Router Redundancy and Scalability Using Clustering

Juha Ranta
Helsinki University of Technology
Telecommunications Software and Multimedia Laboratory
jmranta2@cc.hut.fi

Abstract

Nowadays the availability of networked services is very important for many businesses and it is extremely important that overload or failure of one network component does not prevent the normal usage of all other services. This paper focuses on a clustering technique, Virtual Router Redundancy Protocol, that can be used for deploying resilient and scalable first-hop router systems in local area networks.

KEYWORDS: VRRP, Virtual Router Redundancy Protocol, Router redundancy, High Availability

1 Introduction

Availability of networked services is very important for many different organizations and companies nowadays. Many business and business processes rely heavily on network applications and services provided over Internet Protocol (IP) networks. The global reach of Internet also means that services need to be available 24 hours a day to serve employees, customers and business partners around the world. As a result of this a single network failure can be very costly for businesses.

In order to maintain desired level of quality of service and to minimize the cost of network device failures or service breaks there needs to be some redundant network device or devices that can take over the functions of the primary network device if the primary device should fail.

The gateway routers that interconnect organizations own Local Area Networks (LAN) and the global Internet are probably the most critical single points of failure because they are the only path between these different network segments. Even if there were multiple routes between network segments the individual hosts in that network segment are most probably totally unaware of these different routes and cannot reach different networks if the first-hop router is lost. If one critical route or path becomes unavailable then in the worst case it can stop all other services from functioning if the network topology is badly designed.

This paper will focus on router redundancy and mechanisms that can be used to provide fast failover and scalability within LANs and what kind of technical requirements they set to underlying physical network layer and to actual hosts in that LAN segments. First in Section 2 different techniques that can be used to select the first-hop router are discussed and then in Section 3 a technique that can be used to provide router redundancy is introduced.

2 Discovering the first-hop router

In this section different methods end-host computers in LAN segment can use to discover and select first-hop routers towards the desired destination are briefly discussed. Provided that there are more than one possible first-hop router available several problems (listed below) become evident.

- How to select the first-hop router
- How to detect router failure
- How to switch to new first-hop router when failure is detected

One possibility is to run some dynamic routing protocol software such as Routing Information Protocol (RIP) [9] or Open Shortest Path First (OSPF) [10] in end-host computers and this way they could discover and first-hop router by listening routing messages sent by the actual routers. In case of a router failure they could switch route based on the dynamic routing protocol information available. This approach is impractical for many reasons. First of all there might not be required routing software available for all different platforms. Even if required software would be available running and configuring sophisticated routing software on each host, even on desktop computers, would be error prone and time consuming job as configurations would need to be check for every installation and would most probably introduce new more serious problems as security and software compatibility issues. This would also increase management task since now all hosts that are running routing software should be monitored to make sure that routing software itself is up and running and the hosts would recover from first-hop router failure. If routing information for one reason or another would not be available anymore (eg. change of routing protocols from RIP to OSPF) each of end hosts should be updated and reconfigured. Running dynamic routing protocol software could solve the problem involved with discovering alternative routes in case of failures but as this operation could take something from tens of seconds to a few minutes this alternative can be discarded as too slow, complex and impractical [9, 10].

Another option would be to use Router Discovery Protocol (RDP) and run ICMP router discovery client [2]. In this approach each end-host computer must act actively and listen `RDP advertisement` multicast messages to detect failure of the first-hop router and to find another router if the default route should become unreachable. One should also note that RDP protocol's default advertisement rate (7 to 10 minutes) and advertisement lifetime (21 - 30 minutes) are too long for quick error detection and recovery. Even though these timer values can be set lower RDP is still not best possible solution for this problem since it requires active participation of all hosts and setting advertisement timer values lower would result in more protocol overhead in the network [2].

In most common case end-host network configurations are static. The configurations itself can be acquired either using Dynamic Host Configuration Protocol (DHCP) [3] or using local configuration files. Although DHCP helps with the selection of default route it does not provide any methods that end-host could use for selecting alternative route if the default route should fail. This means that fault tolerance must be provided by the router system itself. Virtual Router Redundancy Protocol (VRRP) is one solution to this problem and it can be deployed without having do any changes to end-host computers and their configurations. Consequently this is the most practical approach since static configurations are supported by most of the TCP/IP implementations and no additional software needs to be run in the end-hosts.

3 Virtual Router Redundancy Protocol

The Virtual Router Redundancy Protocol (VRRP) is an Internet Engineering Task Force (IETF) -standard solution designed to maintain High Availability (HA) of a LAN network without having to run additional software on every end-host. This is achieved by protecting the default gateway function with group of redundant routers where one router acts as the primary router and others as backup routers. If the primary router should fail, a backup router takes over the primary routers function. VRRP configuration "*is designed to eliminate the single point of failure inherent in the static default routed environment*" [7]. Thus VRRP can be used to assure that established paths (route through default gateway) stay open, but VRRP cannot be used for discovering and rediscovering paths like dynamic routing protocols like RIP and OSPF can.

A virtual router consists of two or more "real" routers that are running VRRP. Virtual Router Identifier (VRID) is used for identifying the group that form a virtual router. As long as the primary router is fully functional all network traffic destined to the external networks is handled by the primary router. But when primary router fails one of the backup routers (the one with the highest VRRP priority within this virtual router) takes over the master function and starts forwarding packets as if nothing has happened. VRRP operations are discussed in more details in Section 4.2.

3.1 VRRP scenarios

Sections 3.1.1 and 3.1.2 demonstrate how VRRP can be utilized to assure that static default first-hop routes stay open.

3.1.1 Single default route

Figure 1 illustrates a simple VRRP configuration where a single default route is used within LAN segment. Routers `R1` and `R2` form a virtual router (`VRID1`) where router `R2` is used to backup master router `R1` [7].

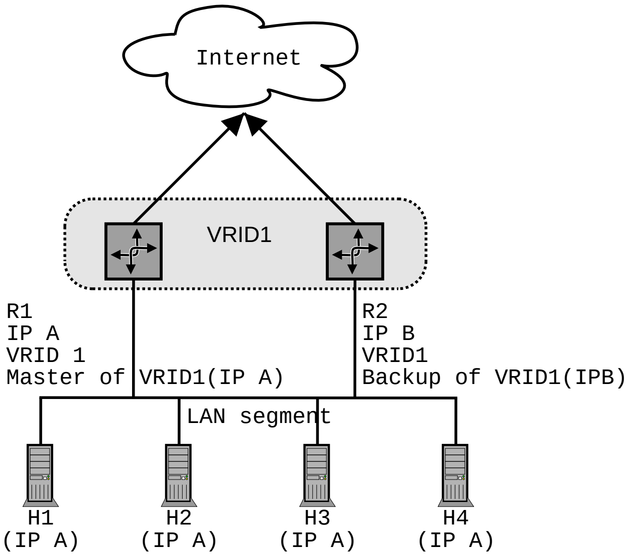

Figure 1: VRRP configuration with single default route

In this setup all network traffic destined to external network is forwarded by primary router `R1`. If router `R1` should fail, then the backup router `R2` of virtual router (`VRID1`) would take over the master router functionality and start routing IP packets normally forwarded by router `R1`. This means that router `R2` would use virtual router's `VRID1` virtual MAC and virtual IP address. End-hosts in LAN segment are totally unaware of this change.

This simple configuration does provide redundancy but it does not add any routing capacity. Router `R2` is completely idle while it is in the backup state and for that reason this is usually not the be desired configuration as valuable resources are not in use.

3.1.2 Multiple default routes

In figure 2 a more realistic and practical VRRP setup is shown. In this setup two routers `R1` and `R2` are used to form two virtual routers (`VRID1, VRID2`). Now `R1` acts as master of virtual router `VRID1` and `R2` as backup. In `VRID2` router `R2` is the master and `R1` acts as the backup.

In this scenario end-hosts `H1` and `H2` in the LAN segment are configured with `IP(A)` as default route and `H3` and `H4` use `IP(B)` as default route. VRRP configuration like this neatly adds scalability and more capacity to routing network functions in LAN segment. This way the valuable resource are in use all the time and it is also more likely that if router `R1` or `R2` becomes unavailable the backup router will work correctly and it's configurations are up to date as it has been active all the time.

One should also note that although VRRP setups with multiple default routes (and backup) will actually make rout-

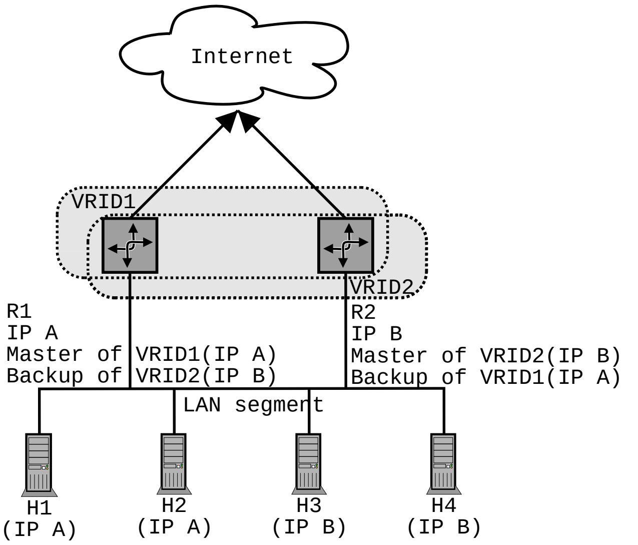

Figure 2: VRRP configuration with multiple default routes

ing resilient and scalable it does not provide any dynamic load sharing metrics.

4 VRRP technical details

VRRP is a IP router failover mechanism and it is based on a specially formed virtual MAC addresses and unique IP address that are shared among routers belonging to same virtual router. Each router within one virtual router would use the same virtual MAC address and IP address when acting actively as a master router. VRRP can be run over variety of LAN media types including Ethernet, FDDI and Token Ring but as VRRP uses addresses both from datalink and network layer it cannot be totally independent of the LAN media type used. Diffences between VRRP and LAN media types are dissussed in Section 5 and Ethernet is used here as an example to clarify the need for both of these addresses.

When VRRP is run over Ethernet media, routers use a virtual MAC address from special MAC address-space. VRRP virtual MAC address on Ethernet always has the format `00:00:5e:00:01:XX` where the last octet (`XX`) reflects the VRID value of that virtual router. It is import to use these virtual MAC addresses to make VRRP work in bridged networks. If virtual MAC addresses were not used learning bridges would learn the bridge port behind which the current master router resides as the end hosts would be using original master router's real MAC address. If the master router should now fail and the backup router would reside behind another bridge port it would never get the packets since they would be forwarded to the port where original master resided as the end hosts would be using old master routers real MAC address. But when virtual MAC addresses are used and the virtual routers MAC address changes with the virtual router IP address to the new master router, learning bridges will notice a station move when a backup router takes over the master router functionality and learning bridge's forwarding tables would be updated to contain new (`MAC, port`) mapping.

For redundancy schema to function properly the following issues have to be considered:

- How routers communicate
- How failover and new master election is performed

VRRP routers use VRRP messages for communication and these messages have important role in all VRRP operations.

4.1 VRRP messages

All VRRP protocol messaging is done using these IP multicast messages. This implies that VRRP can be used on different LAN techniques supporting IP multicasts. VRRP uses a known IPv4 link local multicast address (`244.0.0.18`), specially assigned to this purpose by Internet Assigned Numbers Authority (IANA), for communication between VRRP routers in that (extended) LAN and actual VRRP packets are encapsulated in IP packets [7].

These messages are used for exchancing VRRP priority, VRRP state and VRID information. One should note that only the master of the virtual router is sending out these messages and backup routers are passively listening.

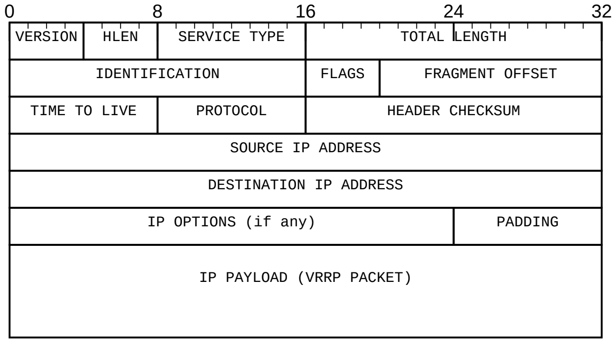

Figure 3: IP packet header

There are four fields in IP header [4] (shown in figure 3) that are important for VRRP. These fieds are:

- Time To Live (TTL)
- Protocol type
- Source IP Address
- Destination IP Address

TTL field of IP Packet is always 255. Since VRRP is intended to be run in within single LAN segment this value can be used to make sure that this IP packet originated from this LAN segment. Any other value for TTL in IP packet containing VRRP data must be dropped immediately. This feature prevents VRRP packets being injected from remote networks.

Protocol type for IP packets containing VRRP packets is 112. This value has been assigned by IANA.

Source IP address of IP header containing VRRP packet is the real IP address of interface from which this packet was sent.

Destination IP address of IP header is the multicast address assigned by IANA (`244.0.0.18`). This multicast address resides in `224.0.0/24` network which is local multicast network and this means that routers are not required to forward these packet eventhough packet TTL has not exceeded [1].

4.1.1 ADVERTISEMENT

For current VRRP protocol version (2) only one packet type is defined. VRRP advertisement message is shown in the figure 4 and it has the following fields [7].

- Version
- Type
- VRID
- Priority
- Count of IP addresses
- Authentication type
- Advertisement interval
- Checksum
- IP addresses
- Authentication data

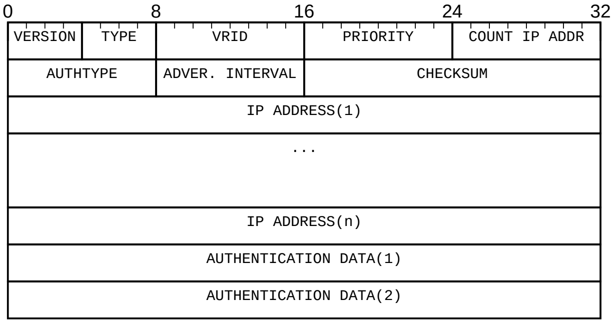

Figure 4: VRRP ADVERTISEMENT

Version number for VRRP is currently 2 [7].

Type field's value for VRRP advertisement is 1. As there are no other VRRP message types defined for VRRP packets containing any other value in this field should be discarded [7].

VRID field contains a integer value (1-255) and it is used to identify the virtual router. [7]

Priority field value (0-255) is used when election master router. Values 0 and 255 have special semantics. The value 0 is used by the current master router to indicate that it is giving up its master status. The value 255 is always the priority of router that owns the virtual router IP. A router with highest priority will always become master a virtual router. Thus owner of the IP address will always be virtual router master if it is functioning properly [7].

Count of IP addresses is the number of **IP Addresses** included in this VRRP packet.

Authentication type is the authentication method used by the virtual router specified by `VRID` field.

- 1. no authentication
  Authentication is not used by the virtual router.
- 2. simple text password
  Virtual router uses the preconfigured clear-text password. VRRP packet is dropped if password does not match with the one configured for this virtual router.
- 3. IP authentication header
  IP Authentication Header is used. More information about `IP Authentication Header` can be found in RFC 2402 [6].

Advertisement interval field is used to define the time interval in seconds that virtual router is using to send `ADVERTISEMENT` messages. Field's value is an integer from range (1-255) but to make fast failover work default value 1 is usually used.

Checksum is computed from the whole VRRP packet and is used to detect transmission errors. Packets with invalid checksum are to be dropped immediately.

IP Addresses are used to inform what IP addresses are handled and protected by this virtual router and these values should be the same for master and all backup routers.

Authentication data is used in case of `simple text password` authentication schema these field contain a password of fixed length (8 octets). In other case octets of this field is set to zero.

4.2 State matchine variables

VRRP virtual router state machine uses few more parameters in addition to those sent on VRRP packets.

Skew time is factor used to create some differences to backup router timers. This is used in a situation when there are multiple backup routers and master router has been lost and the new master router election starts. Now backup routers with high priority start participating in master router election processes faster than those with a lower priority value. Skew time can be calculated with the formula 1.

$$Skew\_Time = ((256 - Priority)/256) \tag{1}$$

Master down interval is a time interval in seconds used to detect master router failure. If no new `ADVERTISEMENT` messages are received within this time interval between two messages, master router is considered to be lost and master router election process starts. The value for this variable is calculated using the formula 2 (where `AI` stands for `Advertisement_Interval`).

$$Master\_Down\_Interval = (3 * AI) + Skew\_Time \tag{2}$$

Now it can be seen that lower values of `Priority` will result in higher values of `Skew_Time` and longer `Master_Down_Interval`s. Thus the backup routers with low priority have the longest delay before they start participating in new master router election process [7].

Preempt mode is a boolean flag indicating whether if backup routers with higher `Priority` than the current master router replace the current master or not. It should be also noted that owner of the virtual router IP address (`R1` in figure 1) will always preempt the current master even if the value of this variable is `false`.

4.3 State machine transitions

There are only 3 different state in VRRP state machine and all transitions from any state to any other state are possible. These states are illustrated in figure 5. When VRRP virtual router is started, the owner of the IP address will become master and starts sending `ADVERTISEMENT` messages. Other routers will go into backup state and they passively listen for these messages sent by master.

If for some reason or another `ADVERTISEMENT`s are not received by the backup routers within `Master_Down_Interval` then backup routers will start the master router election process. When the new master router is elected it will forward the packets destined to this virtual router until new master is elected. This is master router election process describe in more details in section 4.4.

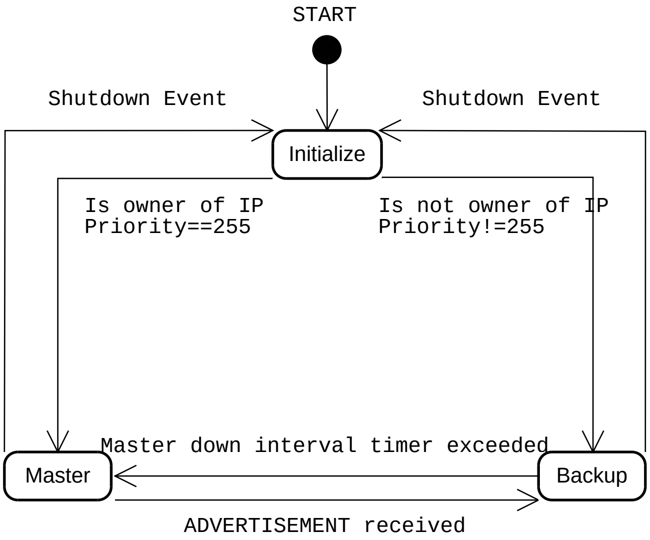

Figure 5: VRRP state machine

4.3.1 Operations in Master state

When router is operating as the virtual router master it is responsible for the forwarding all network traffic destined virtual router and it will perform some additional VRRP tasks [7].

- responds to ARP requests to for the IP addresses protocted by this virtual router

Case	State transition
$Prio_m = 0$	master
$Prio_m < Prio_r$	master
$Prio_m > Prio_r$	backup
$Prio_m = Prio_r$	backup (if $IP_m > IP_r$)
$Prio_m = Prio_r$	master (if $IP_m < IP_r$)

Table 1: VRRP master ADVERTISEMENT transtions

- forwards packets having virtual routers virtual MAC address as destination address
- accepts packets addressed to the IP addresses protected by this virtual router if the router is the real IP address owner.

In addition to these default gateway network funtions master router will perform VRRP specific tasks. Master router will periodically send new `ADVERTISEMENT` packets to inform backup routers that it is still up and running. This event is triggered by `Advertisement_Interval` timer.

When `ADVERTISEMENT` messages are sent the master router will set it's own configured VRRP parameters like `Priority` and `Advertisement_Interval` to the packet and calculate the `Checksum` of the VRRP packet when other fields are set. Also the IP packet headers, in which the VRRP packet is enscapsulated, are set like described in Section 4.1.

If a VRRP router (r) in master state receives an `ADVERTISEMENT` message (m) with priority higher than the local configured priority the router will transition into the backup state and set it's `Master_Down_Timer`. Transitions are summarized in table 1.

4.3.2 Operations in Backup state

VRRP router operation in backup state is listening to `ADVERTISEMENT` messages sent by the current master router. VRRP router in backup state will not send any `ADVERTISEMENT` messages unless really wants to become current master (router is the virtual router IP address owner). During this time backup routers must act in the following manner [7].

- discard all ARP requests for IP addresses that are protected by this virtual router
- discard packets with virtual routers MAC address
- discard IP packets with IP addresses associated with the virtual router IPs.

Like master router backup router have VRRP specific tasks they need to perform. Backup routers must receive `ADVERTISEMENT` messages sent by the current master and set `Master_Down_Timer` accordingly.

When `ADVERTISEMENT` message is received, backup routers will check that IP packet and VRRP message headers are valid. The following conditions must be true or otherwise the packet is discarded.

- IP packet TTL is 255.

- IP packet Protocol is 112.
- VRRP version is 2.
- VRRP checksum must be correct.
- VRID matches to the virtual router identifier that is being protocted by backup router.
- VRRP advertisement interval is same as configured interval.
- If VRRP authentication is used, additional authentication information must be valid.

When a `ADVERTISEMENT` message with `Priority` 0 is received in backup mode, current master is giving up it's master status, backups routers set their `Master_Down_Interval` timer to `Skew_Time` since they now know that there is no master router anymore but the different priorities of separate backups are to be respected. If message's priority is greater than or equal to backup routers then the `Master_Down_Interval` timer is set in normal manner (formula 2). But if message's $Priority_m$ is lower than backup routers $Priority_b$ and backup router `Preempt_Mode` is set to true then backup router will discard this messages. This will soon result in the fact that backup router's `Master_Down_Interval` timer is triggered and new master election process is initiated.

4.4 Master router election process

When backup router's `Master_Down_Interval` timer is triggered the router transitions into master state. One should note, that it is possible that multiple backup routers transition from backup state to master state almost simultaneously. In this state each backup router thinks that it is the new master router and starts sending `ADVERTISEMENT` messages but this will soon converge and the active master router with highest `Priority` and IP address values will be elected as the master and other "wannabe" master router will quicky transition back to the backup state.

5 VRRP and different LAN media types

As mentioned earlier VRRP can operate on many different LAN technologies including Ethernet, FDDI, Token-Ring. Since there are significant implementation specific differences between these separate network media types.

5.1 Ethernet

VRRP functions well over Ethernet and use a predictable MAC address `00:00:5e:00:01:$VRID`. This means that there can be 255 different VRRP virtual routers in single LAN network. Thus virtual MAC addresses of virtual routers must be different must be unique within one LAN. If LAN segment is built up from multiple bridged Ethernet segments and each of the backup routers is behind different port of the bridge, learning bridges will notice a "station move" when new router take over the master router functionality and start sending `ADVERTISEMENT` messages. One should also note that if multiple learning bridges are used to connect different Ethernet segments and each VRRP router is behind different port of these learning bridges the convergence of learning bridges spanning tree algorithm may take noticably longer than VRRP master election.[7].

5.2 FDDI

Fiber Distributed Data Interface (FDDI) [7] is different from Ethernet in many ways. In FDDI messages are passed around the ring through all interfaces connected to it and when messages MAC address matches to the interface's MAC address then the message is removed from the ring and this makes implementing VRRP in FDDI little more complicated.

This problem can be avoided by configuring a unicast MAC filter to FDDI interfaces of virtual router members. This way all master and all backup routers will see required messages [7]. Unfortunately MAC address filtering might not be supported by all FDDI implementations and as a side effect of this filtering FDDI interfaces will no longer remove these packets automatically from FDDI ring.

If MAC address is filtering is not supported in the FDDI interface implementation, then only way is to use real MAC addresses as virtual router MAC addresses and periodically send some messages through learning bridges to make sure that their (`MAC`, `port`) mapping tables stay up to date.

5.3 Token ring

Operating VRRP over Token Ring is quite problematic. There are no general multicast mechanism that would work both over old and new implementations of Token Ring. This means that one of the Token Ring's functional addresses must be used to attain compatibility with all Token Ring implementations and of 31 total functional address only 12 are user-definable [7].

6 VRRP security issues

When mechanisms for redundancy in LAN are deployed some security issues needs to be considered as when the networks have higher availability they most likely are more vulnerable to malicious attacks. VRRP has some built-in security features that can protect VRRP from these malicious attacks.

VRRP messages must always have IP packet TTL value of 255. VRRP message is discarded if any other value for IP TTL field is detected. This prevents VRRP messages from being injected from remote networks. Other obvious TTL value, for IP packets carrying VRRP messages, would have been 1. This way TTL would immediately drop down to 0 and IP packet would not be forwarded any longer [4]. This would not have been a good choice since a malicious person could then inject VRRP messages from remote network. Also the IP multicast address being used in VRRP is a link local address and most of the routers do not forward these packets to other networks. These countermeasures only work against remote attacks and local attacks are still possible.

Also some authentication methods are provided by VRRP but only the `IP Authentication Header` method can really provide some real security againts evil-intentions as it can be used for data integrity and IP packet origin authentication. `Simple text password` authentication does not provide any real security as this clear text password can easily be snooped. This method still can be used to prevent accidental configuration mistakes, like assigning same VRID for multiple virtual routers in same LAN [7].

7 Alternative techniques

There are also other similar techniques for implementing redundancy and fast failover in LAN networks. Cisco systems has develop a propietary protocol called Hot Standby Router Protocol (HSRP) and IP Standby Protocol (IPSTB) was developed for Digital Equipment Corporation's DECNET routers [8, 5].

7.1 HSRP

HSRP was designed to assure that routing functions is available even if one router should fail. In HSRP there are also multiple routers working as one virtual router. There can also be multiple HSRP virtual routers in the same LAN segement. HSRP is little more complicated as a protocol than VRRP is. There are six different operational states in HSRP protocol and three different message types are used for protocol signaling. HSRP messages are sent over UDP port 1985 and in contrast to VRRP all of the HSRP virtual router groups member are always sending out these messages but the failover mechanism itself is also based on use of special MAC addresses [8]

7.2 IPSTB

IPSTB was designed to provide fast failover mechanisms for IP routers. In IPSTB protocol IP/UDP datagrams are used for signaling and failover mechanism is also based on special MAC address where each member of virtual router group uses it's own IP address but answers to ARP requests with with one of the router groups IP addresses. One big difference to VRRP is that there can only be one IPSTB virtual router in one LAN segment. [8]

8 Conclusion

Availability of networked service can be highly increased if redundant network devices that can perform the same function are introduced. VRRP provides a elegant solution for avoiding the single point of failure in LAN segments default gateway. When redundancy schemas like VRRP are deployed it does not only saves from catastrophes that can occur if the only route between LAN segment and rest of the Internet goes down but can also prevent such catastrophes from happening as individual routers can be brought down for maintence whenever needed.

But as always fine pieces of equipment and advanced software cannot guarantee anything unless the people managing, monitoring and maintaining these equipments are qualified and the processes used are properly organized and planned.

References

[1] Z. Albanna, K. Almeroth, D. Meyer, and M. Schipper. Iana guidelines for ipv4 multicast address assignments. RFC 3171, Internet Engineering Task Force, August 2001.

[2] S. Deering. Icmp router discovery messages. RFC 1256, Internet Engineering Task Force, September 1991.

[3] R. Droms. Dynamic host configuration protocol. RFC 2131, Internet Engineering Task Force, March 1997.

[4] J. P. (editor). Internet protocol - darpa internet program protocol specification. RFC 791, Internet Engineering Task Force, September 1981.

[5] P. Higginson and M. Shand. Development of router clusters to provide fast failover in ip networks. *Digital Technical Journal*, 9(3):32–41, 1997.

[6] S. Kent and R. Atkinson. Ip authentication header. RFC 2404, Internet Engineering Task Force, November 1998.

[7] S. Knight, D. Weaver, D. Whipple, R. Hinden, D. Mitzel, P. Hunt, P. Higginson, M. Shand, and A. Lindem. Virtual router redundancy protocol. RFC 2338, Internet Engineering Task Force, April 1998.

[8] T. Li, B. Cole, P. Morton, and D. Li. Cisco hot standby router protocol (hsrp). RFC 2281, Internet Engineering Task Force, March 1998.

[9] G. Malkin. Rip version 2. RFC 2353, Internet Engineering Task Force, November 1998.

[10] J. Moy. Ospf version 2. RFC 2328, Internet Engineering Task Force, April 1998.