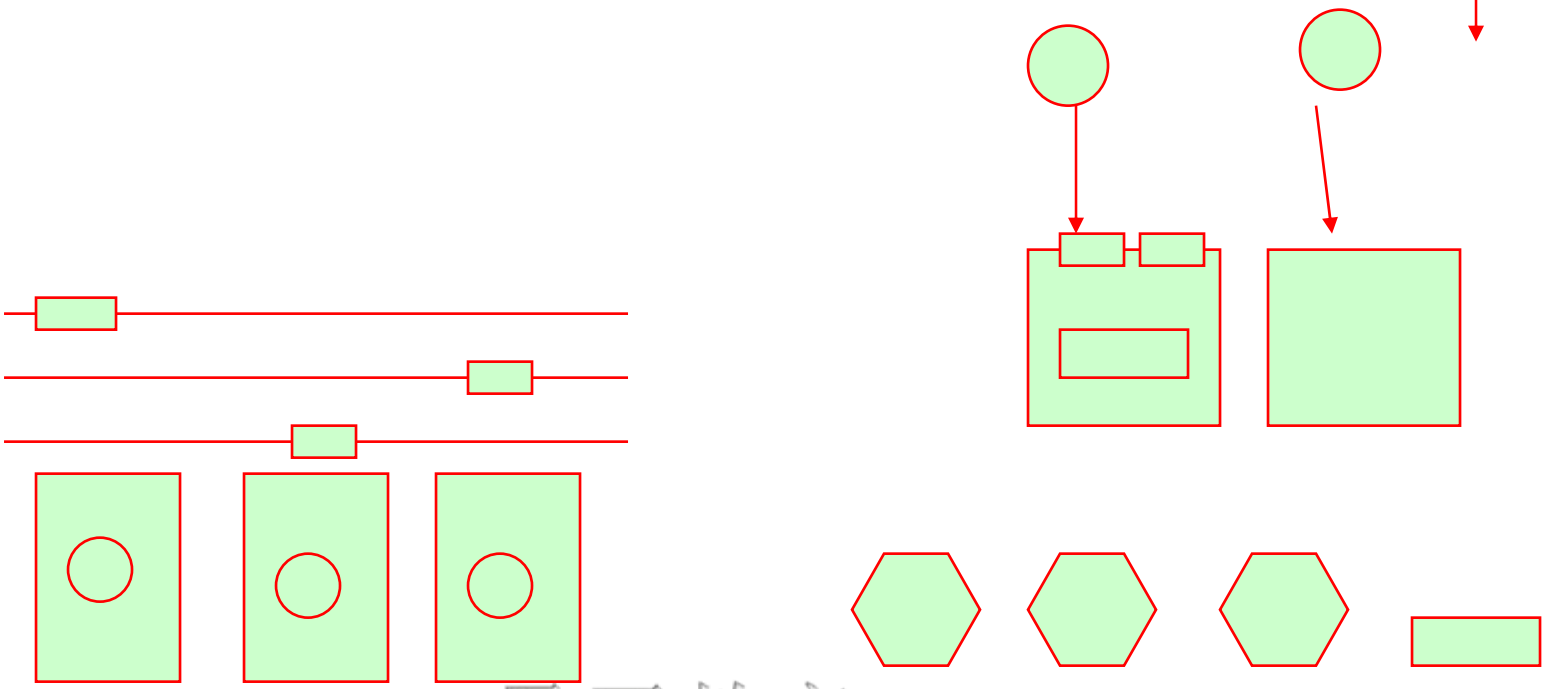
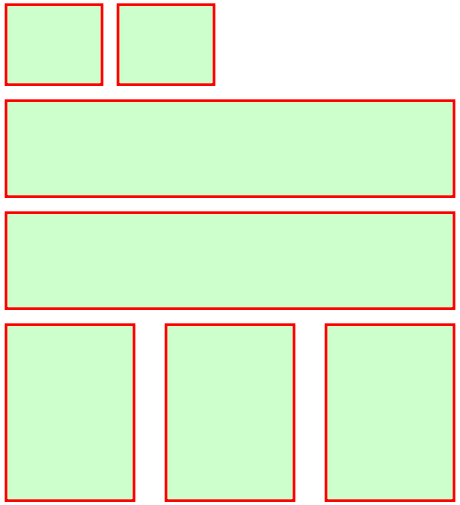
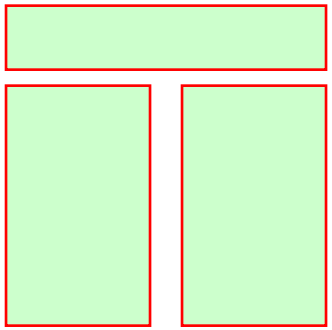
马哥教育

Keepalived

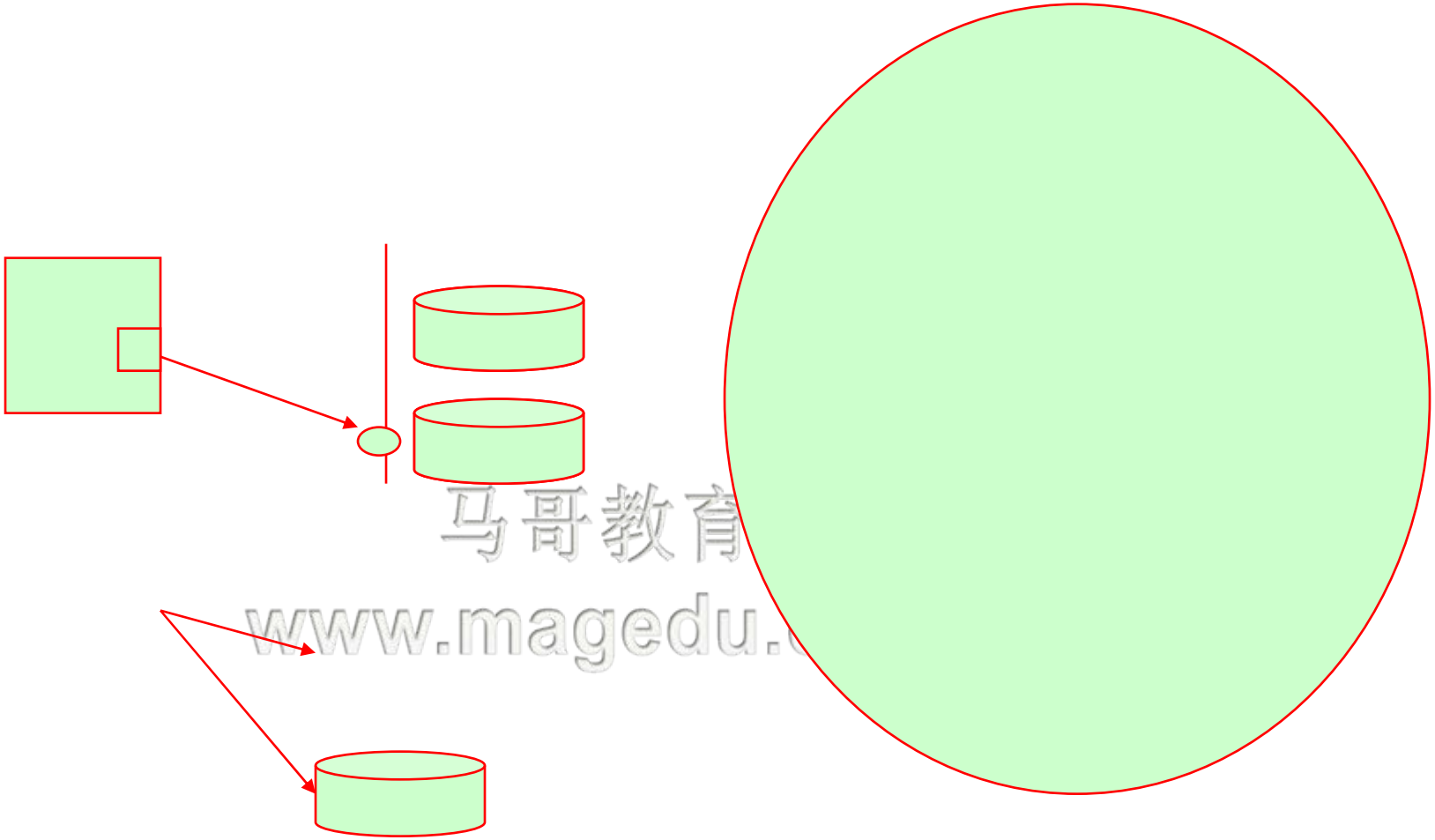
主讲：马永亮(马哥)
QQ群:169777636
客服QQ：2813150558，1661815153
http://www.magedu.com
http://mageedu.blog.51cto.com

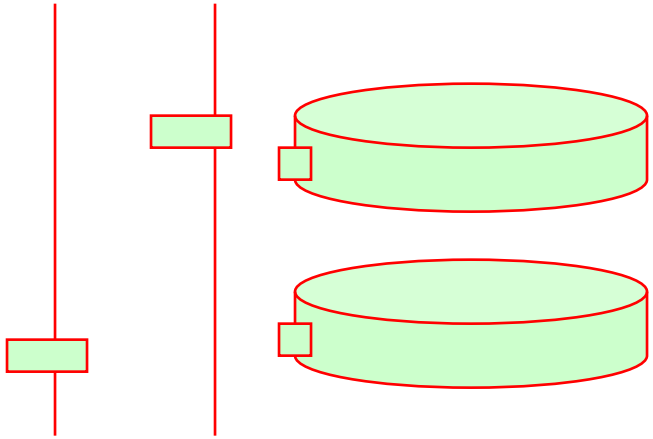


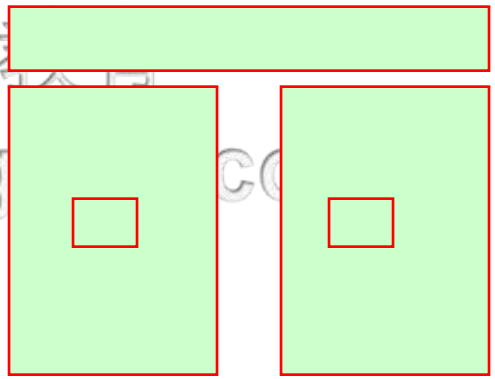




vip, vmac







Agenda

- Keepalived基础
- vrrp协议
- keepalived安装配置
  - keepalived中配置vrrp实例
  - keepalived中配置ipvs

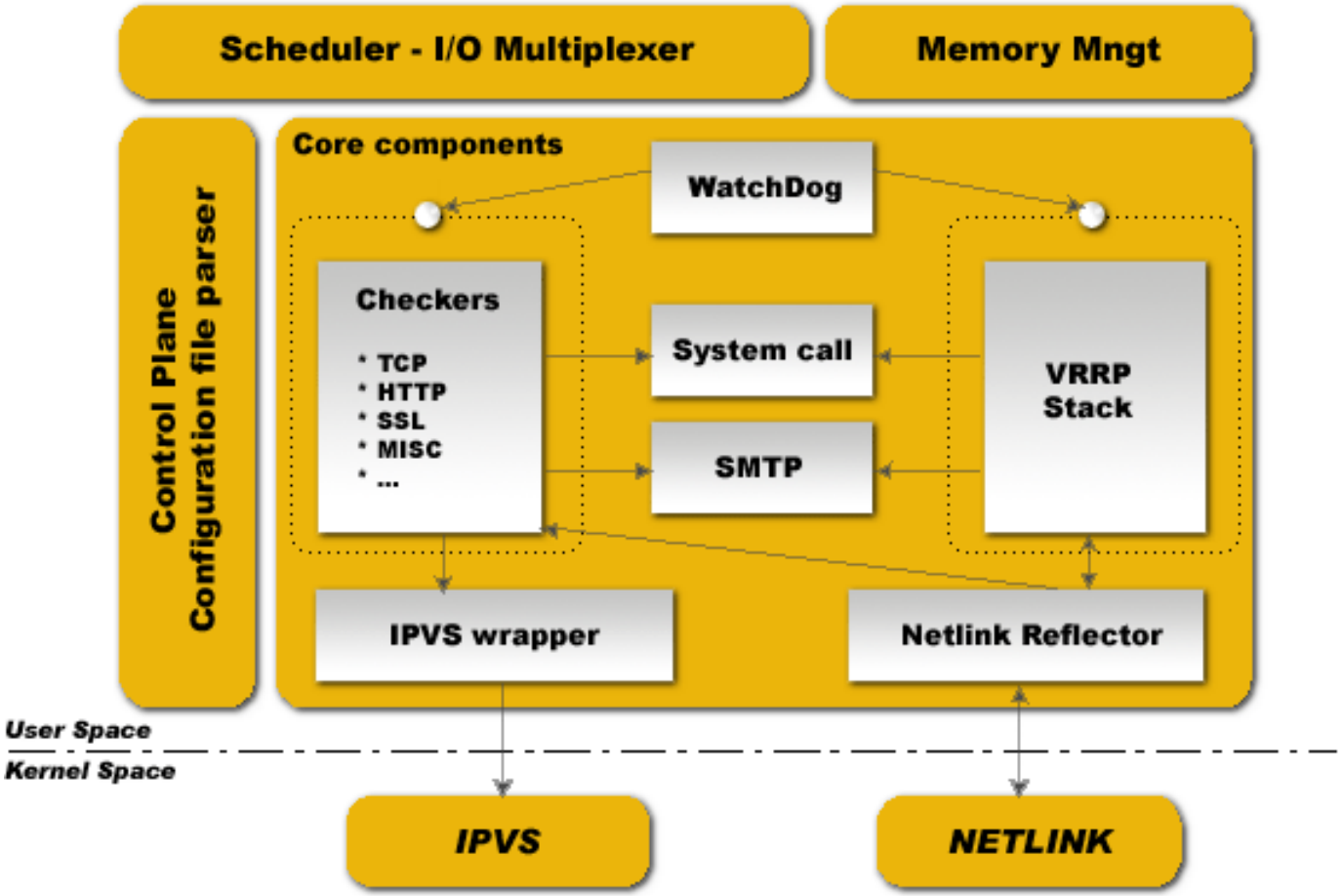
Overview

- Free software, GPLv2
- A routing software written in C
- To provide simple and robust facilities for loadbalancing and high-availability to Linux system and Linux based infrastructures
  - Loadbalancing framework relies on well-known and widely used Linux Virtual Server (IPVS) kernel module providing Layer4 loadbalancing
- Keepalived implements a set of checkers to dynamically and adaptively maintain and manage loadbalanced server pool according their health

Overview

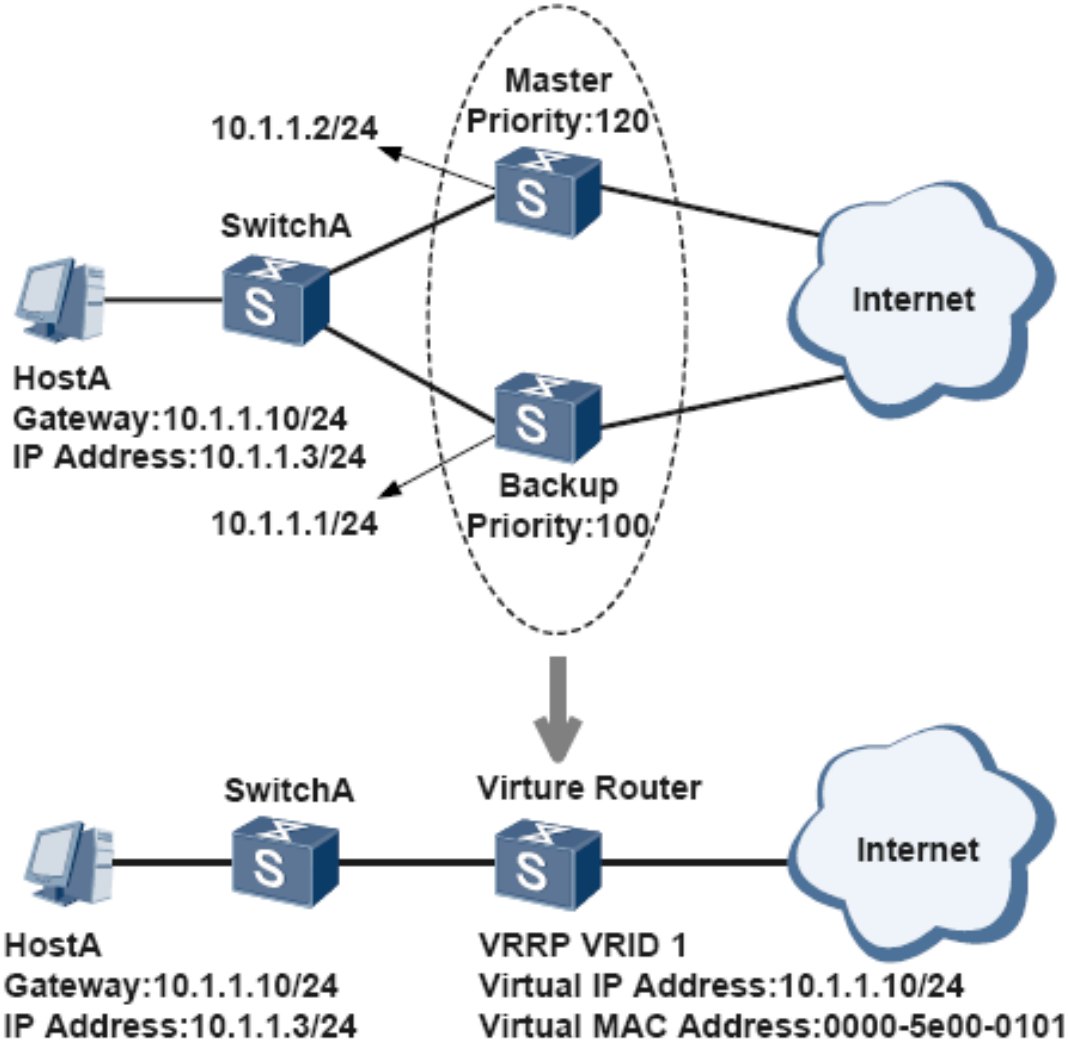
- On the other hand high-availability is achieved by VRRP protocol
  - VRRP is a fundamental brick for router failover
  - Implements a set of hooks to the VRRP finite state machine providing low-level and high-speed protocol interactions
- Keepalived frameworks can be used independently or all together to provide resilient infrastructures

Software Design



- To ensure robustness and stability, daemon is split into 3 distinct processes
  - The global design is based on a minimalistic parent process in charge with forked children process monitoring
  - Then 2 children processes, one responsible for VRRP framework and the other for health checking
- Each children process has its own scheduling I/O multiplexer
- The parent process monitoring framework is called watchdog

vrrp基本概念



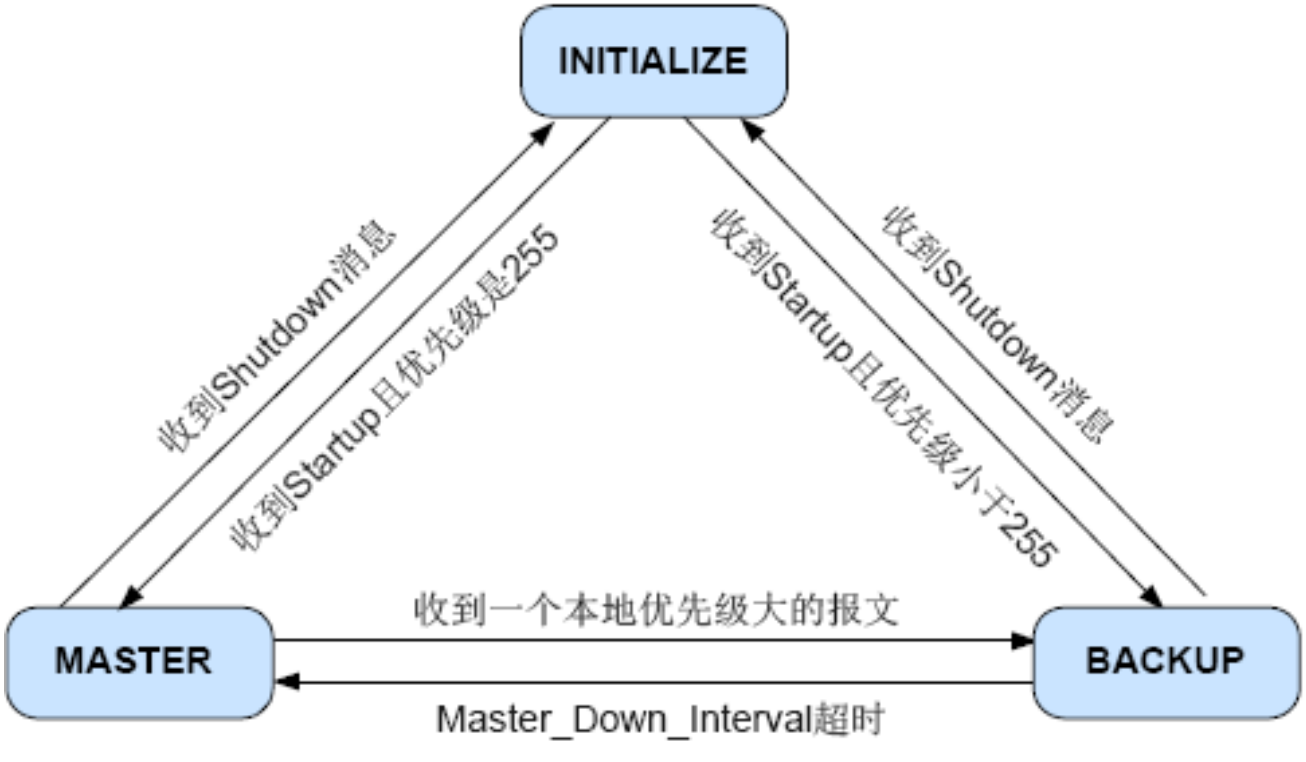
Reference: www.huawei.com/ilink/cnenterprise/download/HW_201057

- 虚拟路由器(Virtual Router)：又称VRRP备份组，由一个Master设备和一个或多个Backup设备组成，被当作一个共享局域网内主机的缺省网关
- VRID：虚拟路由器的标识；拥有相同VRID的一组路由器构成一个虚拟路由器；
- Master：虚拟路由器中承担报文转发任务的路由器；
- Backup：Master路由器出现故障时，能够通过竞选等成为代替Master路由器工作的路由器；
- 虚拟IP：虚拟路由器的IP地址；一个虚拟路由器可以拥有一个或多个IP地址，由用户进行配置；
- IP地址拥有者(IP Address Owner)：如果一个VRRP设备将虚拟路由器IP地址作为真实的接口地址，则该设备被称为IP地址拥有者；如果IP地址拥有者是可用的，通常它将成为Master；

- 虚拟MAC地址(Virtual MAC Address)：虚拟路由器根据虚拟路由器ID生成的MAC地址
  - 一个虚拟路由器拥有一个虚拟MAC地址，格式为：00-00-5E-00-01-{VRID}(VRRP for IPv4)；00-00-5E-00-02-{VRID}(VRRP for IPv6)
  - 当虚拟路由器回应ARP请求时，使用虚拟MAC地址，而不是接口的真实MAC地址
- 主IP地址(Primary IP Address)：从接口的真实IP地址中选出来的一个主用IP地址，通常选择配置的第一个IP地址
  - VRRP广播报文使用主IP地址作为IP报文的源地址

- 优先级(Priority)：虚拟路由器中VRRP设备的优先级
  - 虚拟路由器根据优先级选举出Master设备和Backup设备
- 抢占模式：在抢占模式下，如果Backup设备的优先级比当前Master设备的优先级高，则主动将自己切换成Master
- 非抢占模式：在非抢占模式下，只要Master设备没有出现故障，Backup设备即使随后被配置了更高的优先级也不会成为Master设备

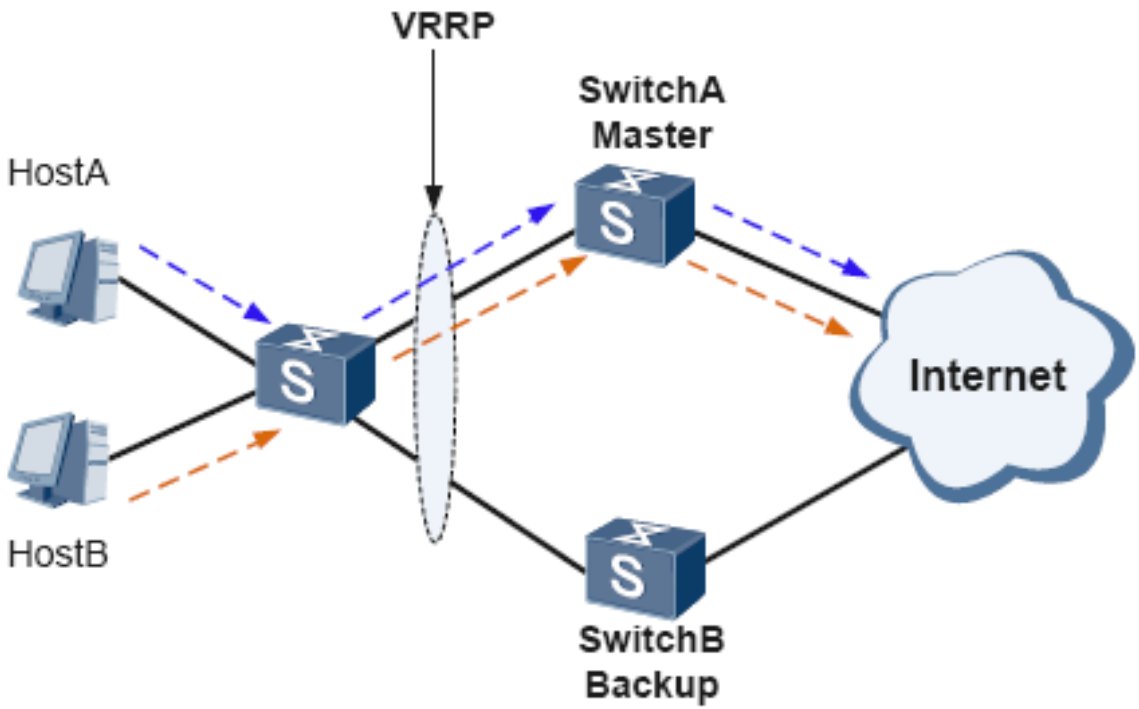
VRRP状态机


INITIALIZE
收到Shutdown消息
收到Startup且优先级是255
收到Startup且优先级小于255
收到Shutdown消息
MASTER
收到一个本地优先级大的报文
BACKUP
Master_Down_Interval超时

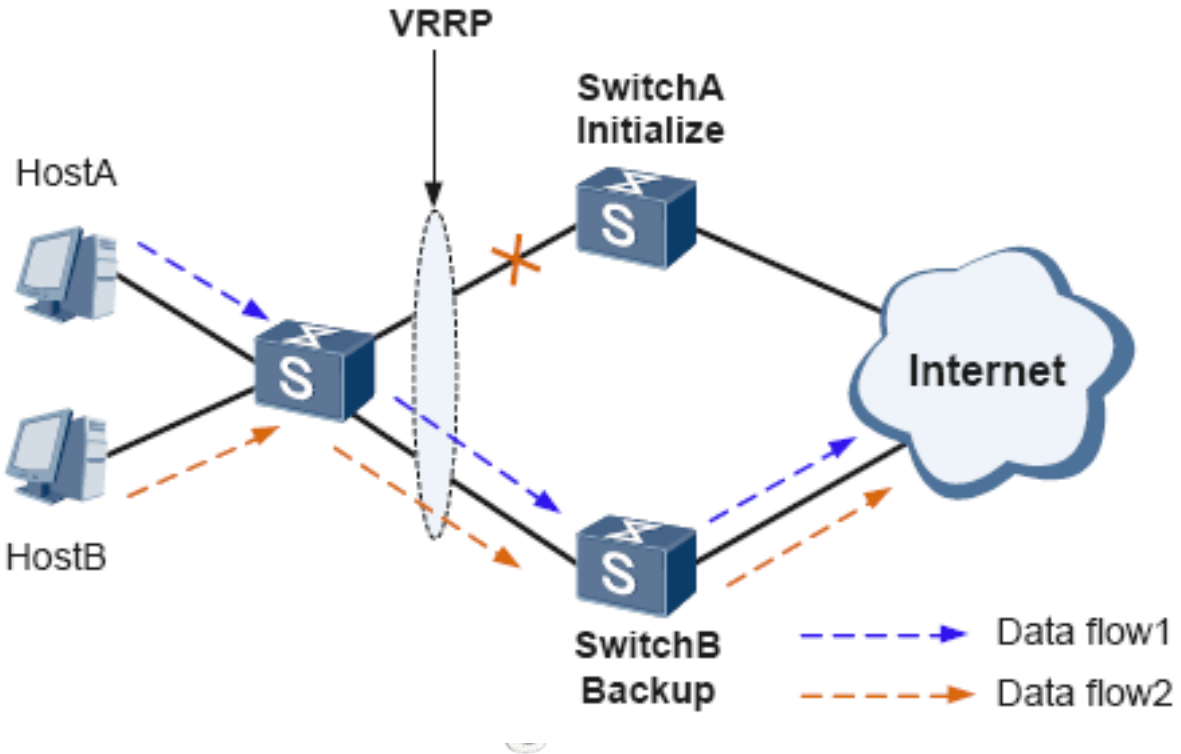
VRRP认证

- 简单字符（Simple）认证：发送VRRP通告报文的交换机将认证方式和认证字填充到通告报文中，而收到通告报文的交换机则会将报文中的认证方式和认证字与本端配置的认证方式和认证字进行匹配。如果相同，则认为接收到的报文是合法的VRRP通告报文；否则认为接收到的报文是一个非法报文，并丢弃这个报文。
- MD5认证：发送VRRP通告报文的交换机利用MD5算法对认证字进行加密，加密后保存在Authentication Data字段中。收到通告报文的交换机会对报文中的认证方式和解密后的认证字进行匹配，检查该报文的合法性。

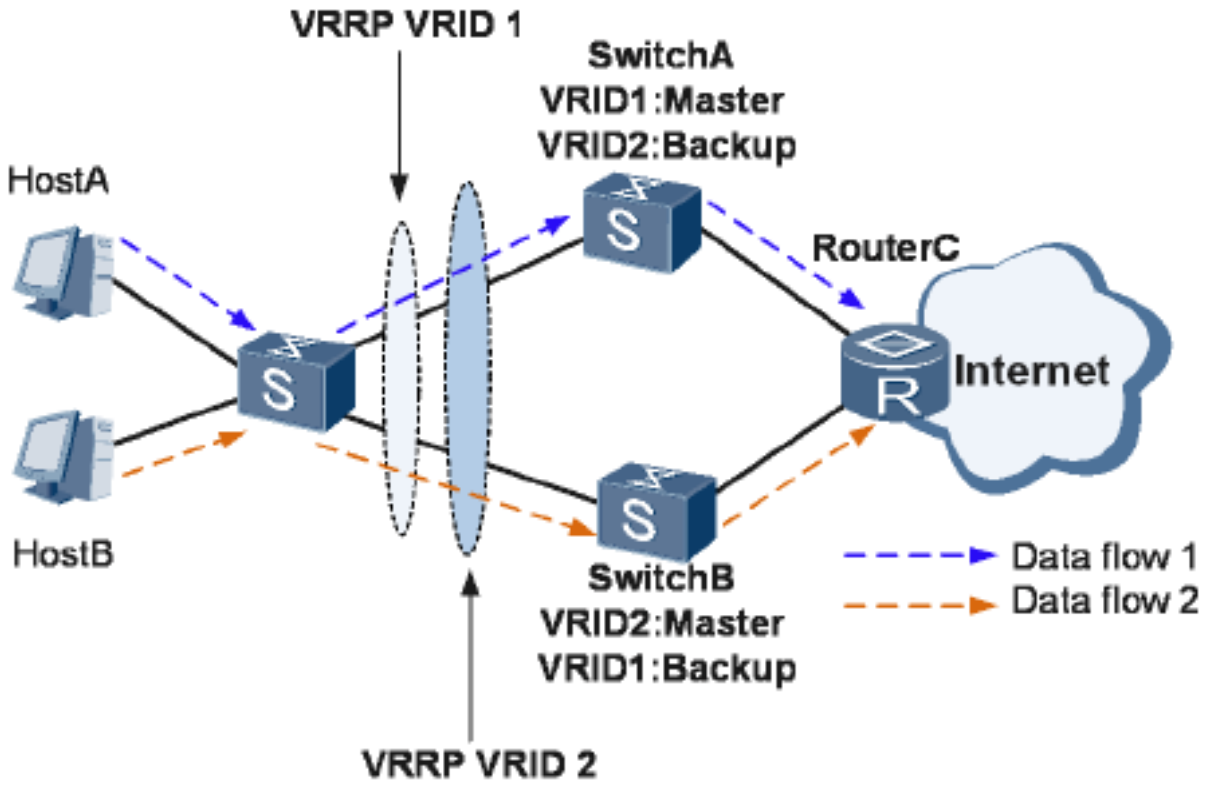
VRRP主备备份



VRRP主备备份(2)


VRRP
SwitchA
Initialize
HostA
Internet
HostB
SwitchB
Backup
Data flow1
Data flow2

VRRP负载分担

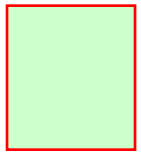

VRRP VRID 1
SwitchA
VRID1:Master
VRID2:Backup
HostA
RouterC
Internet
HostB
SwitchB
VRID2:Master
VRID1:Backup
Data flow 1
Data flow 2
VRRP VRID 2

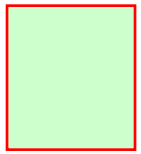
配置keepalived

- 配置文件/etc/keepalived/keepalived.conf
  - GLOBAL CONFIGURATION
    - Global definitions
    - Static routes
  - VRRPD CONFIGURATION
    - VRRP synchronization group(s)
      - String, name of group of IPs that failover together
    - VRRP instance(s)
      - Describes the moveable IP for each instance of a group in vrrp_sync_group
  - LVS CONFIGURATION
    - Virtual server group(s)
    - Virtual server(s)

vrrp_instance

- vrrp_instance VI_NAME {
  - state MASTER|BACKUP
  - interface eth0
  - virtual_router_id 51
  - priority 100
  - authentication {
    - auth_type PASS|AH
    - auth_pass magedu.com
  - }
  - virtual_ipaddress {
    - <IPADDR>/<MASK> brd <IPADDR> dev <STRING> scope <SCOPE> label <LABEL>
  - }
  - virtual_routes {
    - # src <IPADDR> [to] <IPADDR>/<MASK> via|gw <IPADDR> [or <IPADDR>] dev <STRING> scope <SCOPE> tab
  - }
- }





Virtual server(s)

- virtual_server IP port
- virtual_server fwmark int
- virtual_server group string
  - lb_algo rr|wrr|lc|wlc|lblc|sh|dh
  - lb_kind NAT|DR|TUN
  - persistence_timeout <INT>
  - protocol TCP
  - sorry_server <IPADDR> <PORT>

- real_server <IPADDR> <PORT>
- {
  - weight <INT>
  - notify_up <STRING>|<QUOTED-STRING>
  - notify_down <STRING>|<QUOTED-STRING>
  - #HTTP_GET|SSL_GET|TCP_CHECK|SMTP_CHECK|MISC_CHECK
- }

HTTP_GET|SSL_GET

- HTTP_GET|SSL_GET
- {
  - # A url to test, can have multiple entries here
  - url {
    - path <STRING>
    - # healthcheck needs status_code or status_code and digest .Digest computed with genhash, eg digest 9b3a0c85a887a256d6939da88aabd8cd
    - digest <STRING>
    - status_code <INT>
  - }
  - connect_port <PORT>
  - bindto <IPADDR>
  - connect_timeout <INT>
  - nb_get_retry <INT>
  - delay_before_retry <INT>
- }

TCP_CHECK

- TCP_CHECK
- {
  - connect_port <PORT>
  - bindto <IPADDR>
  - connect_timeout <INT>
- }

vrrp_script示例1

- vrrp_script chk_sshd {
  - script "killall -0 sshd" # cheaper than pidof
  - interval 2 # check every 2 seconds
  - weight -4 # default prio: -4 if KO
  - fall 2 # require 2 failures for KO
  - rise 2 # require 2 successes for OK
- }

- vrrp_script chk_http_port {
  - script "</dev/tcp/127.0.0.1/80" # connects and exits
  - interval 1 # check every second
  - weight -2 # default prio: -2 if connect fails
- }

- vrrp_instance VI_1 {
  - …… …… ……
  - track_interface {
    - eth1 weight 2 # prio = +2 if UP
    - eth2 weight -2 # prio = -2 if DOWN
    - eth3 # no weight, fault if down
  - }
  - track_script {
    - chk_sshd # use default weight from the script
    - chk_haproxy weight 2 # +2 if process is present
    - chk_http_port
  - }
- }

vrrp示例2

- vrrp_script chk_haproxy {
  - script "killall -0 haproxy"
  - interval 1
  - weight 2
- }

- vrrp_script chk_mantaince_down {
  - script "[[-f /etc/keepalived/down]] && exit 1 || exit 0"
  - interval 1
  - weight 2
- }

- vrrp_instance VI_1 {
- ……
  - track_interface {
    - eth0
  - }
  - virtual_ipaddress {
    - 172.16.100.1/16 dev eth0 label eth0:0
  - }
  - track_script {
    - chk_haproxy
    - chk_mantaince_down
  - }
  - notify_master "/etc/keepalived/notify.sh master"
  - notify_backup "/etc/keepalived/notify.sh backup"
  - notify_fault "/etc/keepalived/notify.sh fault"
- }

关于马哥教育

- 博客：http://mageedu.blog.51cto.com
- 主页：http://www.magedu.com
- QQ：1661815153，113228115
- QQ群：203585050，279599283


THANK
YOU