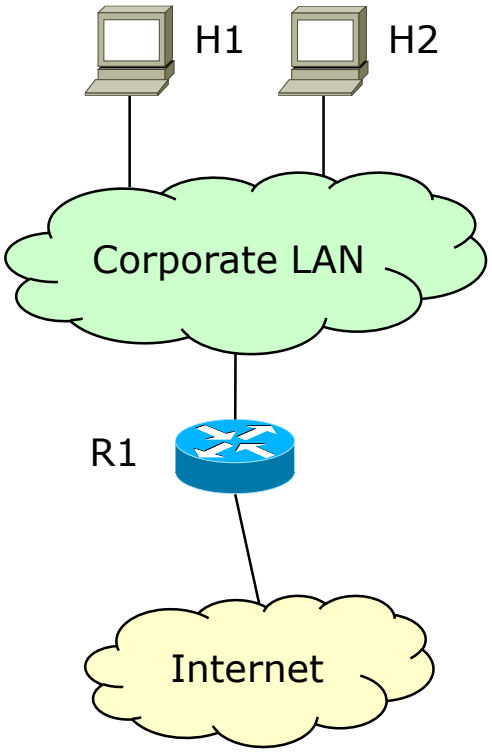
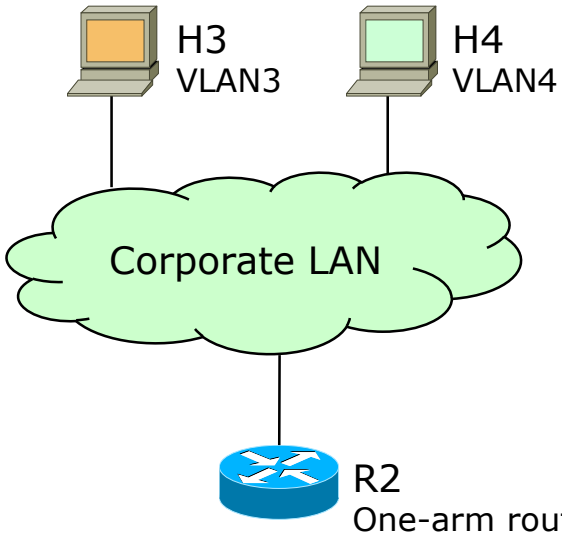
Default gateway redundancy and load balancing in Local Area Networks

Fulvio Risso

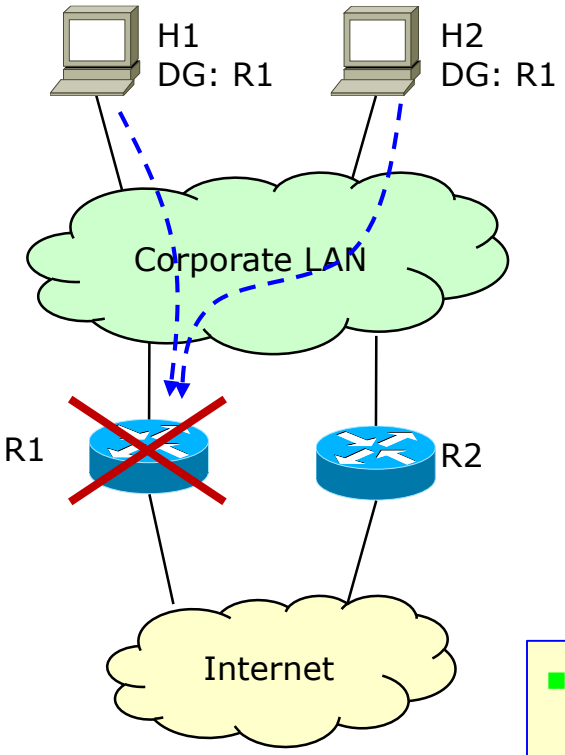
Politecnico di Torino

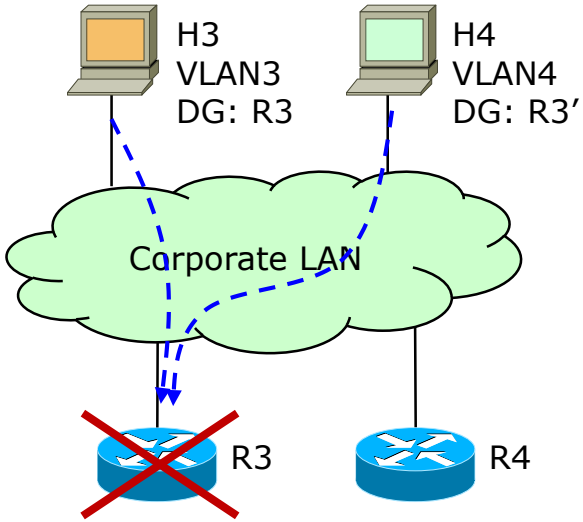
Problem: the router is a single point of failure





The wrong solution: router duplication





- Two routers on the LAN are useless if hosts are not able to switch to another one in case the first fails
- Hosts on the LAN are not able to learn the network topology through Layer 3 routing protocols

Possible solutions to default gateway redundancy

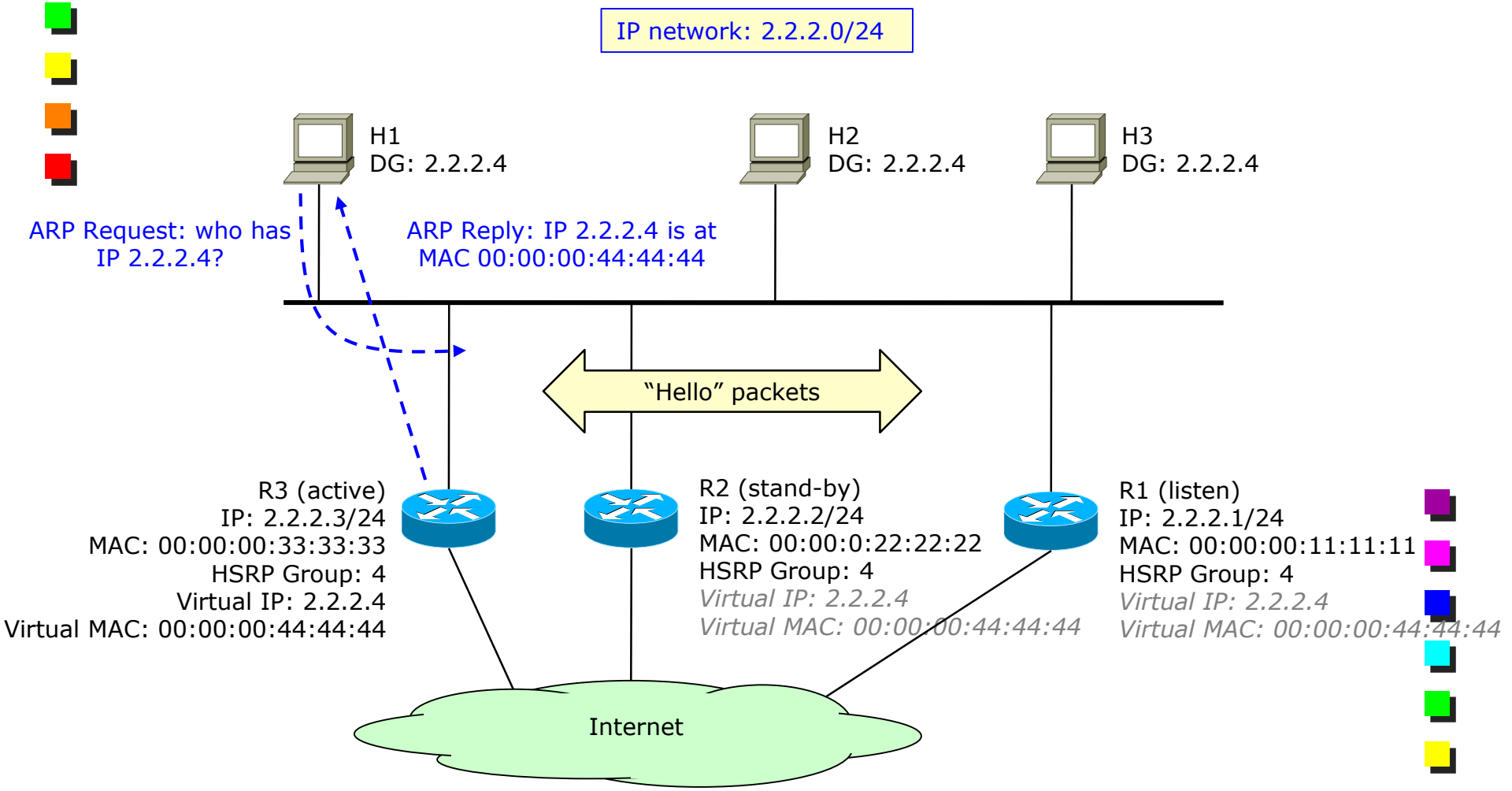
- HSRP (Hot Standby Routing Protocol)
  - Cisco proprietary protocol defined in RFC 2281
- VRRP (Virtual Router Redundancy Protocol)
  - Standard protocol defined in RFC 3768
- GLBP (Gateway Load Balancing Protocol)
  - Cisco proprietary protocol

HSRP: overview

- Objectives
  - First, redundancy of the default gateway
  - Second, load balancing
    - Not done very well

- Operations summary
  - Among several Default gateways, one elected as “master”
  - Very simple *keep-alive* messages (“Hello” packets) to:
    - Elect the “master”
    - Detect when the master fails and replace it with a new master

HSRP: the idea



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
    - May become stand-by in case the active fails
- Active Router
  - The one with the highest Priority (configurable via management)
  - Or (in case of equal priority) the router with highest IP address

HSRP: working operations (2)

- Within each group, we define two additional (shared) addresses
  - A virtual IP address (set by the network admin)
  - A virtual MAC address (automatically generated by the system)
- Only the Active Router can use the above virtual addresses
  - Only the Active router can serve packets coming to those virtual IP/MAC addresses
    - And, of course, any ARP request to the virtual IP
- Standby and Listen routers can serve only packets coming to their own IP/MAC addresses
  - The Standby will start using the virtual addresses only when it is promoted to “active”

HSRP: Hello Packets

- Periodic *Keep-alive* (“Hello”) messages
- Used for the following processes
  - Election process
    - Needed to define which is the Active router
      - Priority
      - (or, if equal) Higher IP address
  - Keep-alive process
    - Needed to keep track of possible failures of the default gateway
- Generated by Active and Stand-by routers
  - Router in Listen state do not generate Hello packets, unless they detect that either the Active or the Stand-by failed

HSRP addresses on the Active Router

- The interface of the Active router has the following assigned addresses:
  - Primary IP address (inserted in the source IP header field)
  - Physical MAC address assigned by the NIC manufacturer
  - Virtual IP address (used by hosts as default gateway)
    - Has to be set by the network admin during HSRP configuration
  - Virtual MAC address allocated to the HSRP Group
    - Well-known MAC address, derived by the HSRP group
    - Automatically generated by HSRP

HSRP addresses: Virtual MAC address

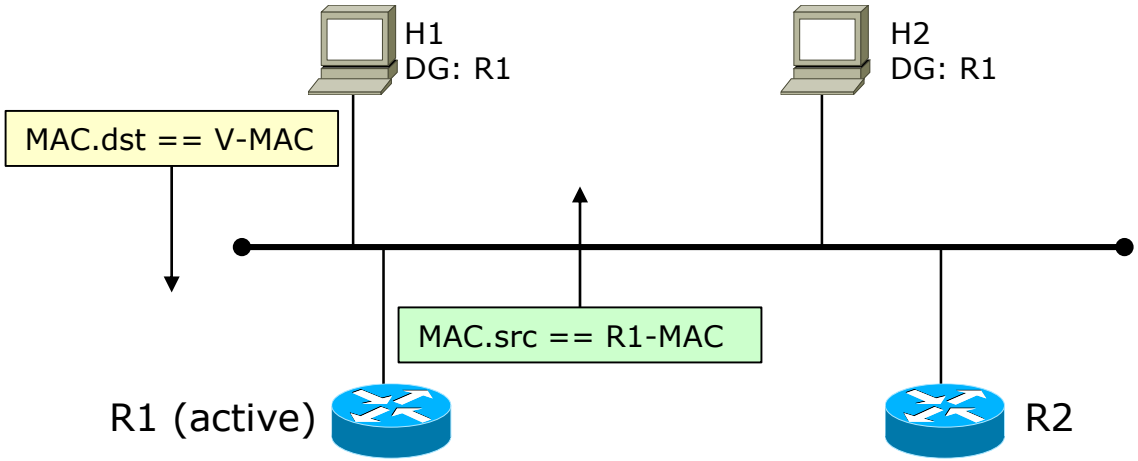
- Taken in charge by the Active router and used to answer to ARP requests
- Unicast address
- Token Ring has only 3 possible values (corresponding to HSRP Groups 0, 1, 2)
  - C0-00-00-01-00-00, C0-00-00-02-00-00, C0-00-00-04-00-00
- Well known virtual MAC address for other LANs (e.g., 802.3, 802.11 etc.)
  - 00-00-0C-07-AC-xx
  - xx represents the HSRP Group
  - 00-00-0C is a Cisco-assigned OUI

HSRP addresses on the Stand-by/Listen Routers

- The interfaces of the Standby and Listen routers have the followings assigned addresses:
  - Primary IP address (inserted in the source IP header field)
  - Physical MAC address assigned by the NIC manufacturer
  - Virtual addresses will become active when the router is promoted as “active”

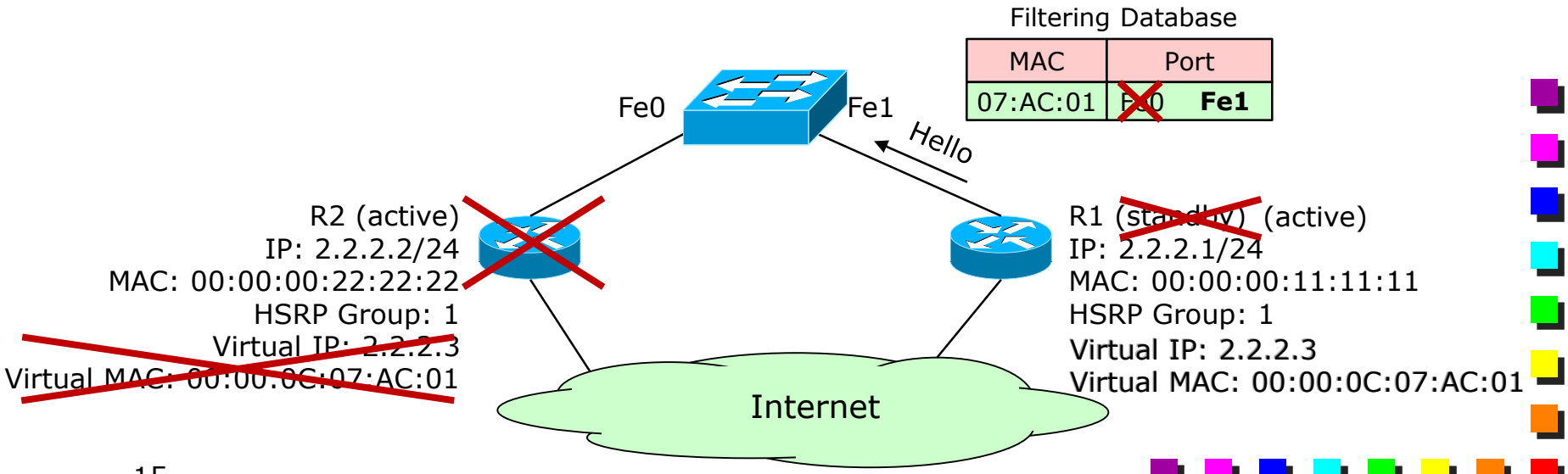
Using MAC addresses when forwarding traffic

- Hosts use the **Virtual MAC** address to deliver their traffic to their Default Gateway
- Routers (also the Active!) use the **Actual MAC** address to deliver the L3 traffic to hosts
  - The forwarding process of a router is not affected by HSRP
  - Hosts will experience *two different MAC addresses* when sending and receiving traffic toward the Default Gateway



HSRP on switched networks

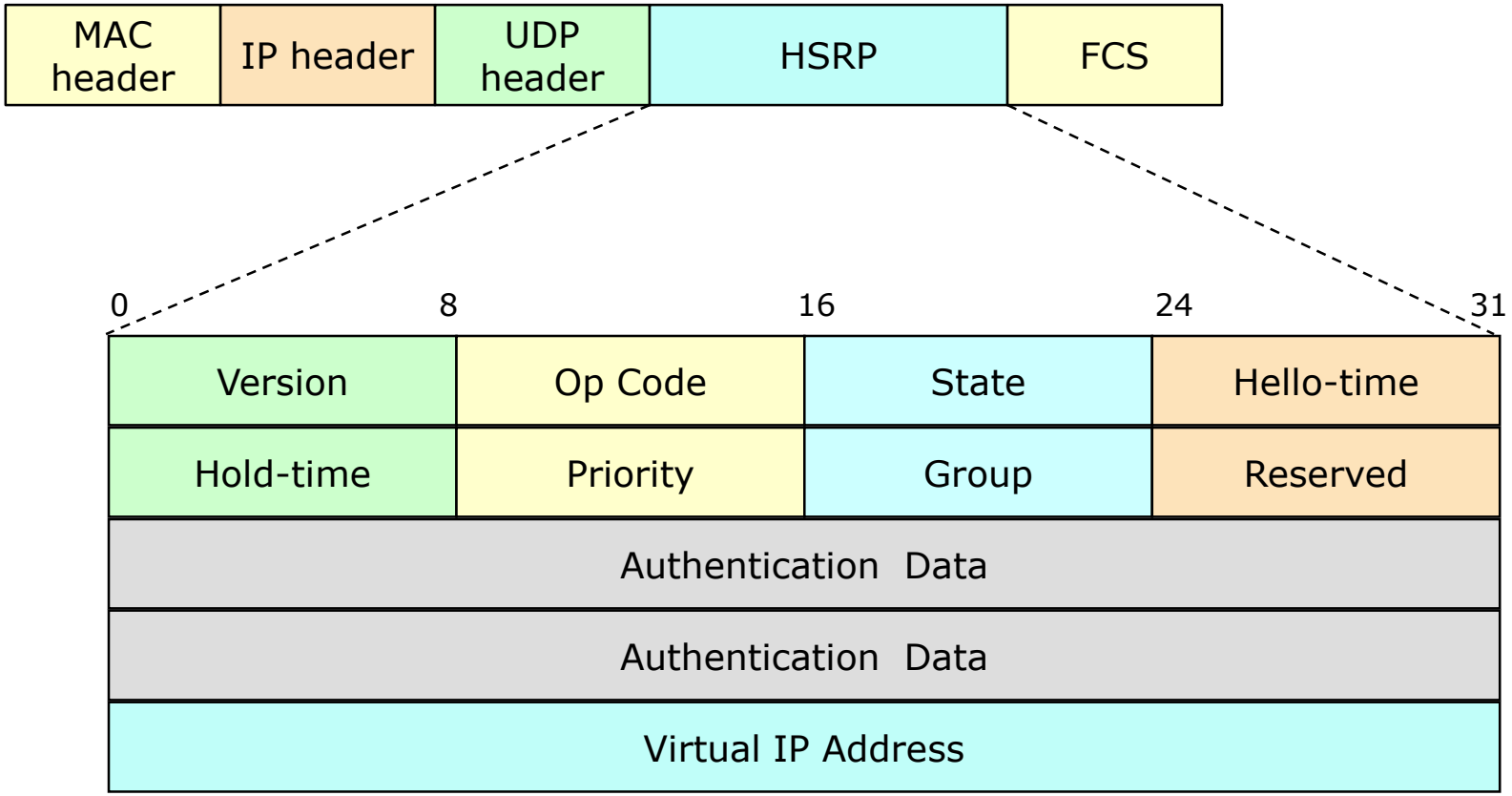
- Need to update the Filtering Database with the location of the Active Router
  - This is done by the Active router that sends HSRP Hello Messages with the Virtual MAC as source
  - When the Active router changes, the new Active router will start sending messages with the Virtual MAC address as source
  - All the switches will update their filtering database accordingly



HSRP and ARP caches

- 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
- Strictly speaking, not needed
  - ARP mapping does not change if the V-IP – V-MAC are used
- The reason stays in the old days
  - In some (old, Token Ring) cases, we may wanted to use the actual MAC address of the router instead of the virtual MAC
  - In those cases, we need to update the ARP cache
    - Although some Operating Systems may not handle it properly (because of the unexpected broadcast MAC address)

HSRP packet format



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 it 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)

- Possible values (cont.) :
  - 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)
- Notes
  - These numbers are set by the active router for the entire LAN
  - The convergence time for an HSRP network is about 10 seconds

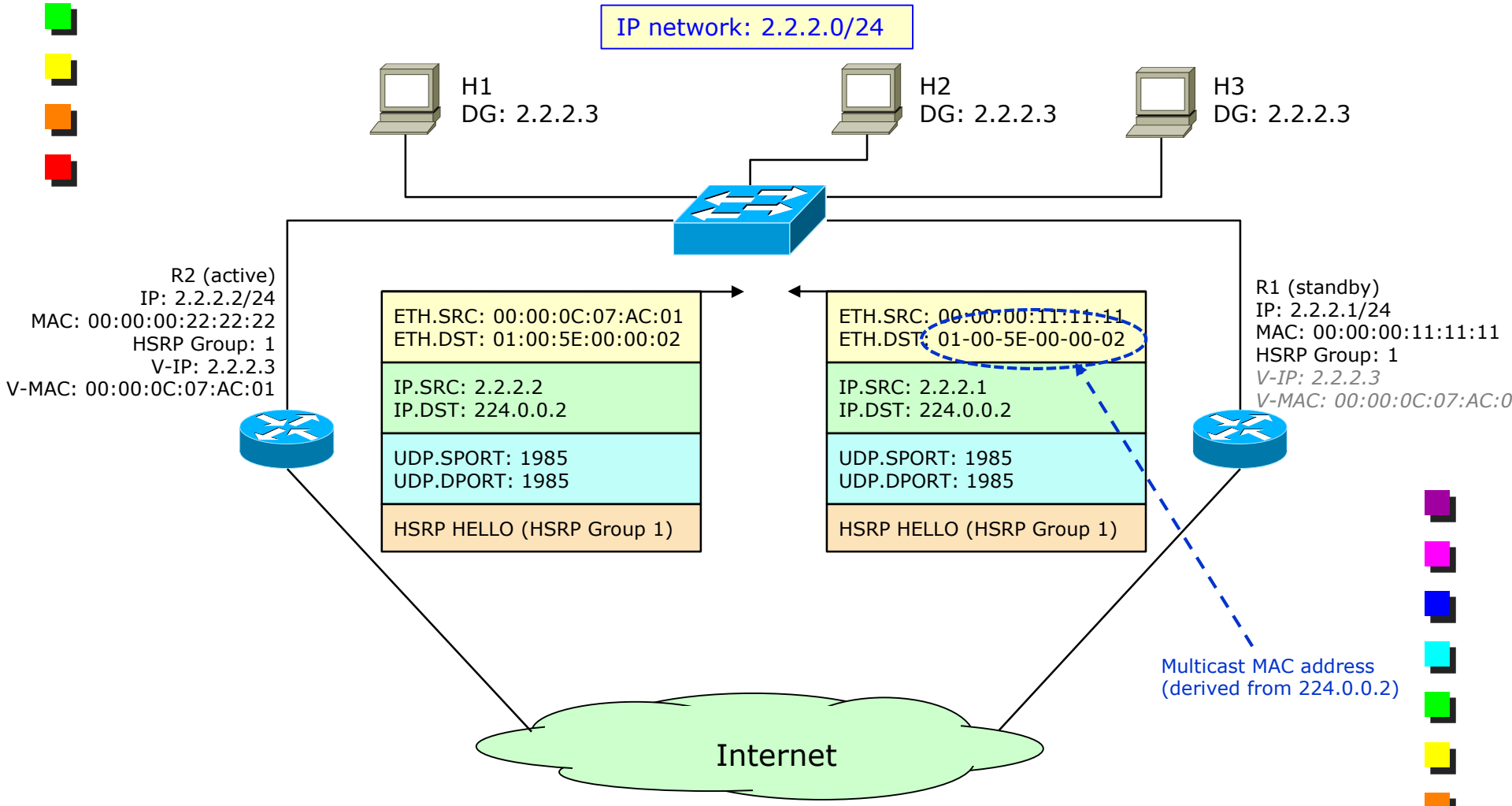
HSRP header: “priority” and “group” fields

- Priority
  - Used to force the election of the router with highest priority (higher number means higher priority)
    - IP addresses used in case of routers with equal priority
  - 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-characters 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
  - The virtual IP address used by this group
    - IP address of the default gateway configured on the hosts
  - If the virtual IP address is not configured on a router, then it may be learned from the Hello message sent from the active router

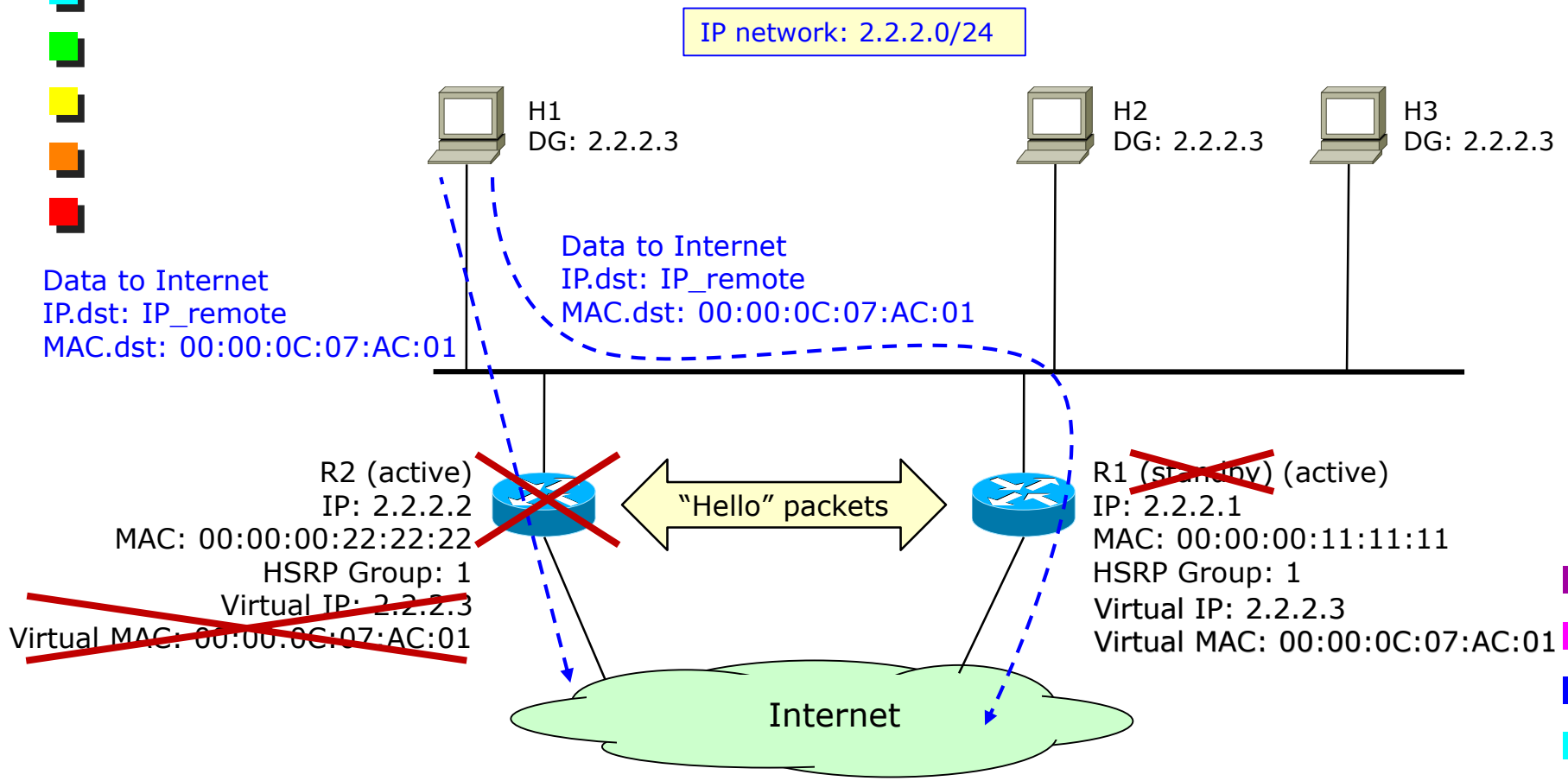
HSRP encapsulation (1)



HSRP encapsulation (2)

- Encapsulated in UDP, src/dst port is 1985
- IP Layer
  - Transmitted to multicast address 224.0.0.2 (“all routers”)
    - Well-known multicast address; does not require IGMP, hence it is propagated across the entire L2 network, even if the IGMP Snooping is active
  - Source IP address is the actual IP address of the router (in order to elect the best router as active)
  - TTL = 1 (packets are not further forwarded by routers)
- MAC layer
  - Destination address: derived from the Destination IP address
  - Source MAC: Virtual MAC address for the group

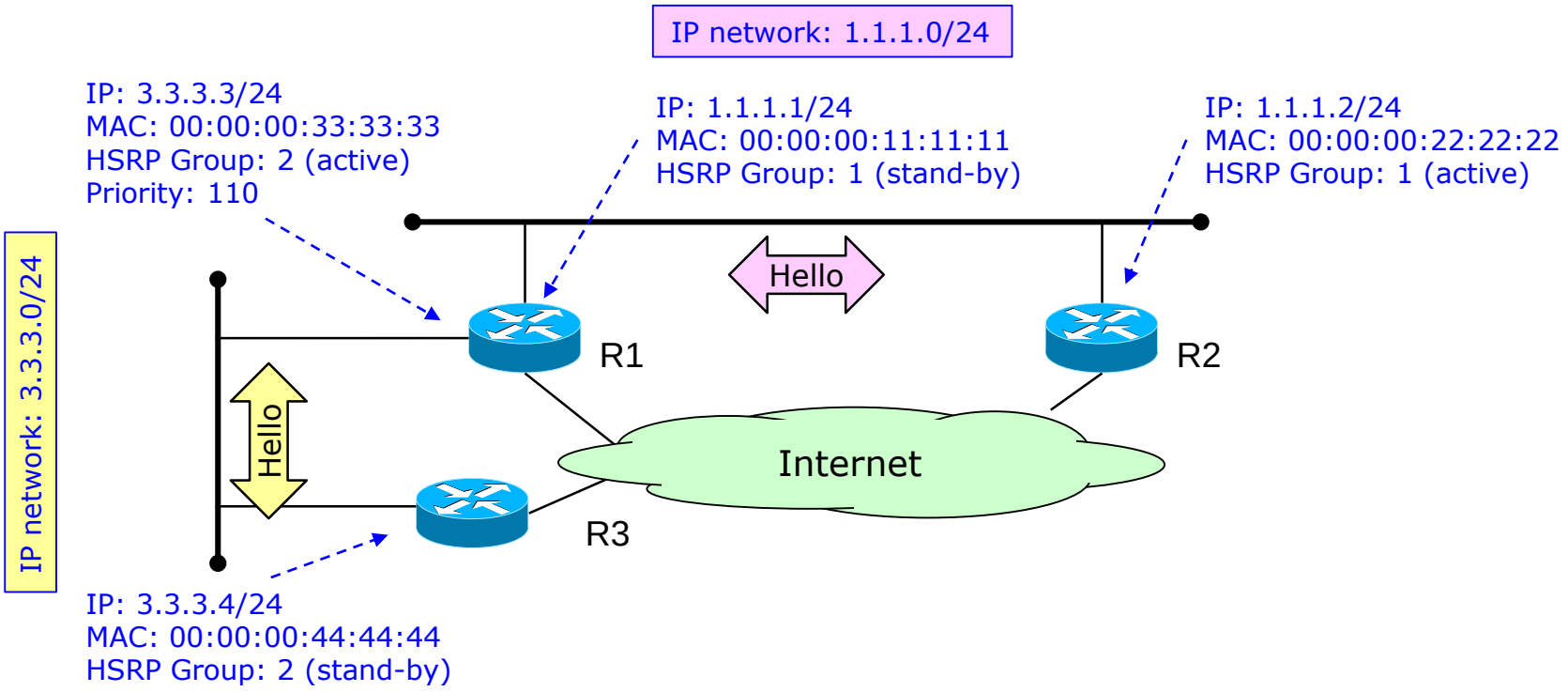
HSRP: fault reaction



Convergence time: depends on the HoldTime value, contained in the Hello packets (i.e., 10 seconds with default values).

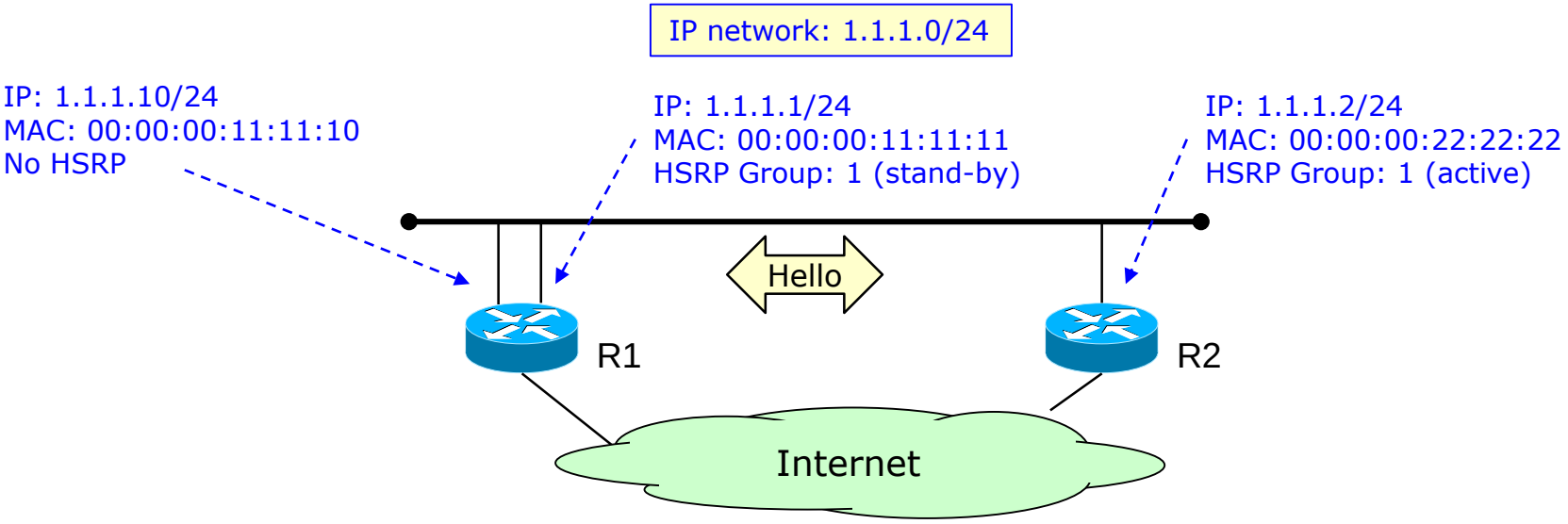
HSRP and interfaces (1)

- Interfaces on different IP networks must have different HSRP groups active
  - Router R1 has 2 HSRP instances, one per IP network
  - The router can assume *different roles* on the different groups



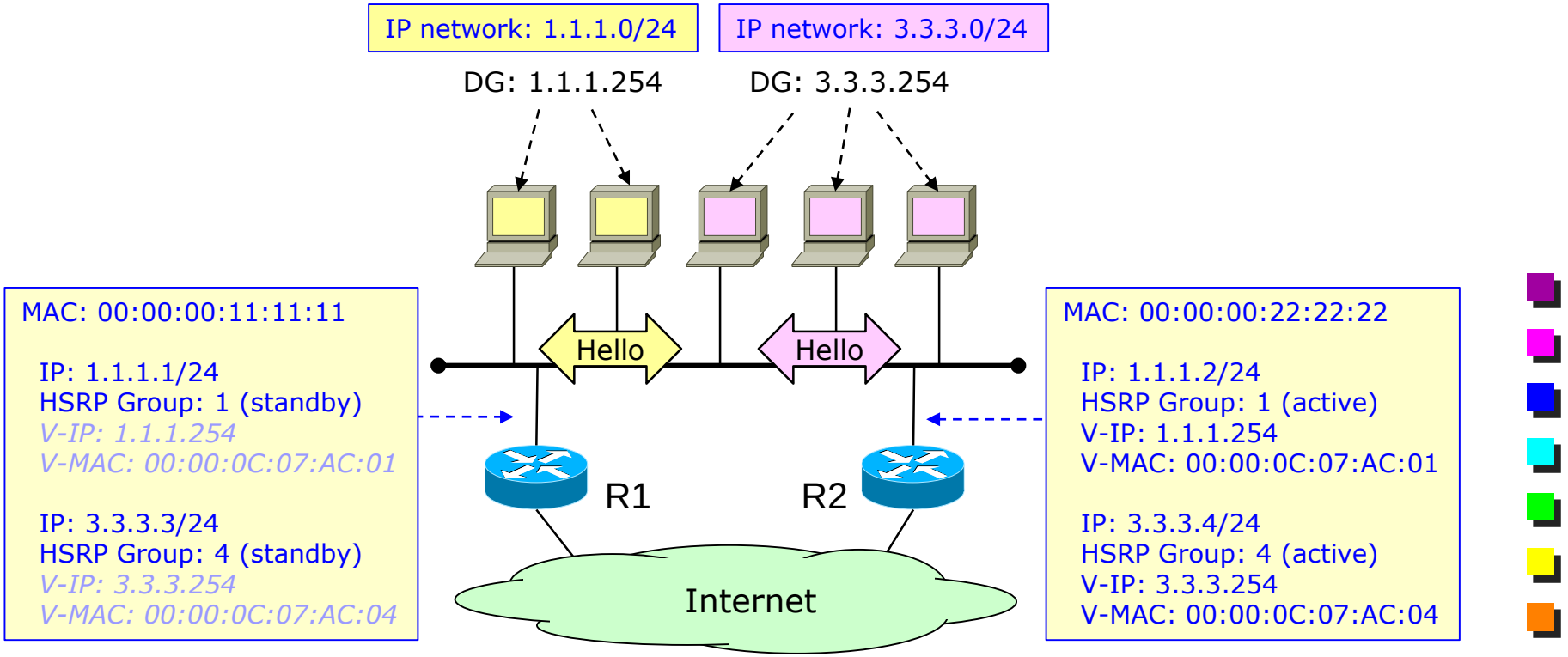
HSRP and interfaces (2)

- HSRP requires to be configured **per-interface**
  - Example: both interfaces on R1 belong to the same IP network, but we would like to enable HSRP only on the first interface



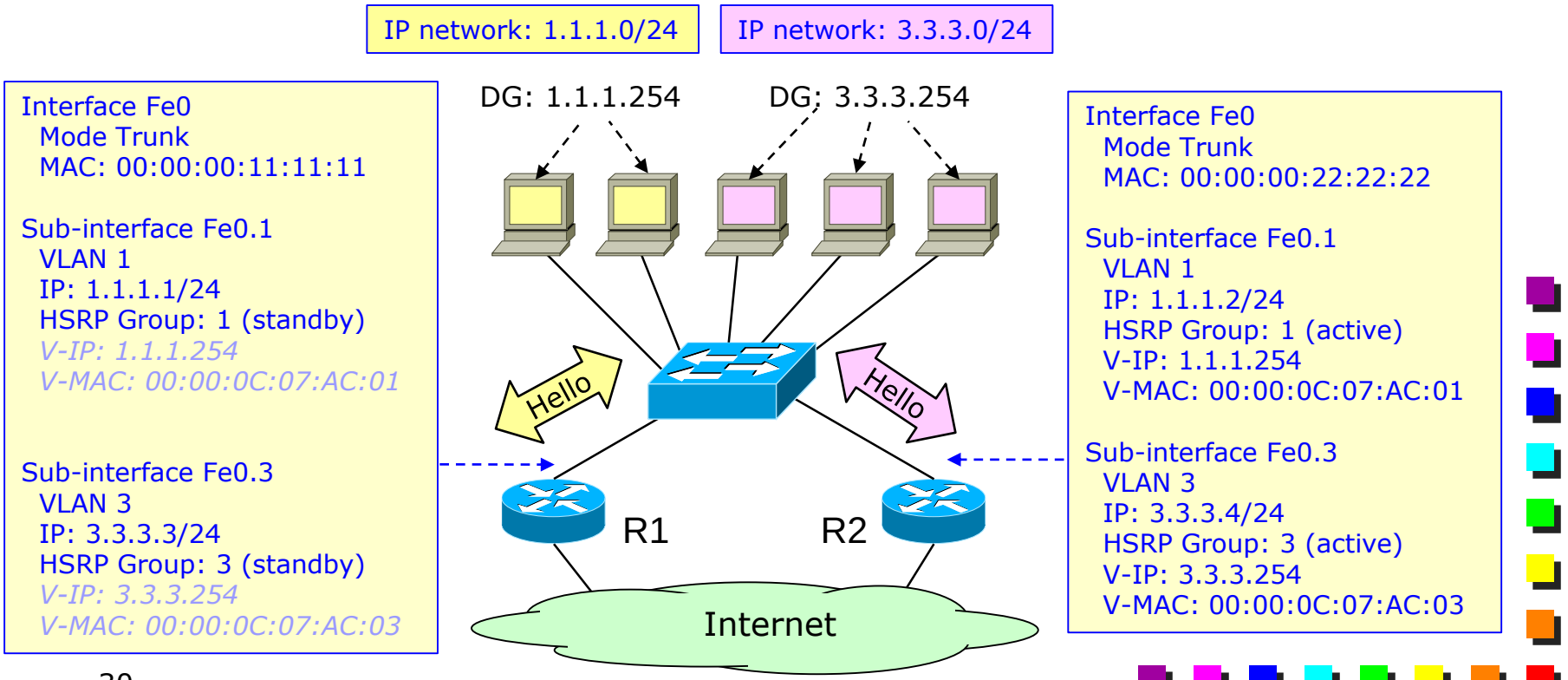
HSRP and IP networks

- HSRP is specific for a given IP network
  - HSRP interfaces belonging to the same group must be reachable at L2
  - A LAN with two Logical IP Networks must use two HSRP groups



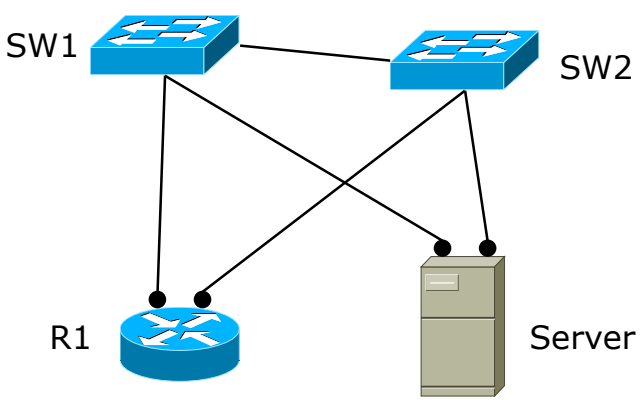
HSRP and VLANs

- Each VLAN is a separated LAN, with its own default gateway
- Multiple HSRP groups are required
  - Each VLAN goes on a specific virtual interface, each one with its own HSRP group



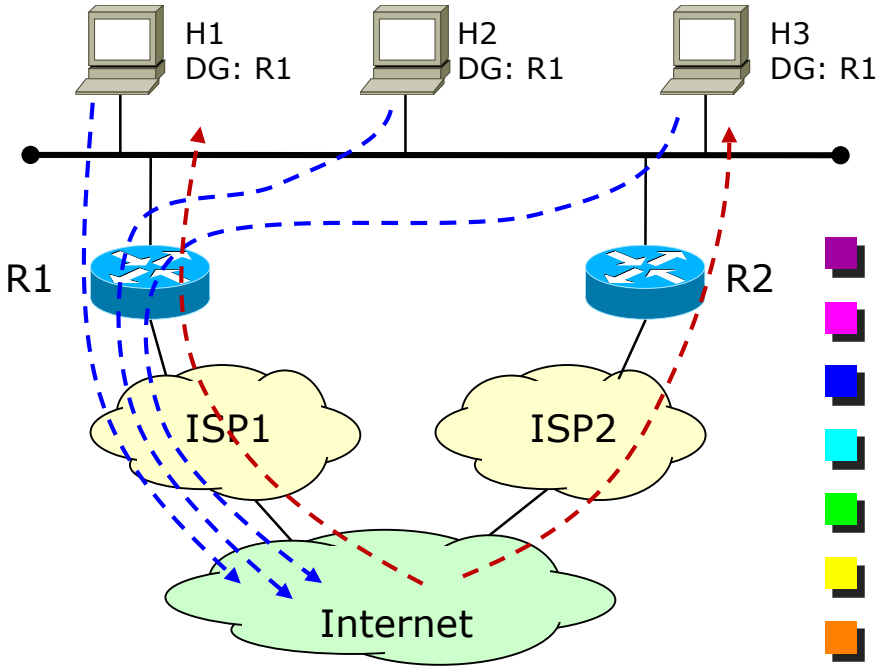
HSRP on dual-homed servers/routers

- HSRP can be used to implement Dual homing servers/routers and achieve single-machine redundancy
  - Two standard NICs
  - HSRP will handle the virtual IP/MAC address on the primary interface
  - Both interfaces must belong to the same LAN
- Other option for redundancy
  - Server (fault tolerant) NIC
    - Not available on routers



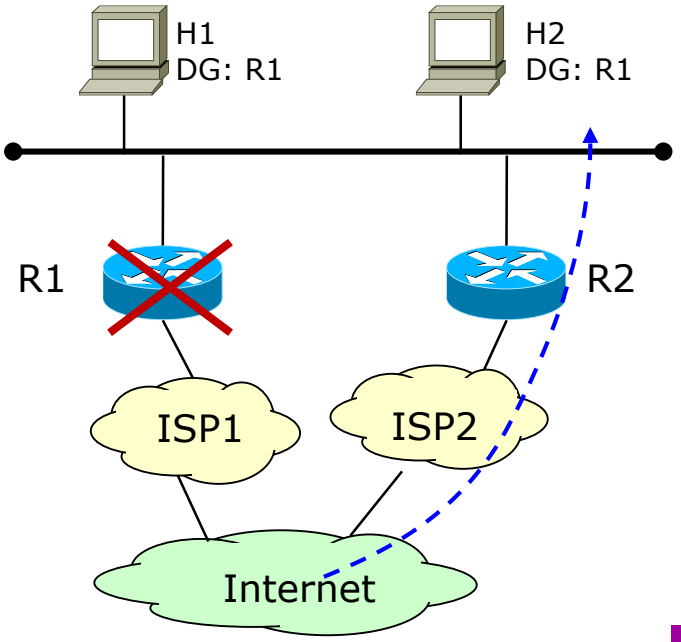
HSRP and asymmetric routing (1)

- Incoming traffic (from internet to the LAN) is routed by the external external routing protocol (OSPF, BGP, etc.)
  - The configuration of the incoming path does not depend on HSRP
  - We may use one or both links based on the configuration of those routing protocols
- Asymmetric routing
  - Ingress and egress paths may be different



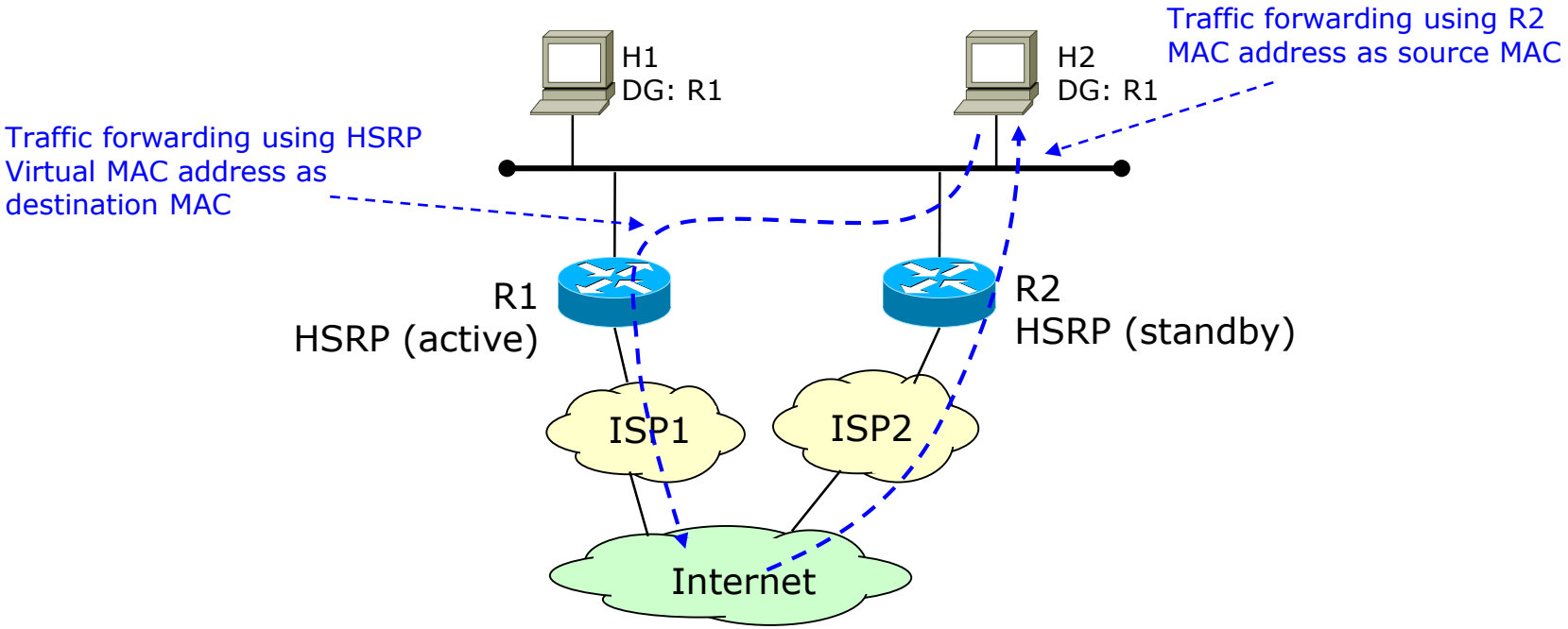
HSRP and asymmetric routing (2)

- Fault detection (for incoming traffic) depends on the external routing protocol
  - Routing protocols can detect failures of any router/path
    - E.g. the failure of a default gateway
  - Protection achieved *also if the LAN does not have HSRP*



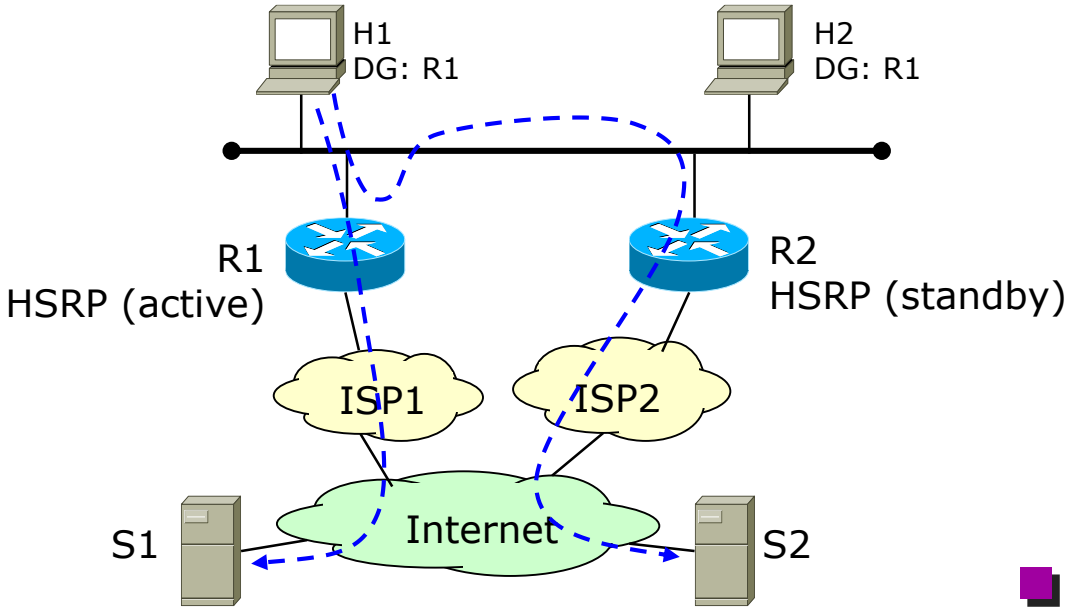
HSRP and traditional L3 routing (1)

- Stand-by (or listen) routers can (will!) route traffic directed to the LAN
  - In fact, this is part of the normal router operations
  - Non-active routers must use their actual MAC address



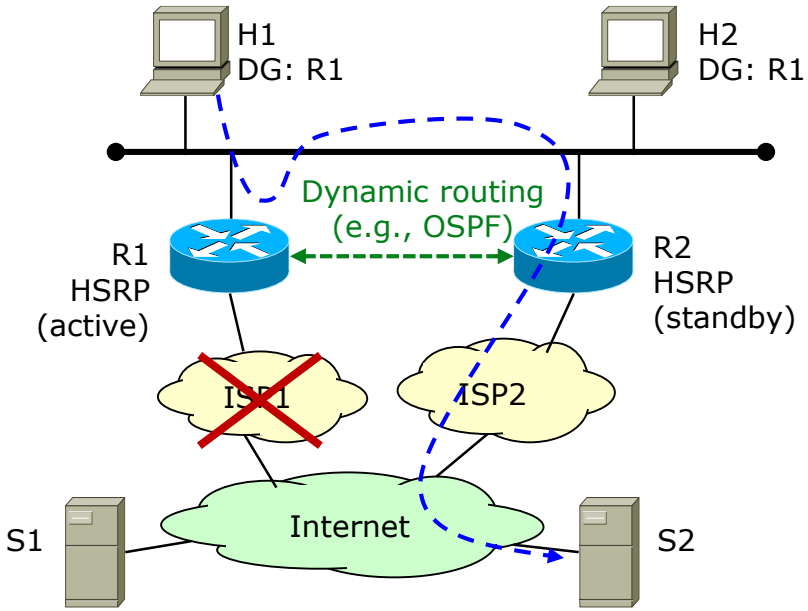
HSRP and traditional L3 routing (2)

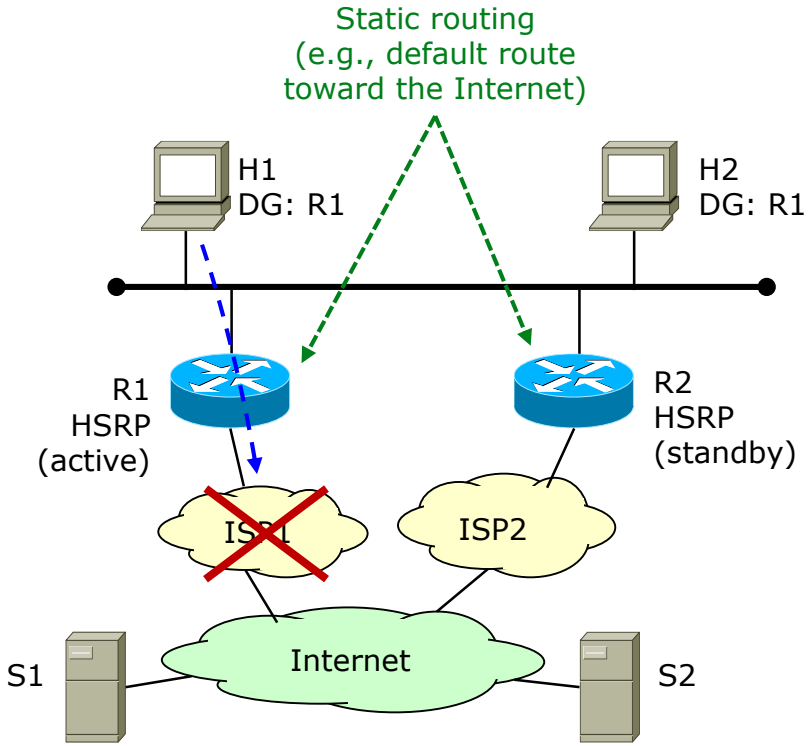
- Outgoing traffic may cross the stand-by router if the best L3 path goes there
  - Outgoing traffic is always sent to the Active router
  - From there, L3 routing protocols select the best path, independently from HSRP settings
  - ICMP Redirect can be generated by R1



DG redundancy and L3 routing: be careful!

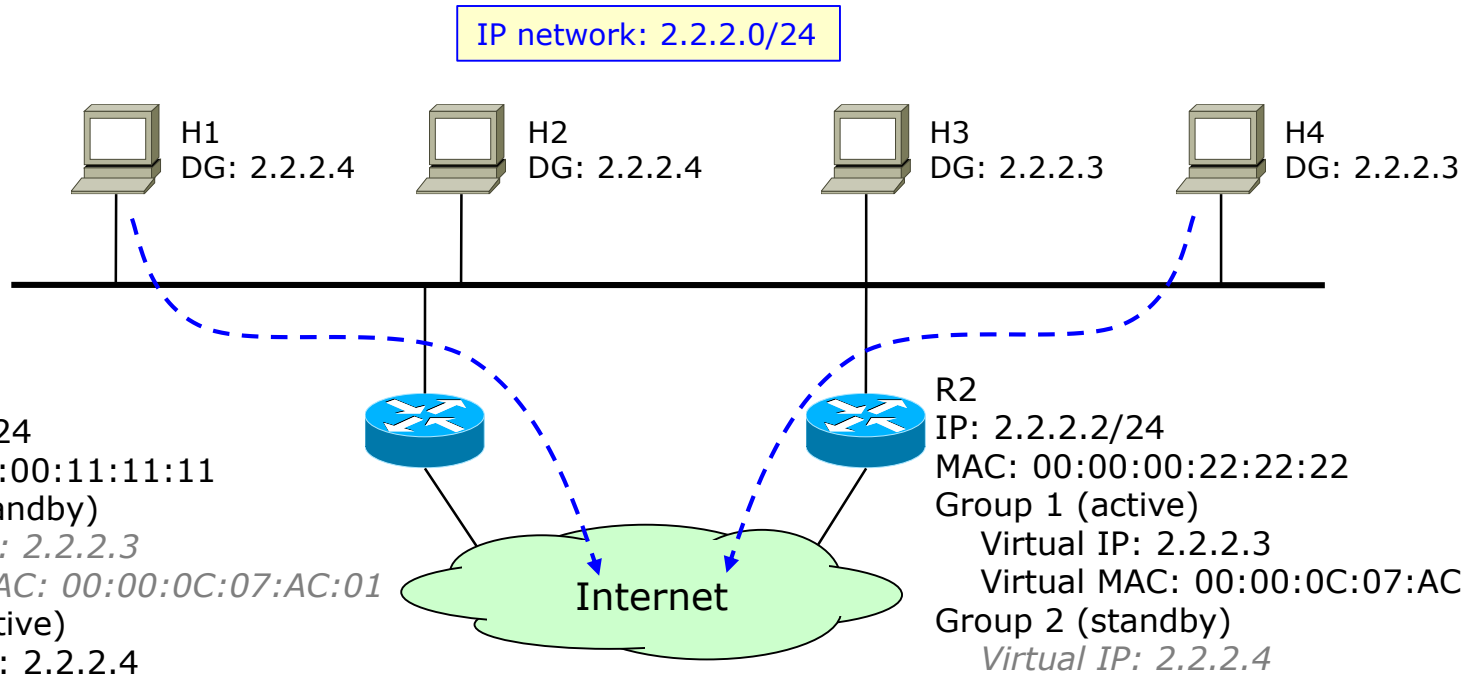
- HSRP protects from a fault of the default gateway
- HSRP does not influence at all the selection of the **exit path** toward the Internet!





HSRP and load sharing (1)

- Only one exit link used
  - Unused bandwidth on the stand-by router
  - Can be solved through a proper HSRP configuration



R1
IP: 2.2.2.1/24
MAC: 00:00:00:11:11:11
Group 1 (standby)
Virtual IP: 2.2.2.3
Virtual MAC: 00:00:0C:07:AC:01
Group 2 (active)
Virtual IP: 2.2.2.4
Virtual MAC: 00:00:0C:07:AC:02

R2
IP: 2.2.2.2/24
MAC: 00:00:00:22:22:22
Group 1 (active)
Virtual IP: 2.2.2.3
Virtual MAC: 00:00:0C:07:AC:01
Group 2 (standby)
Virtual IP: 2.2.2.4
Virtual MAC: 00:00:0C:07:AC:02

HSRP and load sharing (2)

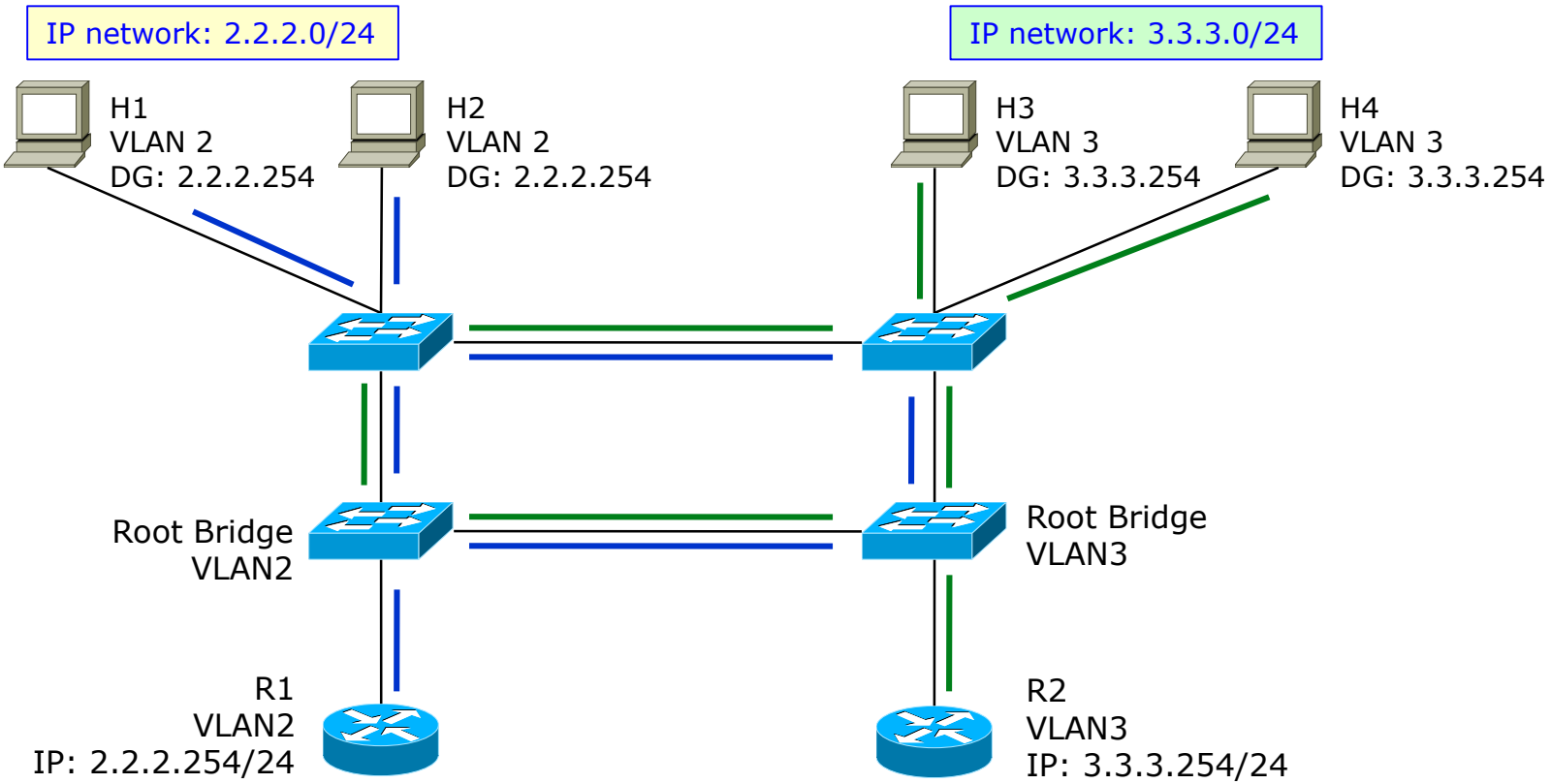
- Multiple HSRP groups
  - Multi-group HSRP (mHSRP)
  - One router active for the first group, the other active for the second group
  - Clients are configured half with the first DG, half with the second
- Problems of the achieved load balancing
  - Load balancing is statically defined by physically partitioning hosts between two different default gateways
    - What about if the traffic among the two groups is unbalanced?
  - Configuration burden
    - Not easy to differentiate the Default Gateway on the clients
    - DHCP usually returns a single DG for all hosts
  - Impact only on the outgoing traffic (asymmetric paths)

HSRP and load sharing (3)

- Is it really needed?
- 1) Default gateways toward the Internet
  - Often corporate networks have much more incoming than outgoing traffic
  - The routing of incoming traffic does not depend on the HSRP configuration (see next slide)
  - In this case, the egress bandwidth of a single link may be enough
- 2) One-arm router for VLAN interconnections
  - May be useful, due to the large amount of traffic present there
  - We can exploit the forwarding capabilities of both routers

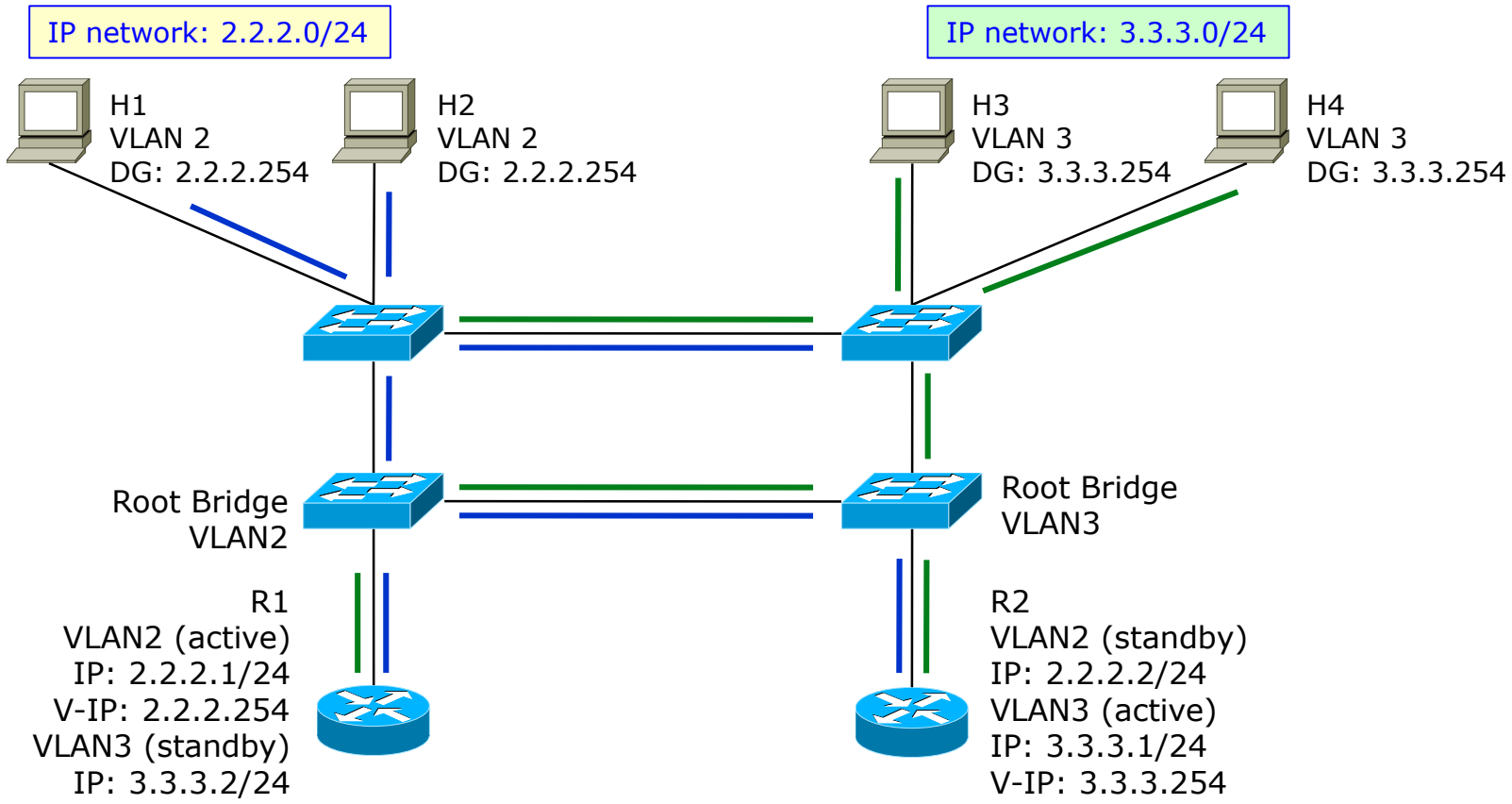
Load sharing with VLANs (1)

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



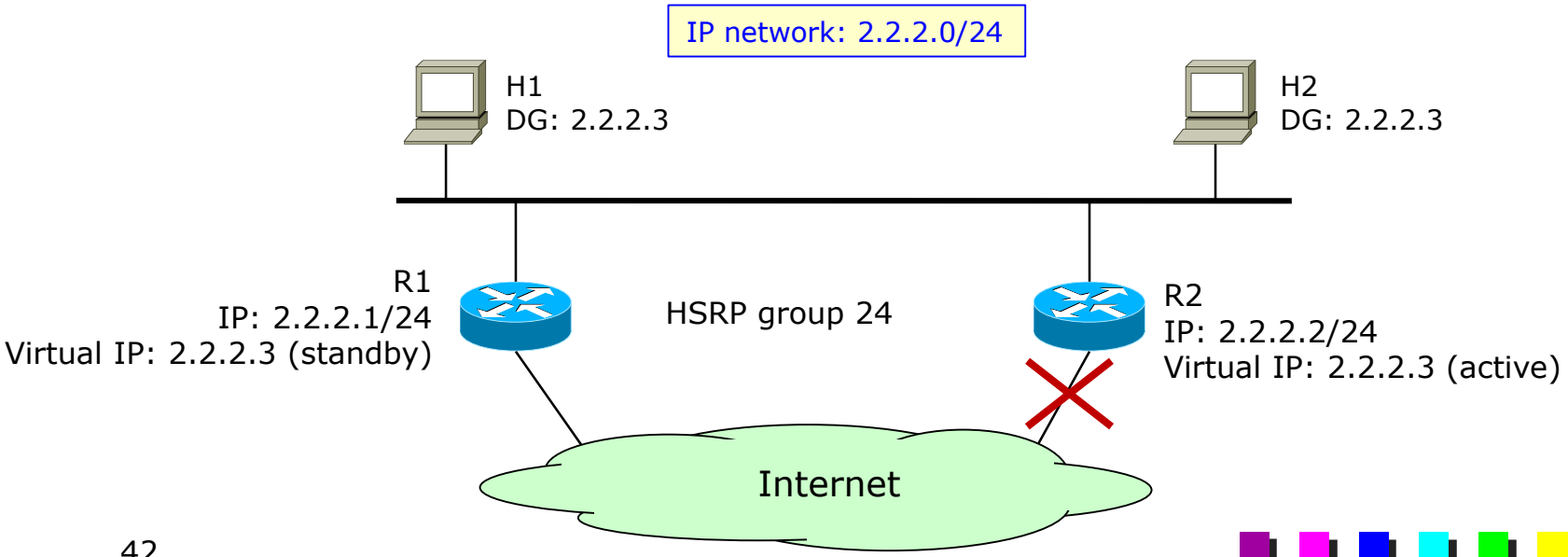
Load sharing with VLANs (2)

- Much better solution
- Two HSRP groups, but in two different VLANs



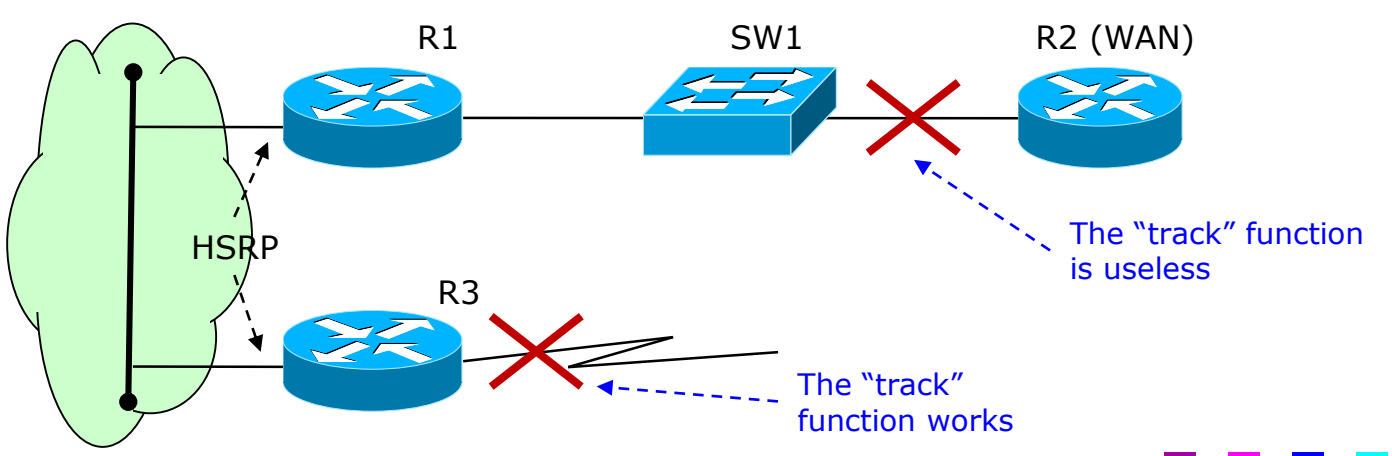
HSRP: "track" function (1)

- 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 (routing protocol know the possible routes)
    - Opposite route works as well (not under HSRP control, though)
  - No problems in connectivity, but additional overhead in forwarding



HSRP: "track" function (2)

- "Track" the status of the physical layer on the interface
  - 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 faults cause the interface to go down
  - E.g. an interface will stay up if connected to an active L2 switch
  - Connectivity still fine thanks to L3 routing (with one more hop)



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

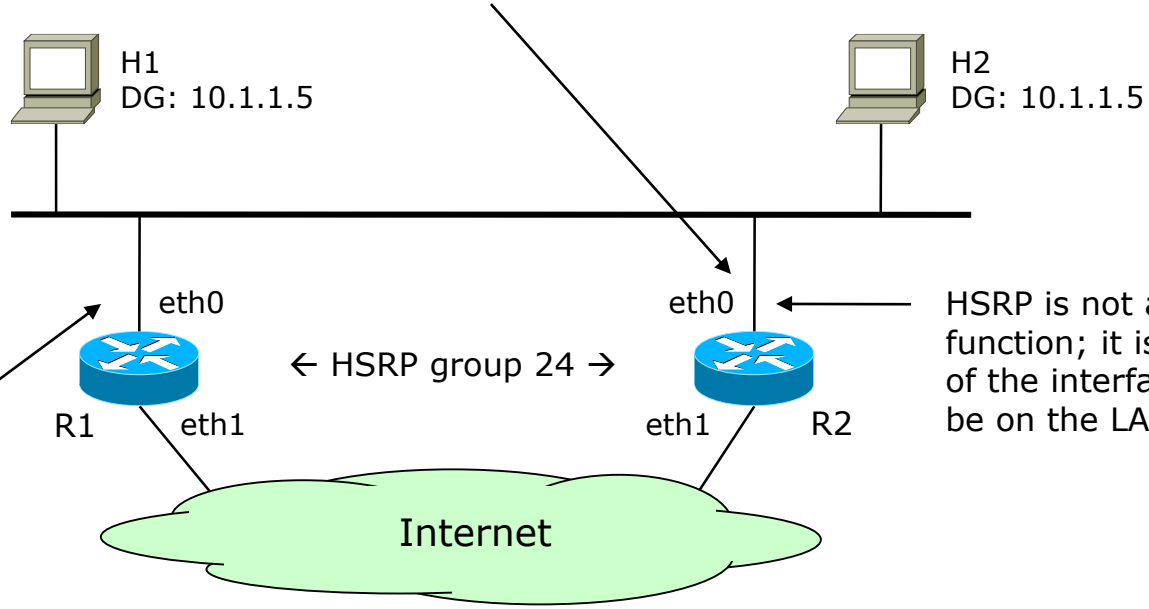
HSRP: convergence

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

HSRP: basic configuration

```
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
```

IP network: 10.1.1.0/24

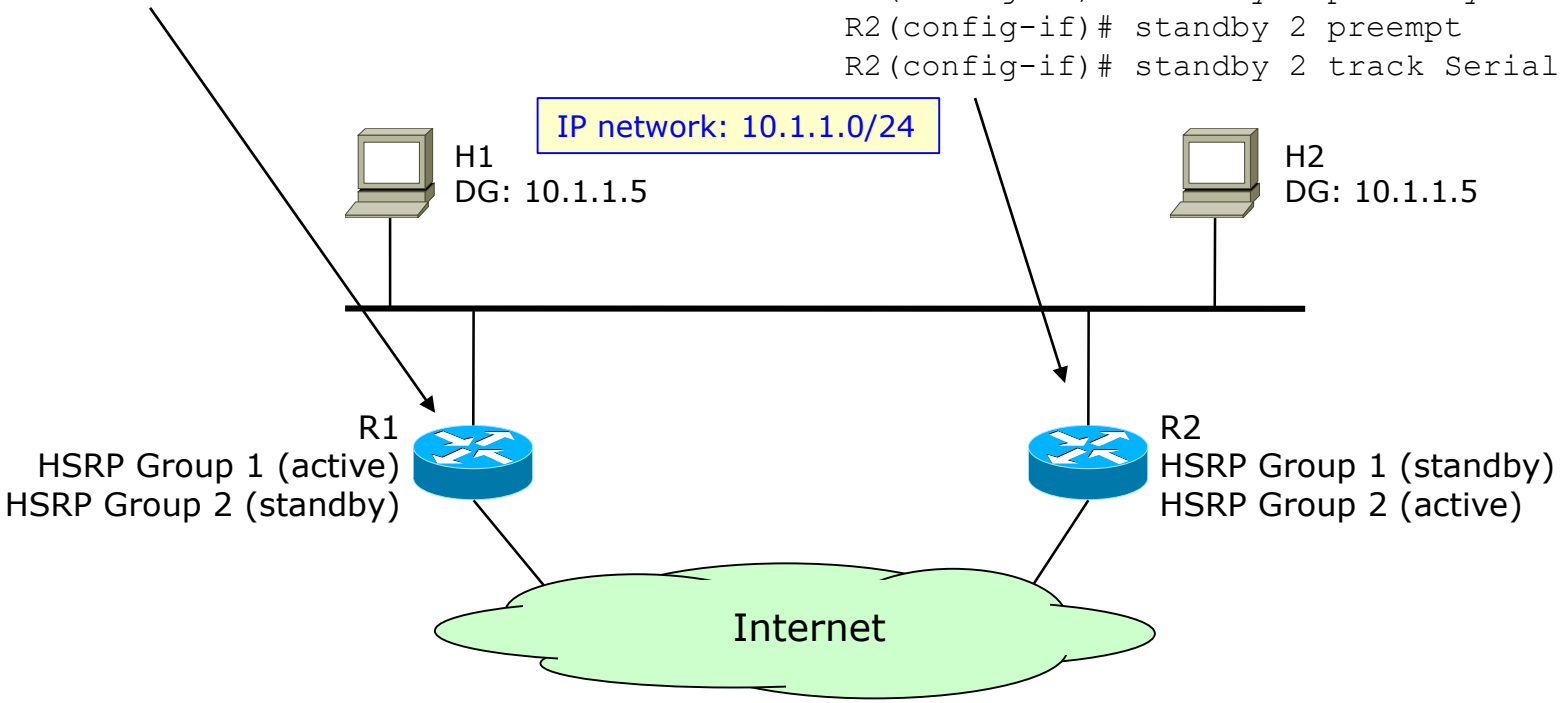



```
R1(config)# interface ethernet 0
R1(config-if)# ip address 10.1.1.1 255.255.255.0
R1(config-if)# standby 24 ip 10.1.1.5
R1(config-if)# standby 24 priority 105
R1(config-if)# standby 24 preempt
```

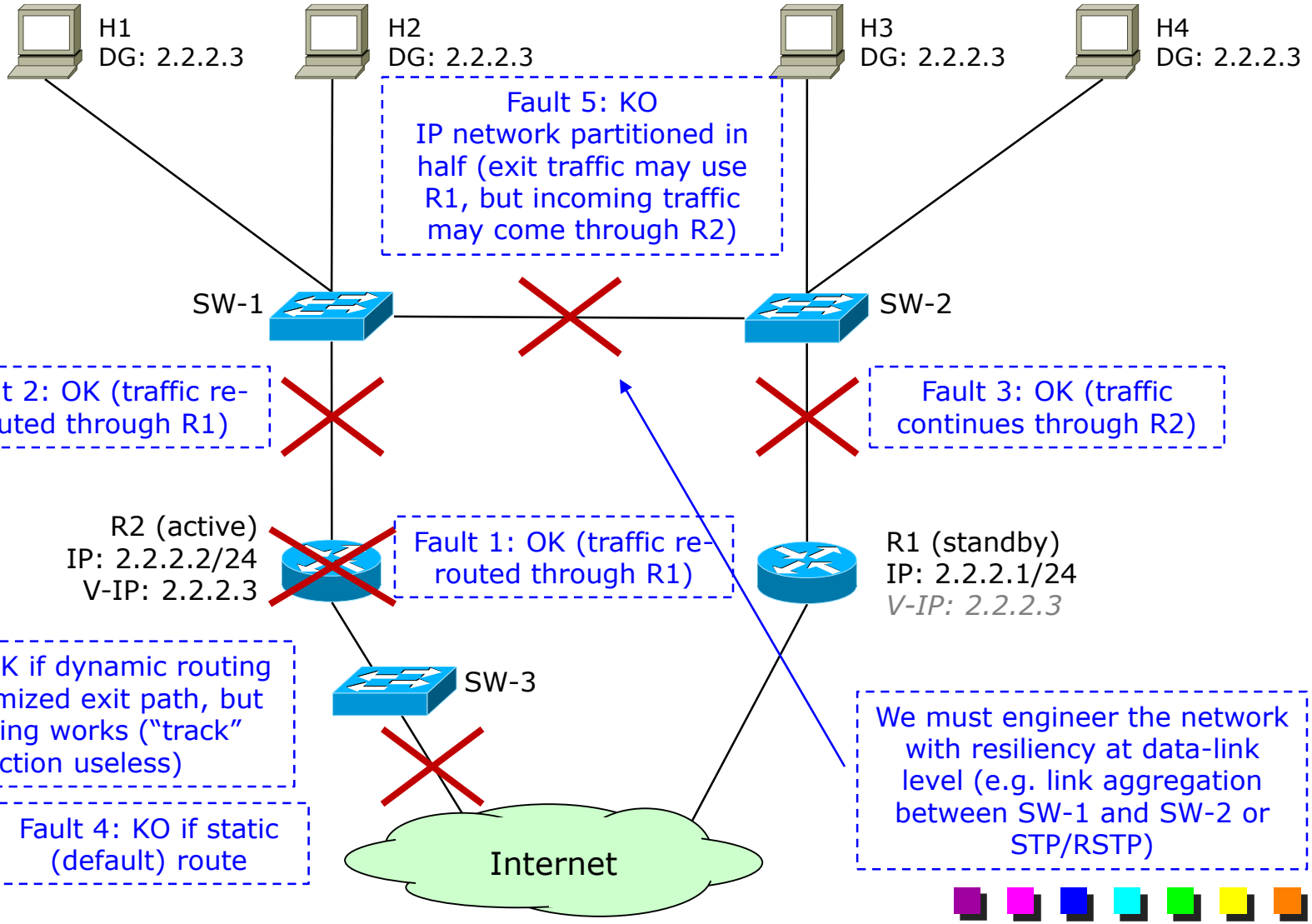
HSRP: advanced configuration

```
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 priority 105
R2(config-if)# standby 2 preempt
R2(config-if)# standby 2 track Serial0
```



HSRP on real LANs: L2 resiliency (1)



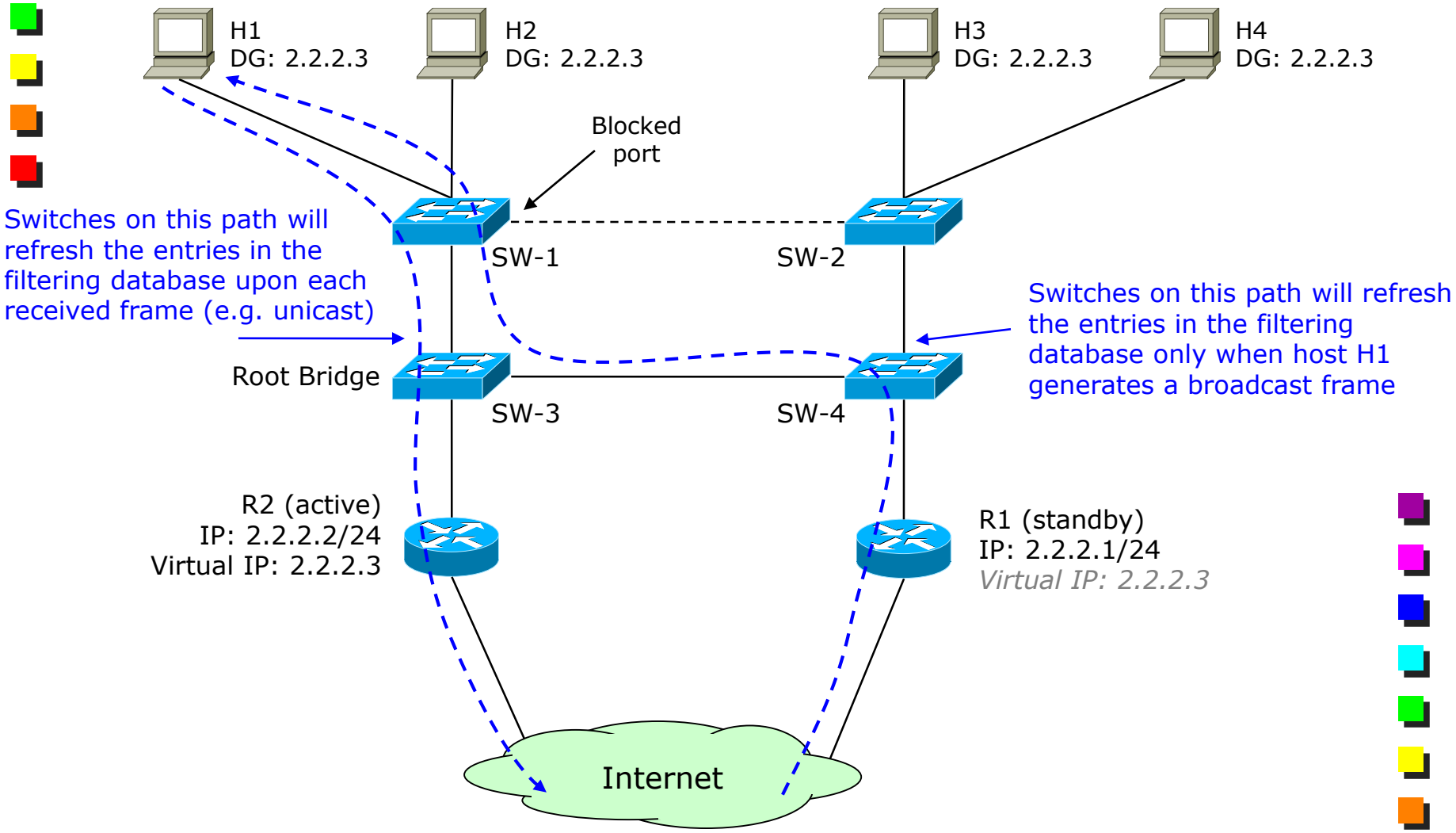
HSRP on real LANs: L2 resiliency (2)

- HSRP does 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 does not protect from some faults on the WAN link

HSRP on real LANs: flooding (1)

- Switches may have incomplete filtering database
  - Some entries may be missing, due to the Aging Time
  - If a frame directed to that host is received, the frame is flooded
- A periodic generation of broadcast frames avoids the problem
  - Some hosts generate a limited number of broadcasts (e.g. UNIX or VmWare)
- Sometimes ARP messages are sent only “occasionally” on the network
  - ARP cache is often 5 minutes or more
  - Max Ageing Time in the filtering database is often 2 minutes

HSRP on real LANs: flooding (2)



HSRP on real LANs: flooding (3)

- A possible pathological situation
  - Let's assume that an hosts A that wants to contact host B has still the mapping (IP(B), MAC(B)) in its ARP cache
  - However, the MAC(B) is no longer in the filtering database of the switches
  - In that case, host A will not send an ARP Request to B, and the MAC frame is generated and sent on the network
    - Flooding
  - Please note that in any case, the ARP Reply may not reach the entire network, although usually this is not a problem
- This situation may be extremely common with HSRP/VRRP

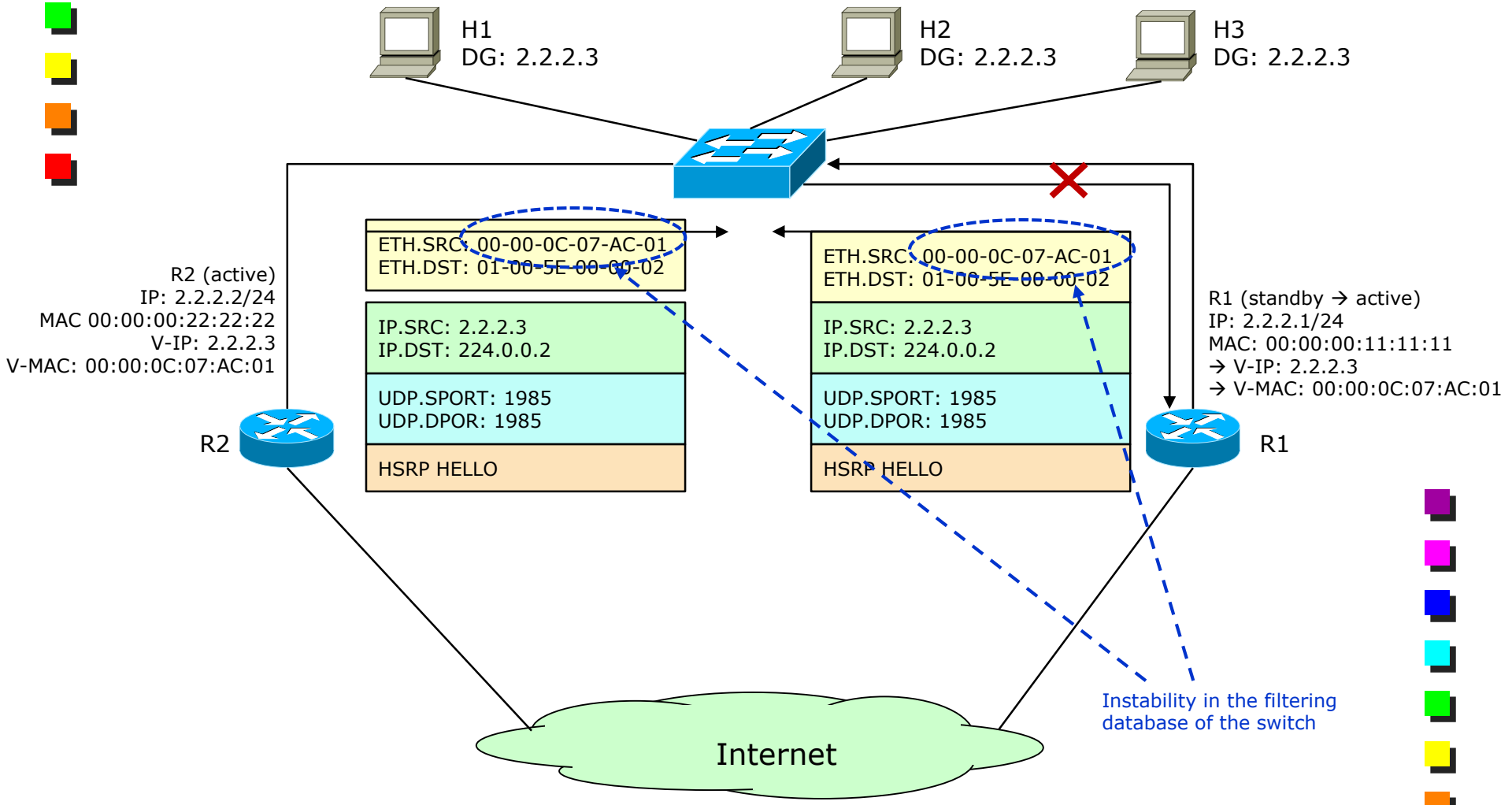
HSRP on real LANs: flooding (4)

- Hypothesis
  - Let us suppose that host H1 does not generate periodic broadcasts on its own
  - Egress and ingress routers are different
    - Asymmetric routing
  - The ARP Cache on router R2 lasts longer than the filtering database
- Description
  - Egress traffic from station H1 updates the filtering database on the exit path
  - The ingress router still has a valid mapping for H1 in its ARP cache
  - Switches on the ingress path do no longer have the H1 entry in the filtering database
  - The only option (for the switch) is to send the frame on all its ports
  - ➔ periodic flooding (every now and then)

HSRP on real LANs: flooding (5)

- Solutions
  - Re-engineer the L2 spanning tree (not really a solution)
  - Force stations to send broadcast frames rather often (< Max Ageing Time)
  - Increase Max Ageing Time on the switches

HSRP on unidirectional links (1)



HSRP on unidirectional links (2)

- R1 will not receive HSRP packet from R2, therefore both will become active
  - Both R1 and R2 will send HSRP Hello using the virtual MAC address as source
- The filtering database on the switch SW-1 will oscillate periodically
- Note that the problem appears despite R1 will not answer to ARP Requests ("*Who has 2.2.2.3?*") because the incoming path is unavailable

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
  - Same way to achieve the Default Gateway redundancy
  - Same way to achieve Load Balancing

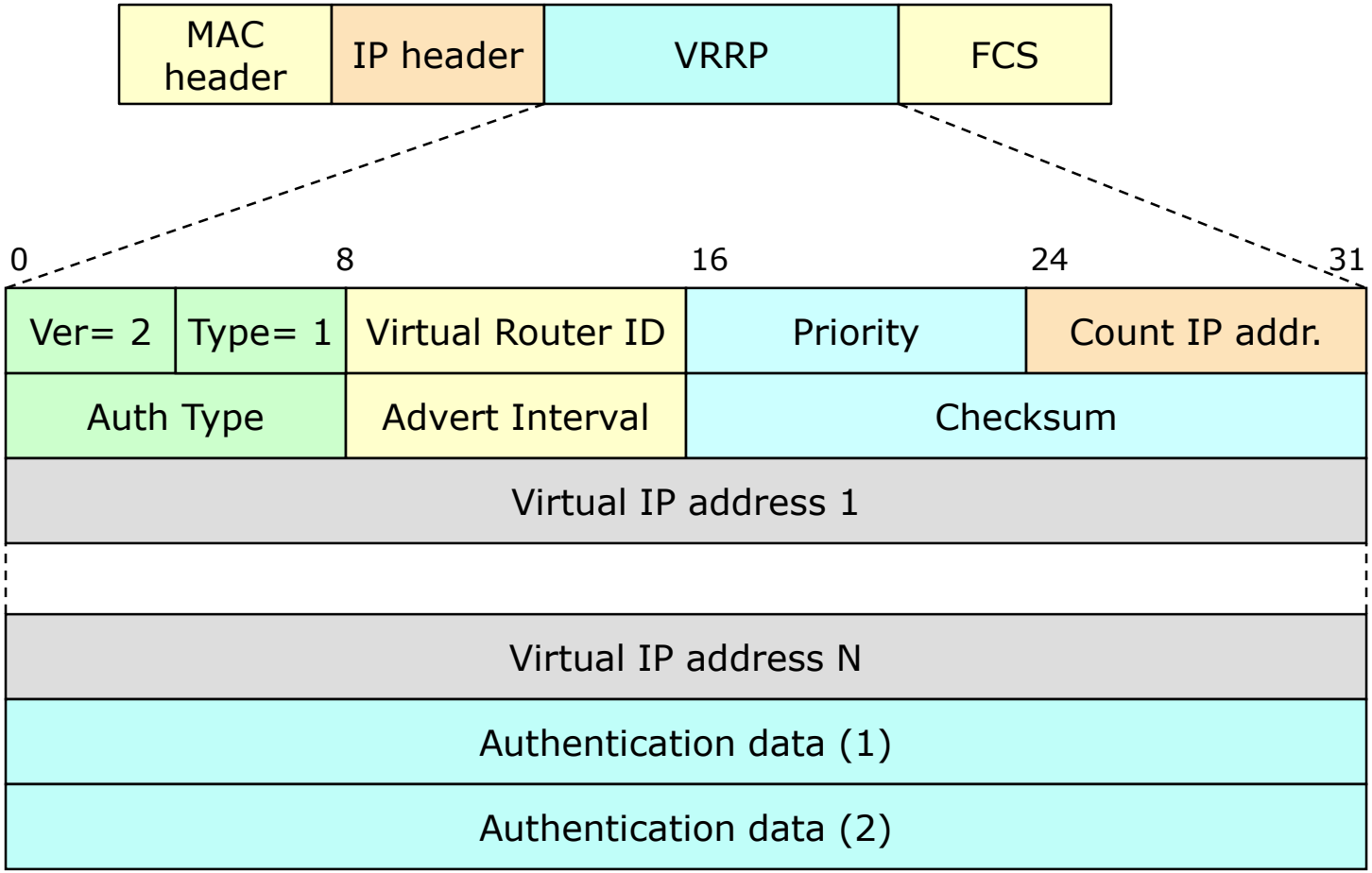
VRRP overview (2)

- 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 (3)

- 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: 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 - Priority) / 256 (seconds)
- Master_Down_Interval
  - (3 * Advertisement_Interval) + 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

VRID	Token Ring Functional Address
1	03-00-02-00-00-00
2	03-00-04-00-00-00
3	03-00-08-00-00-00
4	03-00-10-00-00-00
5	03-00-20-00-00-00
6	03-00-40-00-00-00
7	03-00-80-00-00-00
8	03-00-00-01-00-00
9	03-00-00-02-00-00
10	03-00-00-04-00-00
11	03-00-00-08-00-00

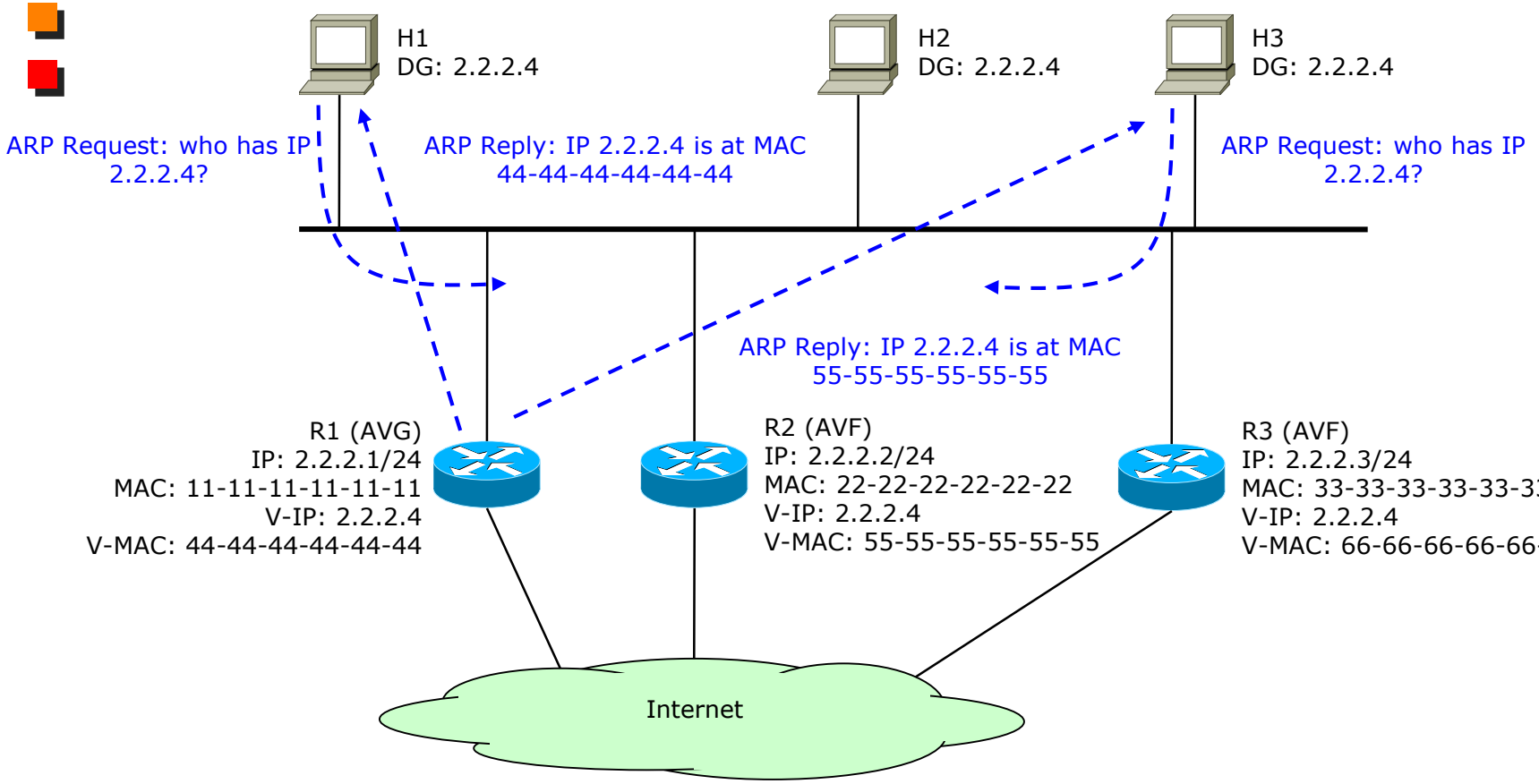
VRRP: convergence

- About 4 sec with default parameters
  - (3 * Advertisement_Interval) + Skew_time
  - User can configure Advertisement Interval and Priority
  - Less flexible than HSRP

GLBP

- Gateway Load Balancing Protocol
- Enhancement (and replacement) of HSRP
- 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 provides a unique virtual router service
  - One IP address
  - Multiple virtual MAC addresses for forwarding
- Cisco proprietary
  - Not even available on the entire product line

GLBP: the idea



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 GLBP 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)
  - In order to move that traffic to another router in case that AVF fails
- 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 the forwarding for that AVF

Load Balancing algorithms

- **None**: GLBP operates like HSRP
- **Weighted**: each GLBP router in the group advertises its weight 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 use the same virtual MAC address as long as the number of AVFs in the GLBP group is constant
  - Used when the router provides also the NAT function to the external world 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 (per VLAN) is used
  - Usually outgoing traffic much smaller than incoming traffic
- GLBP
  - Proprietary, not documented
- Be careful with the L2 network
  - L2 resiliency
  - L2 flooding
  - Other faults
    - “Track” function not effective