hypothesis 1. All $f_{Xi}(h)$ are independently and identically distributed, and $f_{Xi}(h)=f_X(h)$ (i=1,…,N).

That the minimum hopcount equals h means that there is at least one path whose hopcount(s) equals h and the remains are equal to or greater than h. Since $f_{Xi}(h)$ (i=1,…,N) are identical and independent of each other, we have

$$f_U(h,k) = C_N^k f_X(h)^k (\sum_{j>k} f_X(j))^{N-k} \quad (1)$$

and k can be from 1 to N, so

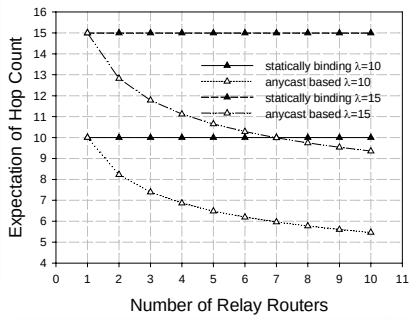
$$f_M(h) = \sum_{k=1}^{N} f_U(h,k) \quad (2)$$

With (1) and (2),

$$f_M(h) = \sum_{k=1}^{N} C_N^k f_X(h)^k (\sum_{j>k} f_X(j))^{N-k} \quad (3)$$

In statically binding scheme, the border router has no information about distance to each of the relay router, and it just configures one of them as default relay router arbitrarily. So the probability distribution of the hopcount from source to relay can be represented with $f_X(h)$. On the other hand, the nearest relay router is selected in anycast based scheme, in which the probability distribution of the hopcount from source to relay is $f_M(h)$.

Recent researches on Internet measurement[9] found that there's a scaling law for the hopcount of the shortest path between two arbitrary nodes in the Internet and it can be approximated by Poisson Law, which is denoted as $f(h)=\frac{e^{-\lambda}\lambda^h}{h!}$, where λ is the average of hopcount in Internet. The observed results based on measurement showed that λ varies from 14 to 20 over different periods and for different continents such as Europe, North America and Asia-Pacific. For the whole Internet over a long period, λ is about 16. Since source and relay routers are two nodes in the Internet, it is reasonable to take $f_X(h)$ as one instance of $f(h)$. Supposed that the average hopcount between source and relay router is λ_1, then $f_X(h)=\frac{e^{-\lambda_1}\lambda_1^h}{h!}$.



Given the probability distribution of hop count, we can further calculate the expectation of hop count from source to relay router. Figure 5 illustrates the expectation of hop count with the variation of number of relay routers for λ_1=10 and λ_1=15 respectively. It is clear that anycast based scheme improves the efficiency of relay routing.

5.2 Reliability

As the Internet's inter-domain routing protocol, the Border Gateway Protocol (BGP) is crucial to the overall reliability of the Internet. Faults in BGP implementations or mistakes in the way it is used have been known to disrupt large regions of the Internet. However, the configuration errors are pervasive, with 200~1200 prefixes (0.2%~1.0% of the BGP table size) suffering from misconfiguration each day[10]. As for the relay routing system, the reliability refers to the probability that packet from one routing domain to another domain can be successfully relayed by a relay router. In the scheme of statically binding a relay router at the border router, when the default relay router goes down, this border router can not reach another routing domain any more even if other relay routers still work. On the contrary, in the anycast based relay routing system, the border router will be served as long as there is one relay router working, since the anycast routing system is aware of the state of all relay routers and will always provide one available relay router for the border router. Supposed that there are N relay routers in the system, and the reliability of each relay router is P_i (i=1,…,N), then the reliability of the system is P_i in statically

binding scheme, and $1-(1-P_i)^N$ in anycast based scheme. As an instance, if N is 5 and $P_i(i=1,\ldots,5)$ is 0.8, the reliability of the system will be 0.8 in statically binding scheme and 0.99968 in anycast based scheme. Obviously, anycast mechanism greatly improves the reliability of the relay routing system.

5.3 Overhead

The anycast based relay routing system enables auto-configuration at border routers, and it doesn't require any modification in the end hosts and transit routers, nor it requires keeping any additional state in the routers. Moreover, the anycast routing is maintained by the existing unicast routing system with a simple extension. So it is nearly compatible with existing network infrastructure. The only overhead of the anycast based relay routing system is the cost of processing of LSRR option in the border routers and relay routers. However, LSRR option is a standard IP option that is supported by all router implementation, and processing this option is light weight. In summary, the overhead of the system is quite low.

6 Conclusion

In this paper, we propose an anycast based relay routing system, which enables auto-configuration at border routers and improves the system reliability and the efficiency of relay routing. This system is transparent to end hosts and transit routers, and is compatible with the existing network infrastructure with quite low overhead. Moreover, the architecture of this system can be applied in both IPv4 and IPv6 network.

Although this system is designed for relay routing, the design rationale is not limited to relay routing, but can be adopted in the deployment of replicated servers that provide proxy service, e.g. web proxy service. Designing a general architecture of anycast proxy system is left to be our future work.

Acknowledgement We are indebted to Dr. David Lee, Hao Ruibing, Huang Dawei and Ma Juntao, for the insightful comments and stimulating discussions.

References:

[1] Partridge C, Mendz T, Milliken W. Host anycast service. IETF RFC 1546, 1993.

[2] Katabi D, Wroclawski J. A framework for scalable global IP-Anycast (GIA). In: Proc. of the ACM SIGCOMM. 2000.

[3] Xuan D, Jia W, Zhao W. A routing protocol for anycast messages. IEEE Trans. on Parallel and Distributed Systems, 2000,11(6): 571−588.

[4] Basturk E, Engel R, Haas R, Kandlur D, Peris V, Saha D. Using network layer anycast for load distribution in the Internet. Techical Report, IBM T.J. Watson Research Center, 1997.

[5] Bhattacharjee S, Ammar MH, Zegura EW. Application-Layer anycasting. In: Proc. of the IEEE INFOCOM. 1997.

[6] Deering S, Hinden R. Internet Protocol Version 6 (IPv6) Specification. IETF RFC 2460, 1998.

[7] Postel J. DARPA Internet program protocol specification. IETF RFC 791, 1981.

[8] Savage S, Anderson T, Aggarwal A, David Becker N, Cardwell A, Collins EH, John Snell A, Vahdat GV, Zahorjan J. Detour: A case for informed Internet routing and transport. IEEE Micro, 1999,19(1):50−59.

[9] Begtasevic F, Van Mieghem P. Measurements of the hopcount in Internet. In: Proc. of the Passive and Active Measurement. 2001.

[10] Mahajan R, Wetherall D, Anderson T. Understanding BGP misconfiguration. In: Proc. of the ACM SIGCOMM. 2002.