L2 Switch Feature for Virtual Router Redundancy Protocol Fast Convergence

Hiroshi Matsuda
Faculty of Econoinformatic, Himeji Dokkyo University
Hyogo, JAPAN

ABSTRACT
Computer networks have played a central role at any circumstances in society today. So it is necessary for them to be available at any time. To this end we often use redundant technologies such as duplex systems to overcome failures. Virtual Router Redundancy Protocol (VRRP) is one of the redundant technologies, which gives detours to end hosts in case of trouble. Because of its advantages such that end hosts have no requirement for using VRRP it has been widely used in many Local Area Networks. In this paper we propose a new method on VRRP. Our method makes VRRP convergence time shorter in a typical network design: it means that through using our method computer networks recovery faster when failures occurred.

General Terms
Network Design, Router, L2 switch

Keywords
First Hop Redundancy Protocol, VRRP, Gratuitous ARP

1. INTRODUCTION
In most of Local Area Networks (LAN), end hosts such as personal computers are connected to L2 switches, which are referred to as switching hubs in regard to Ethernet technology. And through L2 switches end hosts connect a predefined router. This router is referred to as a default gateway and any end host must send an Ethernet frame to its default gateway whenever the destination address of the frame doesn't belong to the same network as the source address of the frame.

Unavailability of default gateways, therefore, would practically separate end hosts from LAN and make them isolated. So, to avoid isolations of end hosts several methods such as routing protocol and ICMP router discovery protocol (IRDP) have been developed so far. Those methods, however, are disadvantageous; some end hosts cannot run any routing protocol or it takes too long to converge in using IRDP.

To overcome those shortcomings First Hop Redundancy Protocols (FHRP) are developed. End hosts, in general, know only the IP address of the default gateway assigned to them, which may be configured either manually or through dynamic host configuration protocol (DHCP). FHRPs presuppose this situation, which means that there is no need for end stations to run any routing protocols. Thanks to this precondition and other advantages such as fast convergence time, FHRPs have been widely used in many LANs. And in several areas they have been under intense investigation [1]-[3].

In terms of development entities, FHRPs are divided into two groups : One group is called vendor-specific protocol, which includes Hot Standby Router Protocol(HSRP) developed by Cisco Systems[4] and the other one is VRRP, which is a standard protocol developed by the Internet Engineering Task Force(IETF) [5]. But in concept both protocols are alike and use the same idea called by “Virtual Routers”.

In this paper we propose a new method on VRRP. Our method makes VRRP convergence time shorter in a typical network design: it means that through using our method computer networks recovery faster when failures occurred. This paper is organized as follows: Section two gives a brief description of VRRP enough to understand this paper. And section three describes the network model considered in this paper and points out a problem when VRRP is running on this network model. Section four proposes our method to solve the problem and finally section five gives conclusion.

2. VRRP PRINCIPLE FOR OPERATION
In this section we give a brief description of VRRP enough to understand this paper. And at the end of this section we summarize the key points of VRRP from the view point of our proposed method.

VRRP is a standard protocol developed by the Internet Engineering Task Force (IETF) and has been widely used in many LANs. As stated in the previous section, in TCP/IP protocol suits end hosts must send an Ethernet frame to their default gateways whenever the destination address of the frame doesn't belong to the same network as the source address of the frame. Thus, unless the host knows the Medium Access Control (MAC) Address of the default gateway assigned to the host, it must send an Address Resolution Protocol (ARP) request packet destined to broadcast address to obtain it. VRRP utilizes this mapping between the MAC address and IP address of the assigned default gateway.

In VRRP operation several routers constitute a group. And in the group each router has its own priority and one router with the highest priority is selected as a representative of the group. VRRP defines how each router selects a master in a group by themselves. This representative router and the other routers in the group are respectively referred to as a master and backups. Once a master selected, it behaves as a default gateway. This means that to an ARP request for the IP address of the assigned default gateway only a master replies to it with the MAC address predefined in VRRP. Here note that the MAC address is not the one physically burned in an interface of the master. The MAC address is defined in [5] as follows: 00-00-5E-00-01-{VRID} (in hex in internet standard bit-order), where {VRID} means VRRP group number from 0 through FF. Consequently, the end host getting an ARP reply from the master regards the answering master as the assigned default gateway to the end host.

Once a master determined, then, how do other routers in the group know that the master is still alive? To show its status to any routers in a group, a master periodically sends a tiny packet. This tiny packet is referred to as an advertisement packet. Because its destination address of an advertisement packet is the multicast address predefined in VRRP, L2 switches receiving an advertisement packet floods it through all ports to any router in a group. Therefore, unless getting no Advertisement packets in the predefined interval, any router

other than the master evaluates that the master is unavailable and begins the election procedure to replace the dead master. This predefined interval is referred to as "master down interval" [5]. After the election procedure the new master sends a gratuitous ARP frame [6]. The source MAC address of this frame is the predefined MAC address stated above. This frame enables L2 switches to relearn where the new master is.

And in addition to the above basic functions, VRRP provides a useful mechanism which forces a backup to change its role immediately; a backup would become a master in a fairy shorter time than usual when it receives the advertisement packet whose priority is zero.

Finally, we summarize the key points of VRRP operation as follows.

P1: The source MAC address of any master is predefined.

P2: A master periodically sends an advertisement frame as a multicast frame.

P3: When a backup changes its role the new master sends a gratuitous ARP frame.

P4: An advertisement frame with priority zero makes a backup change its role immediately.

Our method utilizes these key points to make convergence in VRRP

3. NETWORK MODEL

In this section we show the network model considered in this paper and point out a problem when VRRP is running on this network model.

Consider the network in Fig.1, which has two routers and one L2 switch.

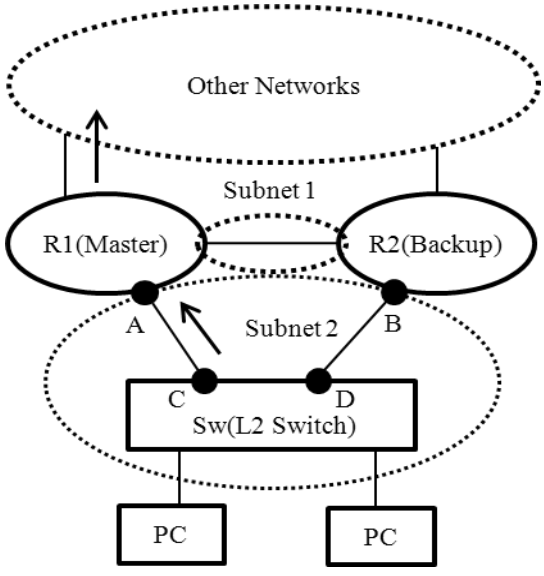

Fig 1: Network Model

In this network both routers run VRRP at the interfaces in subnet 2 and Router1 is configured as a master which uses the real interface address A as the primary IP address in VRRP: end stations in subnet 2 use the IP address A as their default gateway address. Note here that link aggregation method such as IEEE802.3ad [7] can be used on the link which connects each router to the L2 switch.

In Fig.1 subnet 1 could be equal to subnet 2 to avoid complicated configurations or several L2 switches could be in subnet 2 to accommodate many end stations. The configurations on those conditions, however, would require running such an L2 protocol as rapid spanning tree protocol (RSTP) [8] to ensure no loops at L2 level in the network. This means that it takes a longer convergence time during system failures on those networks. So in Fig.1 subnet1 is defined as different from subnet 2. In real networks this configuration is widely used with the same reason.

Now suppose that a failure occurred at the link between Router1 and Sw1 and made them separated. In this case, the backup router, Router 2, cannot in principle recognize that the master router is currently unavailable until Master Down Interval has elapsed (Master Down Interval is the predefined time interval in [5] and about three times the interval of an advertisement frame).

Considering running applications such as IP telephony and web conference system, which have little tolerance for time delay, it is desirable to make unavailable time as short as possible in modern networks. Then, could we make Master Down Interval as short as we want? Unfortunately it could not because any L2 switches and routers have some buffers in them. If a sequence of advertisement frames should be in some buffers during Master Down Interval, a backup router judges that a master is dead and changes its role. And then if some advertisement frame arrives the new master it changes its role immediately again. In the end, repeating this role changing causes significant dysfunction in the network. In the next section we propose a new method to overcome this problem.

4. PROPOSED METHOD

In this section we propose a new method to make VRRP convergence time shorter. Although essentially VRRP is a protocol implemented on routers, our method adds a new feature to a L2 switch placed between two VRRP-routers.

Consider the network in Fig.1 again, where two routers R1 and R2 are running VRRP in subnet 2 and R1 is a master. In this situation the following facts are easily seen:

1. If the link between the Router R1 and the L2 switch Sw should be disconnected, Sw can immediately notice it.
2. From P1 and P2 presented in section three, Sw can know which interface a master is connected and what group in VRRP the master belongs dynamically from MAC addresses.
3. From P3 presented in section three, Sw can know which interface a new master is connected.

Our method utilizes these facts and is broken down into the following step-by-step procedure.

1. In advance keep a temporary buffer in Sw to hold some received frames and specify the group number which you want to track if you need it. Suppose R1 as the group master. Note that it is possible to track some groups concurrently although extra buffers required.
2. Immediately after the link connected to R1(a master) should be disconnected, Sw moves the frames, which are stored in the buffer at C or determined to forward to C, to the temporary buffer in Sw.
3. Sw sends an advertisement frame for the group, which has priority value zero.

4. Immediately after Sw learns dynamically the interface (D) connected to the new master (R2) by the gratuitous ARP frame sent by R2, Sw moves the frames in the temporary buffer to the interface D.

Fig.2 shows a simplified description of our method.

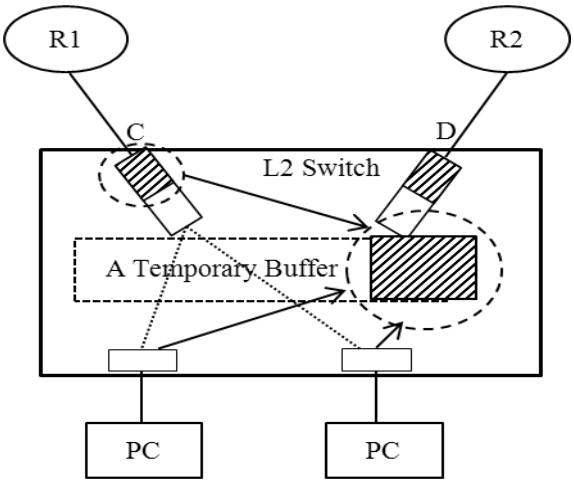

Fig 2: Simplified Description of Our Method

As above stated, our method makes a backup change its role without waiting Master Down Interval. It, therefore, shorten an unavailable time of network in principle. Note here that our method doesn't require any modification of VRRP itself. In addition to this, unlike IGMP snooping [9], our method imposes no additional feature which investigates data included in any Ethernet frame upon L2 switches; All that we have to know consist in Layer 2 level. Furthermore our method requires no additional line between the two routers to convey their statuses each other unlike Virtual Switching System (VSS) [10].

To differentiate L2 switches as commercial products it is useful to implement our method into them as a fast-recovery feature. In this case a network administrator needs only to configure one command once for the following reasons. Now suppose that an L2 switch has an interface whose bandwidth is 1Gbps and connected to a router. Because it is initially evaluated that this bandwidth is sufficient to work, given assumed permissible unavailable time length we can estimate the temporary buffer size necessary to this interface when failures occurred. For example, if the time length is assumed to be 300ms the temporary buffer size needs about 37.5Mbyte with ignoring preamble and IFG(Inter Frame Gap) in Ethernet frame. And all other information needed is dynamically learned. So, just configuring one command, such as "vrrpfc 300", automatically sets aside this temporary buffer in this L2 switch.

Besides we can indicate that our method enables any L2 switches to disable the interface connecting to a master and make a backup change its role depending on such a threshold as latency, or reliability on the interface; just sending an advertisement frame with priority zero works after an L2 switch disables the interface. Since tracking ability has added to original VRRP protocol specification originally by some vendors any routers running the "extended-VRRP" could do the same thing. However, since in general a router has a lot of interfaces it is considered better for an L2 switch connecting to the router to track its interfaces and judge their status from the viewpoint of load distribution (Fig.3).

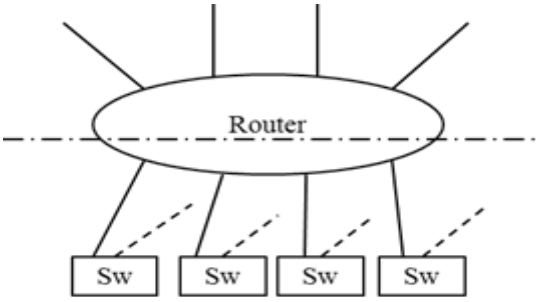

Fig 3: Load Distribution among L2 Switches

5. CONCLUSIONS

In this paper we have proposed a new method which makes VRRP convergence time shorter in a typical network model. Our method adds a new feature to a L2 switch deployed between two routers using VRRP. Our method enables a VRRP backup router to replace a master router immediately without waiting the protocol-defined time. And our method utilizes VRRP mechanism itself without any modification and imposes no additional feature which investigates data included in any Ethernet frame upon L2 switches. Furthermore our method requires no additional line between the two routers to convey their statuses each other. Therefore, consider all the various factors together, implementing our method on L2 switches would be one of the cost effective approaches to realize fast convergence. Our future work includes estimations of necessary temporary buffer sizes on several conditions.

6. REFERENCES

[1] J.Ranta," Router Redundancy and Scalability Using Clustering", Seminar on Internetworking, 2004

[2] N.H.Bhagat," Virtual Router Redundancy Protocol-A Best Open Standard Protocol in Maintaining Redundancy", International Conference on Web Services Computing (ICWSC), 2011

[3] W.Wu,K.Wang,Rong.Jan,and C.Huang," A Fast Failure Detection and Failover Scheme for SIP High Availability Networks", Dependable Computing, 2007. PRDC 2007. 13th Pacific Rim International Symposium, 2007

[4] T.Li, B.Cole, P.Morton and D.Li," Cisco Hot Standby Router Protocol (HSRP)", RFC2281, March 1998

[5] S. Nadas," Virtual Router Redundancy Protocol (VRRP) Version 3 for IPv4 and IPv6", RFC5798, March 2010.

[6] David C. Plummer," An Ethernet Address Resolution Protocol -- or --Converting Network Protocol Addresses to 48.bit Ethernet Address for Transmission on Ethernet Hardwar", RFC826, November 1982

[7] IEEE standard 802.1AX-2008, November 2008

[8] IEEE standard 802.1D-2004, section 17, June 2004

[9] M. Christensen, K. Kimball, and F. Solensky," Considerations for Internet Group Management Protocol (IGMP) and Multicast Listener Discovery (MLD) Snooping Switches", RFC4541, May 2006

[10] Cisco Systems,"Cisco Data Center Interconnect Design and Implementation Guide", http://www.ciscosystems.biz/en/US/solutions/collateral/ns340/ns517/ns224/ns949/ns304/ns975/data_center_interconnect_design_guide.pdf, 2009