Redundancy and load balancing at L3 in Local Area Networks

Fulvio Risso

Politecnico di Torino
Copyright notice

- This set of transparencies, hereinafter referred to as slides, is protected by copyright laws and provisions of International Treaties. The title and copyright regarding the slides (including, but not limited to, each and every image, photography, animation, video, audio, music and text) are property of the authors specified on page 1.

- The slides may be reproduced and used freely by research institutes, schools and Universities for non-profit, institutional purposes. In such cases, no authorization is requested.

- Any total or partial use or reproduction (including, but not limited to, reproduction on magnetic media, computer networks, and printed reproduction) is forbidden, unless explicitly authorized by the authors by means of written license.

- Information included in these slides is deemed as accurate at the date of publication. Such information is supplied for merely educational purposes and may not be used in designing systems, products, networks, etc. In any case, these slides are subject to changes without any previous notice. The authors do not assume any responsibility for the contents of these slides (including, but not limited to, accuracy, completeness, enforceability, updated-ness of information hereinafter provided).

- In any case, accordance with information hereinafter included must not be declared.

- In any case, this copyright notice must never be removed and must be reported even in partial uses.
Default gateway redundancy (1)

- **Objective**
  - Provides redundancy of the default gateway
  - Hosts on the LAN are not able to learn the network topology through Layer 3 routing protocols
Default gateway redundancy (2)

- Solutions
  - HSRP (Hot Standby Routing Protocol)
    - Cisco proprietary protocol defined in RFC 2281
  - VRRP (Virtual Router Redundancy Protocol)
    - Standard protocol defined in RFC 3768
  - GLBP (Gateway Load Balancing Protocol)
    - Cisco proprietary protocol
HSRP: overview

- Objectives
  - First, redundancy of the default gateway
  - Second, load balancing

- HSRP packets
  - Required for the “keepalive” function between routers
  - Encapsulated in UDP, port 1985
  - Transmitted to multicast address 224.0.0.2
  - TTL = 1 (packets are not further forwarded by routers)
HSRP: the idea

ARP Request: who has IP 2.2.2.4?
ARP Reply: IP 2.2.2.4 is at MAC 44-44-44-44-44-44

"Hello" packets

R1 (active)
IP: 2.2.2.1
MAC: 11-11-11-11-11-11
Virtual IP: 2.2.2.4
Virtual MAC: 44-44-44-44-44-44

R2 (stand-by)
IP: 2.2.2.2
MAC: 22-22-22-22-22-22

R3 (listen)
IP: 2.2.2.3
MAC: 33-33-33-33-33-33
HSRP: working operations (1)

- Routers that provide redundancy belong to the same “group”
  - Each group emulates a single virtual router
- Routers can be
  - Active: the one that has the right to serve the LAN
  - Stand-by: the one that takes place in case the Active fails
  - Listen: the others that are neither Active nor Stand-by
- Within each group, we define two additional (virtual) addresses
  - A common IP address
  - A common MAC address
HSRP: working operations (2)

- The Active router interface has the followings assigned addresses:
  - Primary IP address (inserted in the source IP header field)
  - Virtual IP address (used by hosts as default gateway)
  - Physical MAC address assigned by the NIC manufacturer
  - Well-Known MAC address allocated to the HRSP Group

- The Standby router interface has the followings assigned addresses:
  - Primary IP address (inserted in the source IP header field)
  - Physical MAC address assigned by the NIC manufacturer
HSRP: working operations (3)

- The Active Router
  - Is the only device allowed to use the virtual addresses
    - Only the Active router can reply to the virtual IP/MAC addresses
  - Do no longer use its IP/MAC addresses (unless an explicit request for that address is received)
- The stand-by router replies only to its own IP/MAC addresses
  - It will start using the virtual MAC when it is promoted to “active”
- The keep-alive message is used to define which router is actually in “active” mode
HSRP: Well-Known MAC address

- Act as a Virtual MAC Address
- Take in charge by the Active router and used to answer at the received ARP request packets
- Token Ring has only 3 possible groups (corresponding to HSRP Group 0, 1, 2)
  - C0-00-00-01-00-00
  - C0-00-00-02-00-00
  - C0-00-00-04-00-00
- Well known virtual MAC address for other LAN type (example 802.3, 802.11 etc.)
  - 00-00-0C-07-AC-xx
    - xx represents the HSRP Group
HSRP packets over the LAN

R1 (active)
- IP: 2.2.2.1
- MAC: 11-11-11-11-11
- V-IP: 2.2.2.3
- V-MAC: 00-00-0C-07-AC-01

H1
- DG: 2.2.2.3
- ETH.SRC: 00-00-0C-07-AC-01
- ETH.DST: 01-00-5E-00-00-02
- IP: 2.2.2.3
- IP.DST: 224.0.0.2
- UDP.SPORT: 1985
- UDP.DPORT: 1985
- HSRP HELLO

H2
- DG: 2.2.2.3
- ETH.SRC: 22-22-22-22-22-22
- ETH.DST: 01-00-5E-00-00-02
- IP: 2.2.2.3
- IP.DST: 224.0.0.2
- UDP.SPORT: 1985
- UDP.DPORT: 1985
- HSRP HELLO

H3
- DG: 2.2.2.3
- ETH.SRC: 00-00-0C-07-AC-01
- ETH.DST: 01-00-5E-00-00-02
- IP: 2.2.2.3
- IP.DST: 224.0.0.2
- UDP.SPORT: 1985
- UDP.DPORT: 1985
- HSRP HELLO

R2 (standby)
- IP: 2.2.2.2
- MAC: 22-22-22-22-22-22

Internet

Multicast MAC address (derived from 224.0.0.2)
HSRP: fault reaction

Data to Internet
IP.dst: IP_remote
MAC.dst: 33-33-33-33-33-33

"Hello" packets

R1 (active)
IP: 2.2.2.1
MAC: 11-11-11-11-11-11
Virtual IP: 2.2.2.3
Virtual MAC: 33-33-33-33-33-33

R2 (standby)
IP: 2.2.2.2
MAC: 22-22-22-22-22-22

R3
DG: 2.2.2.3

Internet
HSRP and switches: Filtering DB update

- A gratuitous ARP Reply (in broadcast) is sent when a router becomes Active
  - Source MAC: Virtual MAC address
  - Destination MAC: broadcast
  - ARP Reply: Virtual IP is at Virtual MAC

- Switches across the entire network will update their filtering database
HSRP and network utilization

- Only one exit link used
  - Unused bandwidth on the stand-by router
- Asymmetric routing
  - The external routing protocol can use both paths
HSRP for server/single router redundancy

- HSRP is a per-interface configuration
  - No matter if interfaces are on two devices or on just one
- Can be used to implement Dual homing servers/routers
  - Two standard NICs
  - HSRP will handle the virtual IP/MAC address on the primary interface
- Other option
  - Server (fault tolerant) NIC
    - Not available on routers
HSRP and VLANs

- HSRP provides redundancy of the default gateway
- VLANs require multiple default gateways (one per VLAN)
- → we need multiple HSRP groups
HSRP and load sharing (1)

R1
IP: 2.2.2.1
MAC: 11-11-11-11-11-11
Group1 (active)
Virtual IP: 2.2.2.3
Virtual MAC: 33-33-33-33-33-33
Group2 (stand-by)

Internet

H1
DG: 2.2.2.3

H2
DG: 2.2.2.3

H3
DG: 2.2.2.4

H4
DG: 2.2.2.4

R2
IP: 2.2.2.2
MAC: 22-22-22-22-22-22
Group1 (standby)
Group2 (active)
Virtual IP: 2.2.2.4
Virtual MAC: 44-44-44-44-44-44
HSRP and load sharing (2)

- Multiple HSRP groups
  - Multi-group HSRP (mHSRP)
  - One router active for the first group, the other active for the second group
  - Clients are configured half with the first DG, half with the second
- Note that load sharing in ingress does not depend on HSRP
  - Ingress traffic is handled by the routing protocol on the WAN side (e.g. BGP multi-homing)
- Is it really needed?
  - Often corporate networks have much more incoming than outgoing traffic
  - In this case, the egress bandwidth of a single link may be enough
HSRP and load sharing (3)

- Problems of the achieved load balancing
  - Load balancing is statically defined by physically partitioning hosts between two different default gateways
    - What about if one group generates much more traffic than the other?
  - Configuration burden
    - Not easy to differentiate the Default Gateway on the clients
    - Usually DHCP is used, and this usually returns a single DG for all hosts

- Asymmetric routing (again)
  - Ingress and egress paths may be different
L2 and load sharing

- We can achieve the same results of mHSRP
- However, no protection in case the router fails
HSRP packet format

- MAC header
- IP header
- UDP header
- HSRP
- FCS

- Version
- Op Code
- State
- Hello-time
- Hold-time
- Priority
- Group
- Reserved

- Authentication Data
- Authentication Data
- Virtual IP Address
HSRP header: “opcode” field

- Describes the type of message contained in this packet
- Possible values
  - 0 = Hello
    - The router is running and is capable of becoming the active or standby router
  - 1 = Coup
    - The router wants to become the active router
  - 2 = Resign
    - The router does no longer want to be the active router
    - “Coup” and “Resign” are not necessarily used together
      - E.g. a router that has highest priority can send the “coup” message to take over the current router, but this router goes into “speak” state without sending any “resign”
HSRP header: “state” field (1)

- Describes the current state of the router sending the message
- Possible values
  - 0 = Initial
    - This is the starting state and indicates that HSRP is not running
  - 1 = Learn
    - The router has not determined the virtual IP address and is still waiting to hear from the active router
  - 2 = Listen
    - The router knows the virtual IP address, but is neither the active router nor the standby router
HSRP header: “state” field (2)

- State field:
  - 4 = Speak
    - The router sends periodic Hello messages and is actively participating in the election of the active and/or standby router
  - 8 = Standby
    - The router is a candidate to become the next active router and sends periodic Hello messages
    - At most one router can be in Standby state (for each group)
  - 16 = Active
    - The router is currently forwarding packets that are sent to the group virtual MAC address
    - Each group has at least one (and only one) router in Active state
HSRP header: “Hello time” and “Hold time”

- **Hello-Time**
  - Period between the Hello messages sent by the routers
  - In case no active routers exists, a router uses its configured value (default: 3 seconds)

- **Hold-Time**
  - Validity of the current Hello message
  - When this timer expires, the Standby router proposes itself as Active router
  - In case no active routers exists, a router uses its configured value (default: 10 seconds)

- **Corollary**
  - The convergence time for an HSRP network is about 10 seconds
HSRP header: “priority” and “group” fields

- Priority
  - Used to force the election of the router with highest priority (higher number means higher priority)
  - In the case of routers with equal priority the router with the higher IP address wins
  - Default Priority is 100

- Group
  - Group ID the current HSRP instance is referring to
  - For Token Ring, values between 0 and 2 are valid
  - For other media, values between 0 and 255 are valid
    - Max 255 groups
HSRP header: Authentication and Virtual IP Address

- **Authentication Data**
  - This field contains a clear-text 8 character password
  - If no authentication data is configured the default text is “cisco”
  - Really simple authentication; mostly used to differentiate multiple instances of HSRP within the same LAN

- **Virtual IP address:**
  - Used by hosts as default gateway IP address
  - The virtual IP address used by this group
  - If the virtual IP address is not configured on a router, then it may be learned from the Hello message from the active router
HSRP: Preemption capability

- Parameter configured on each router
- If a router has higher priority than the active router and preemption is configured, it MAY take over as the active router using a Coup message
  - Without preemption, the currently Active router will stay active until it has a fault
  - A router configured with the highest priority cannot force the current Active router to resign, unless preemption is used
HSRP: basic configuration

R1(config)# interface ethernet 0
R1(config-if)# ip address 10.1.1.1 255.255.255.0
R1(config-if)# standby 24 ip 10.1.1.5
R1(config-if)# standby 24 priority 105
R1(config-if)# standby 24 preempt

R2(config)# interface ethernet 0
R2(config-if)# ip address 10.1.1.2 255.255.255.0
R2(config-if)# standby 24 ip 10.1.1.5
R2(config-if)# standby 24 preempt

HSRP is not a router-wide function; it is a propriety of the interface that must be on the LAN.
HSRP: “track” function (1)

- Problem: a failure on the WAN link does not trigger the Stand-by router to take place
  - Packets are sent to R1, from there to R2
  - No problems in connectivity, but additional overhead in forwarding
**HSRP: “track” function (2)**

- This function dynamically decrease the HSRP Priority when a tracked interface goes down
  - By default HSRP algorithm decrease the Priority by 10 when the link-layer of a tracked interface goes down
  - Be careful: only some types of faults cause the interface to go down
    - E.g. an interface will stay up if connected to an active L2 switch
HSRP: advanced configuration

R1(config)# interface ethernet 0
R1(config-if)# ip address 10.1.1.1 255.255.255.0
R1(config-if)# standby 1 ip 10.1.1.5
R1(config-if)# standby 1 priority 105
R1(config-if)# standby 1 preempt
R1(config-if)# standby 1 track Serial0
R1(config-if)# standby 2 ip 10.1.1.6
R1(config-if)# standby 2 preempt
R1(config-if)# standby 2 track Serial0

R2(config)# interface ethernet 0
R2(config-if)# ip address 10.1.1.2 255.255.255.0
R2(config-if)# standby 1 ip 10.1.1.5
R2(config-if)# standby 1 preempt
R2(config-if)# standby 1 track Serial0
R2(config-if)# standby 2 ip 10.1.1.6
R2(config-if)# standby 2 preempt
R2(config-if)# standby 2 track Serial0
VRRP overview (1)

- A “smart” clone of HSRP, with care taken not to infringe any Cisco patent
- Functioning, philosophy is exactly the same as HSRP
- Minor differences
  - Packet encapsulated in IP, protocol type 112 (no longer in UDP)
  - Transmitted to multicast address 224.0.0.18
  - Different MAC addresses associated to each group
  - TTL = 255
    - A VRRP router receiving a packet with the TTL not equal to 255 must discard the packet (only one possible hop)
- Active/Standby → Master/Backup
- Hello Messages → Advertisement messages
- HSRP Group → Virtual Router ID (VRID)
- Some timers (see later)
VRRP overview (2)

- Major (?) differences
  - Each master VRRP router can control more than one IP Address
  - A VRRP Router may backup one or more virtual routers
  - Any of the virtual router's IP addresses on a LAN can then be used as the Default Gateway by end-hosts
  - Support multiple logical IP subnets on a single LAN segment
  - For any VRID a single Master Router is elected the remaining routers are selected as Backup Routers (no longer routers in “Listen”)
  - Only the Master router sends Advertisement packets
  - The master router may have the same address as the group virtual router address
  - “tracking” not available
  - “preempt” is the specified behavior
VRRP overview (3)

- Default gateway redundancy is achieved in the same way as HSRP
- Load balancing is achieved the same way
VRRP: packet format (1)
VRRP: packet format (2)

- IP header
  - Source IP: real IP address of the interface the packet is being sent from
  - Destination IP: 224.0.0.18

- Type
  - The type field specifies the type of this VRRP packet. The only packet type is:
    - 1 Advertisement

- VRID
  - The Virtual Router Identifier (VRID) field identifies the virtual router this packet is reporting status for
    - Allowed values: 1-255
VRRP: packet format (3)

- **Priority**
  - Router with highest priority will become the master
    - In case of a tie, the router with the highest real IP becomes master
  - Priority = 255: assigned automatically to the router that has the same address as the virtual router
    - The router will be the master router (known as the “virtual address owner”)
  - Priority = 1-254: normal priority values
  - Priority = 0 → the current router does not participate in VRRP
    - Also advertised during an orderly shutdown of a master in order to speed-up Backup promotion (no need to timeout)
  - The default priority value for VRRP routers backing up a virtual router is 100
VRRP: packet format (4)

- Count IP Addrs
  - number of IP addresses contained in this VRRP Advertisement

- Authentication Type
  - Unused in the current version (RFC 3768)
  - Removed because operational experience showed that they did not provide any real security and would only cause multiple masters to be created
  - 0 = No Authentication
VRRP Timers

- **Advertisement Interval**
  - Time interval (in seconds) between Advertisements
  - default value = 1 s (HSRP was 3)

- **Skew_Time**
  - \((256 - \text{Priority}) / 256\) (seconds)

- **Master_Down_Interval**
  - \((3 \times \text{Advertisement\_Interval}) + \text{Skew\_time}\)
  - Time interval for Backup to declare Master down (seconds)
  - After that time, a new Master is elected
  - In case an orderly shutdown of a master is detected, backup waits only for the skew time
VRRP: Virtual MAC Address

- Well known virtual MAC address for any LAN except Token Ring (e.g. 802.3, 802.11 etc.)
  - 00-00-5E-00-01-xx
    - xx represents the VRID
  - OUI changed (C0-00-00 is owned by Cisco)
VRRP: Virtual MAC Address for Token Ring

<table>
<thead>
<tr>
<th>VRID</th>
<th>Token Ring Functional Address</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>03-00-02-00-00-00</td>
</tr>
<tr>
<td>2</td>
<td>03-00-04-00-00-00</td>
</tr>
<tr>
<td>3</td>
<td>03-00-08-00-00-00</td>
</tr>
<tr>
<td>4</td>
<td>03-00-10-00-00-00</td>
</tr>
<tr>
<td>5</td>
<td>03-00-20-00-00-00</td>
</tr>
<tr>
<td>6</td>
<td>03-00-40-00-00-00</td>
</tr>
<tr>
<td>7</td>
<td>03-00-80-00-00-00</td>
</tr>
<tr>
<td>8</td>
<td>03-00-00-01-00-00</td>
</tr>
<tr>
<td>9</td>
<td>03-00-00-02-00-00</td>
</tr>
<tr>
<td>10</td>
<td>03-00-00-04-00-00</td>
</tr>
<tr>
<td>11</td>
<td>03-00-00-08-00-00</td>
</tr>
</tbody>
</table>
HSRP/VRRP: convergence

- HSRP: about 10 sec with default parameters
  - Hold Time
  - User can configure Hello-Time and Hold-Time to improve this value

- VRRP: about 3 sec with default parameters
  - \((3 \times \text{Advertisement\_Interval}) + \text{Skew\_time}\)
  - User can configure Advertisement Interval and Priority
  - Less flexible than HSRP
HSRP/VRRP on real LANs: L2 resiliency (1)

Fault 1: OK (traffic re-routed through R2)
R1 (active)
IP: 2.2.2.1
Virtual IP: 2.2.2.3

Fault 2: OK (traffic re-routed through R2)

Fault 3: OK (traffic continues through R1)
R2 (standby)
IP: 2.2.2.2

Fault 4: also the “track” function is useless in this case

Fault 5: IP network partitioned in half (exit traffic may use R1, but incoming traffic may come through R2)

We must engineer the network with resiliency at data-link level (e.g. link aggregation between S1 and S2 or STP/RSTP)
HSRP/VRRP on real LANs: L2 resiliency (1)

- HSRP/VRRP do not protect from all faults on the L2 network
- Solutions: use STP/RSTP or link aggregation
  - The latter is the best for its reduced convergence time
  - STP may take 50s to converge; during this period, malfunctioning may occur
    - IP networks still partitioned
- HSRP/VRRP do not protect from some faults on the WAN link
HSRP/VRRP on real LANs: flooding (1)

Switches on this path will refresh the entries in the filtering database at each received frame (e.g. unicast).

Switches on this path will refresh the entries in the filtering database only when host H1 generates a broadcast frame.
In L2 networks, some switches may miss some entries in the filtering database:

- Usually, hosts talk with a limited number of other hosts
- Most of the time they send data to the default router
- In case host does not generate broadcast (e.g. UNIX or VmWare), large portions of the network may not know the MAC address of the host
HSRP/VRRP on real LANs: flooding (3)

- This situation may be extremely common with HSRP/VRRP
  - Egress and ingress routers may be different, hence the switch connected to the ingress router may not know the MAC address of the host
  - The only option (for the switch) is to send the frame on all its ports
  - ➔ flooding

- Solutions
  - Re-engineer the L2 spanning tree (not really a solution)
  - Force stations to send broadcast frames rather often (< max_age)
  - Increase max_age on the switches
HSRP/VRRP on unidirectional links

R1 (active)
IP: 2.2.2.1
MAC: 11-11-11-11-11-11
V-IP: 2.2.2.3
V-MAC: 00-00-0C-07-AC-01

R2 (standby \(\rightarrow\) active)
IP: 2.2.2.2
MAC: 22-22-22-22-22-22
\(\rightarrow\) V-IP: 2.2.2.3
\(\rightarrow\) V-MAC: 00-00-0C-07-AC-01

Instability in the filtering database of the switch
GLBP

- Enhancement (and replacement) of HSRP
- Cisco proprietary
  - Not even available on the entire product line
- Automatic load balancing across default gateways
  - Traffic is distributed across multiple routers
  - No configuration problems (such as in mHSRP) in assigning multiple default gateways to clients and creating multiple groups
- Same first-hop failure recovery capability of HSRP
- A group of routers provide a unique virtual router service
  - One IP address
  - Multiple virtual MAC addresses for forwarding
GBLP: the idea

ARP Request: who has IP 2.2.2.4?
ARP Reply: IP 2.2.2.4 is at MAC 44-44-44-44-44
ARP Request: who has IP 2.2.2.4?
ARP Reply: IP 2.2.2.4 is at MAC 55-55-55-55-55

R1 (AVG)
IP: 2.2.2.1
MAC: 11-11-11-11-11-11
V-IP: 2.2.2.4
V-MAC: 44-44-44-44-44-44

R2 (AVF)
IP: 2.2.2.2
MAC: 22-22-22-22-22-22
V-IP: 2.2.2.4
V-MAC: 55-55-55-55-55-55

R3 (AVF)
IP: 2.2.2.3
MAC: 33-33-33-33-33-33
V-IP: 2.2.2.4

H1
DG: 2.2.2.4

H2
DG: 2.2.2.4

H3
DG: 2.2.2.4

Internet
GLBP Functions

- **Active Virtual Gateway (AVG)**
  - GLBP members elect one router to be the AVG for the group
  - AVG replies to ARP requests for the virtual IP from clients
  - AVG assigns virtual MAC addresses to the active virtual forwarders

- **Active Virtual Forwarder (AVF)**
  - Each router in the VRRP group (up to 4 per group) routes packets forwarded to its assigned virtual MAC address
GLBP Operation

- An AVG router is elected within each GLBP group
- The AVG allocates a distinct virtual MAC address to each member (the AVFs)
- If a client ARPs the virtual IP address, the AVG responds with one of the virtual MAC addresses assigned to the AVFs
  - Clients now send their frames to one of the AVFs
- If an AVF fails, another AVF takes over forwarding for that AVF
Load Balancing Algorithm

- Four possibilities
  - None: GLBP operates like HSRP
  - Weighted: each GLBP router in the group will advertise its weighting and assignment; the AVG will act based on that value
    - Used in case the exit links have different capacities
  - Host dependent: this ensures that a host will be guaranteed to use the same virtual MAC address as long as the number of AVFs in the GLBP group is constant
    - Used when using stateful Network Address Translation because it requires each host to be returned the same virtual MAC address each time it sends an ARP request for the virtual IP address
  - Round robin: each AVF MAC address is used sequentially in ARP replies for the virtual IP address
Conclusions

- HSRP/VRRP widely used in practice
  - Simple and effective
- Often, a single group is used
  - Usually outgoing traffic much smaller than incoming traffic
- GLBP
  - Proprietary, not documented
- Take care of L2 issues
  - L2 resiliency
  - L2 flooding
  - Undetected faults