FortiGate HA 原理及配置

版本	1.0
时间	2011 年 12 月
作者	胡丹丹(ddhu@fortinet.com)
支持的版本	FortiOS v3.x,v4.x
状态	草稿

目录

1.目的

HA(High Availability)指的是通过尽量缩短或完全避免因日常维护操作(计划)和突发的系统崩溃(非计划)所导致的停机,以提高系统和应用的可用性。

HA 也提供负载均衡在 HA 成员中分配会话及流量实现平衡系统负载以及设备性能的最大化。

FortiGate 支持 3 种方式的 HA:

FortiGate Cluster Protocol (FGCP):FortiGate 专有集群协议;

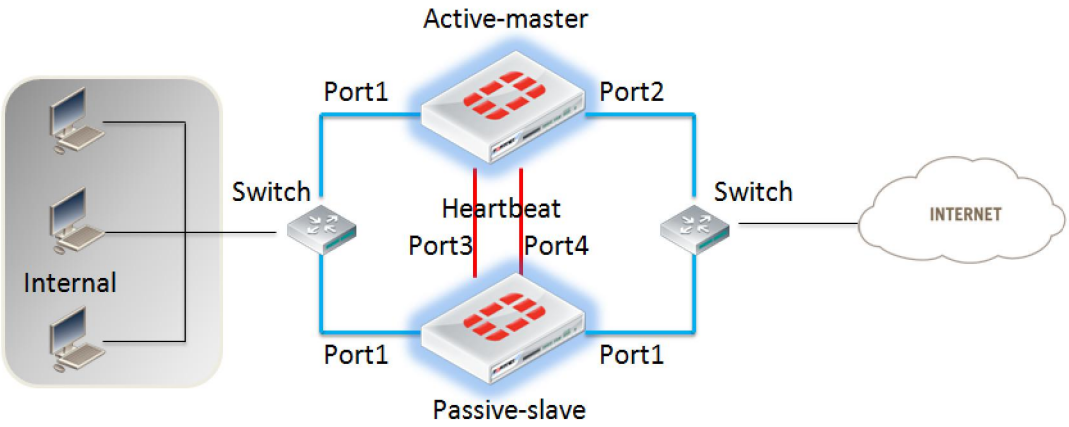
Virtual Router Redundancy Protocol(VRRP):虚拟路由器冗余协议;

TCP Session synchronization:TCP 的会话同步。

当系统出现各种状况时,设备将如何运作？流量及会话由哪台设备处理？什么类型的会话将失效？完全了解 HA 原理,才能在计划或非计划故障中保证系统在预期状态下运行。本文就 FortiGate 常见的 FGCP 模式 HA 进行说明。

2.环境介绍

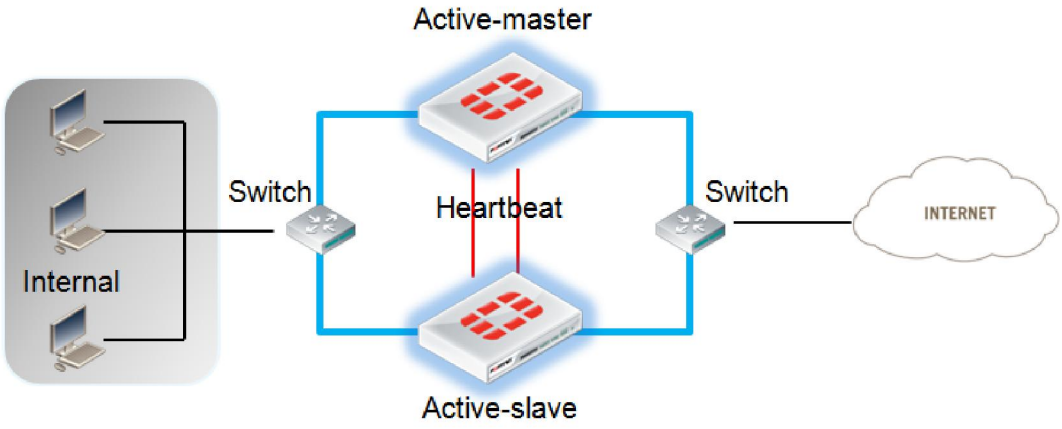
本文使用 2 台 FortiGate 310B 进行说明,本文使用的系统版本为 FortiOS v4.0MR2 Patch8。



3.HA 的工作模式

3.1 主动-主动模式

主动-主动模式,以下称为主主模式(AA),HA 成员都处于工作状态下,任意设备故障后,其他设备仍然正常工作。

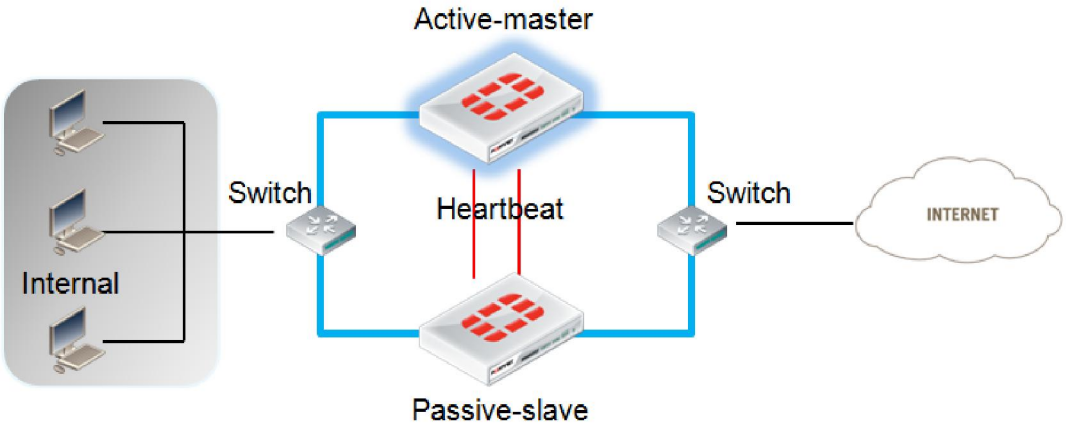


AA 模式下在 HA 状态中查看到 HA 的角色，有主设备及从设备,通常会被认为工作在主被模式下,实际上主主下设备虽然都在工作,仍会有一台作为集群的主设备用来控制和分配流量和会话给集群中的其他设备。

AA 模式默认情况下仅负载均衡 UTM 的流量,所以在下不使用 UTM 功能时建议使用 AP 模式。

3.2 主动-被动模式

主被模式(AP),HA 主设备处于工作状态下,从设备处于备份状态,主设备故障后,从设备切换至主设备继续工作。



4.主设备的选举

4.1 常规模式下的主设备选举

HA 组建开始时,HA 成员将基于以下条件自动协商选举主设备,当主设备选举完成后,其他设备成为从设备。

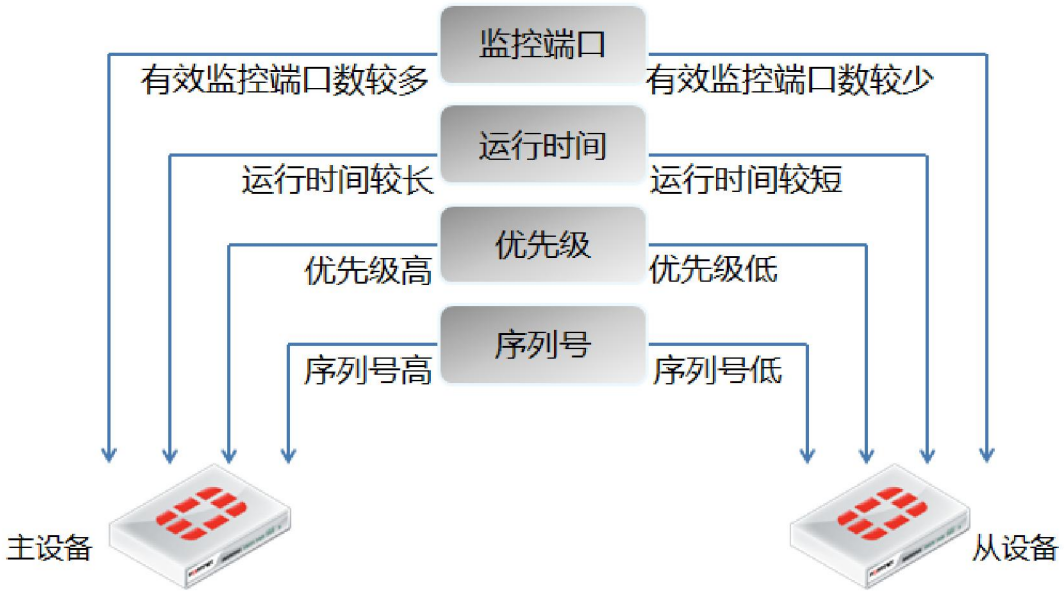
选举顺序从步骤 1 至步骤 4:

1.监控端口:有效监控端口数相等,考虑下一条件;

2.运行时间:运行时间相等,考虑下一条件,运行时间差在 5 分钟以内不考虑在内;

3.设备优先级:设备优先级相同,考虑下一条件;

4.设备序列号:如果以上条件都相同,则设备序列号最大的将被选举为主设备。



HA 运行时间(Age time):当设备启动或从监控端口检测到链路失效或端口失效,Age time 被重置。如果监控端口检测到链路失效,那么该设备的 age time 重置,将会比其他设备的 age time 较小,也就不能在新的设备选举过程中胜出。

在所有的 HA 环境中,HA 成员可能无法同时启动,FGCP 会忽略 5 分钟之内的时间差,以保证此时间差不影响主设备的选举。

查看 HA 运行时间差

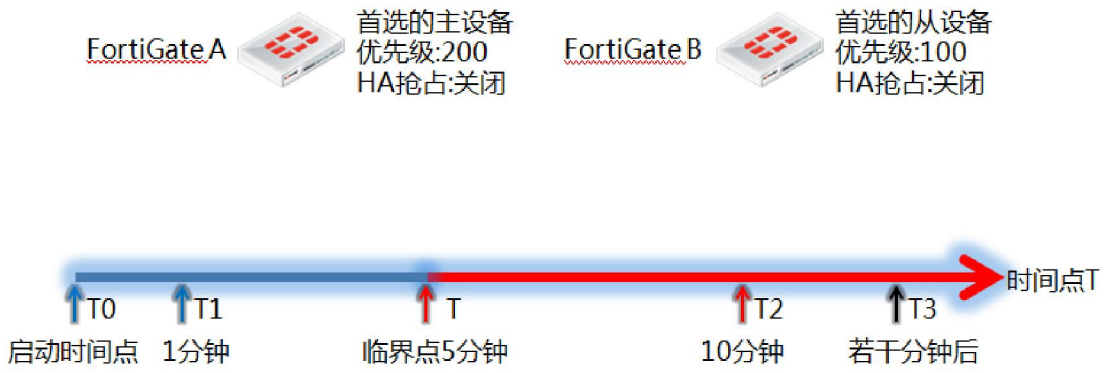
ha-b-118 # di sys ha dump 1
HA information.
vcluster id=1, nventry=2, state=work, digest=2.ec.cf.4e.a0.5a...
ventry
idx=0,id=1,FG300B3909600118,prio=130,-100,claimed=0,override=0,flag=1,time=0,mon=0 mondev=port7,50port5,50port2,50port1,50
ventry
idx=1,id=1,FG300B3908600981,prio=120,-100,claimed=0,override=0,flag=0,**time=252**,mon=0

ha-b # exe ha manage 1
ha-a-981 $ di sys ha dump 1
HA information.
vcluster id=1, nventry=2, state=work, digest=2.ec.cf.4e.a0.5a...
ventry
idx=1,id=1,FG300B3908600981,prio=120,-100,claimed=0,override=0,flag=1,time=0,mon=0

mondev=port7,50port5,50port2,50port1,50
ventry
idx=0,id=1,FG300B3909600118,prio=130,-100,claimed=1,override=0,flag=0,**time=-252**,mon=0

其中 time 即为 HA 的时间差,此处为 ha-a-981 晚于设备 ha-b-118 启动时间 25.2 秒,即 HA 的时间差。

HA 的时间差作为主设备选举的第二条件,对于 HA 的主从关系是非常重要的因素,甚至优先于设备的优先级,以下为关于时间差的具体描述示例:



1.设备启动时: A 和 B 启动,此时由于两者运行时间差小于 5 分钟,HA 选举 A 作为主设备,因为 A 的优先级高于 B;

2.时间点 1 分钟 T1: A 设备监控接口失效,B 被选举成为主设备,在 A 监控接口恢复后,A 重新作为主设备加入 HA,因为 A 的优先级高于 B,且 HA 时间差仍小于 5 分钟;

3.时间点 10 分钟 T2: A 设备监控接口失效,B 被选举成为主设备,在 A 监控接口恢复后,A 重新作为从设备加入 HA,因为 A 的优先级虽然高于 B, HA 时间差大于 5 分钟;

diag sys ha reset-uptime 可以重置 HA 运行时间,以期系统通过预期的优先级设定重新将高优先级选举为主设备。

4.2 抢占模式下的主设备选举

如果希望某台设备一直作为主设备工作,除了给该设备设置较高的优先级以外,也可以启用 HA Override(抢占),这样,即便设备失效,再恢复之后,将无视 HA 的运行时间因素,仍能当选为主设备。

Override 开启情况下,如果主设备重启,从设备会成为主设备,原主设备重启恢复后,将重新通过选举,仍工作在主设备模式下。HA override 开启时,如果在从设备选举为主设备这段时间内更改配置,那么在原主设备恢复后,这段时间的配置将被原配置同步,导致丢失。

5.HA 的虚拟 MAC 地址

5.1 虚拟 mac 地址的组成及排序

HA 组建时,FGCP 会指定给主设备各个接口虚拟 MAC 地址,当主设备失效切换时,从设备将获得同样的虚拟 MAC 地址。虚拟 MAC 使用以下格式:

00-09-0f-09-<group-id_hex>-<vcluster_integer><idx>

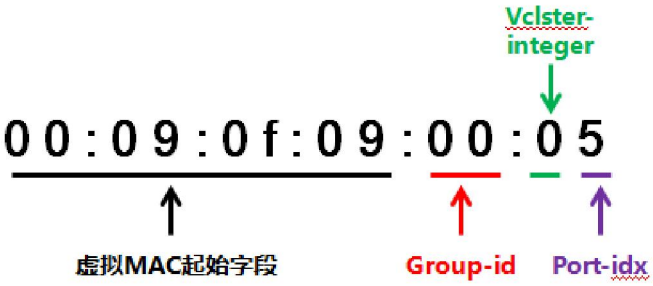
通过命令查看虚拟 MAC 地址

ha-b-118 # get hardware nic port5

...

MAC: 00:09:0f:09:00:05 **#虚拟 MAC 地址**

Permanent_HWaddr: 00:09:0f:89:29:29 #真实 MAC 地址



group-id: 0-63,默认为 0

Vcluster_integer: 0,2　　vcluster 1 为 0,vcluster 2 为 2;如果未启用 vdom,系统将给 root 域使用 vcluster1,即 vcluster_interger 为 0。

Idx:端口号,以以下顺序排列

- port1 virtual MAC: 00-09-0f-09-00-00
- port10 virtual MAC: 00-09-0f-09-00-01
- port2 virtual MAC: 00-09-0f-09-00-02
- port3 virtual MAC: 00-09-0f-09-00-03
- port4 virtual MAC: 00-09-0f-09-00-04
- port5 virtual MAC: 00-09-0f-09-00-05
- port6 virtual MAC: 00-09-0f-09-00-06
- port7 virtual MAC: 00-09-0f-09-00-07
- port8 virtual MAC: 00-09-0f-09-00-08
- port9 virtual MAC: 00-09-0f-09-00-09
- port11 virtual MAC: 00-09-0f-09-00-0a
- port12 virtual MAC: 00-09-0f-09-00-0b

5.2 如何避免虚拟 mac 地址冲突

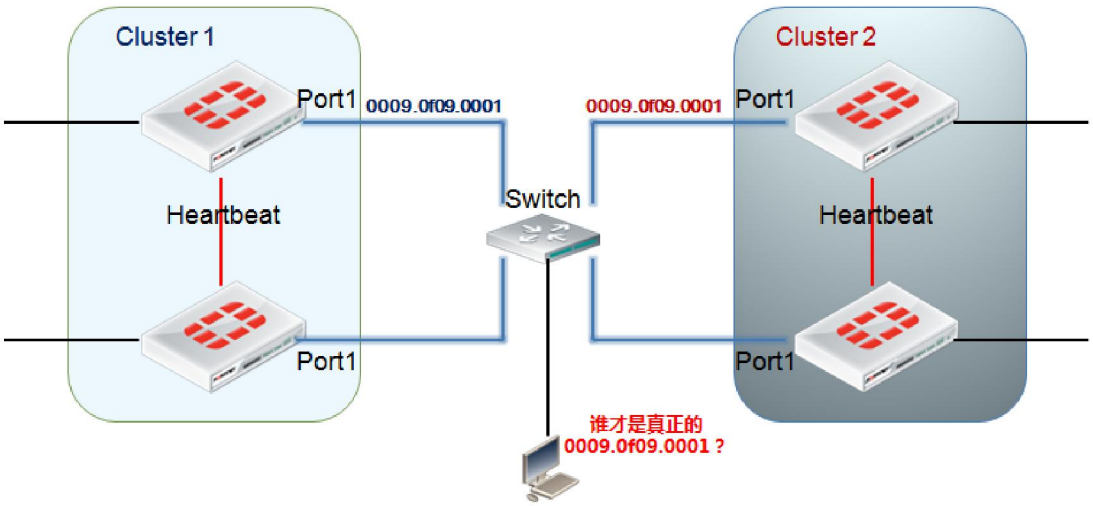
当一个广播域内存在两个或多个 HA 集群时,需要使用 HA 的 group-id 来标识各个 HA,否则如果使用相同的 HA group-id,FGCP 将分配给不同 HA 主设备相同的虚拟 MAC 地址,导致 MAC 地址冲突。

如果两个 Cluster 的虚拟 mac 地址都被分配为 0009.0f09.0001,交换机将出现 2 个相同的 MAC 地址分别连在不同的接口上。

```
1      0009.0f09.0001      DYNAMIC      Gi1/0/1
1      0009.0f09.0001      DYNAMIC      Gi1/0/4
```

出现这种情况后,分别 ping 两个 cluster 的 port1 地址将会出现下面的情况

```
Cluster_1   Cluster_2
reply       timeout
reply       timeout
reply       timeout
timeout     reply
timeout     reply
reply       timeout
reply       timeout
timeout     reply
timeout     reply
```



通过命令行修改 HA group-id

```
config sys ha
    set group-id 12
end
```

6.HA 的配置

6.1HA 配置之前的准备

HA 配置之前需要确保以下情况:

1.硬件型号,软件版本必须一致;

系统软件版本查看,下线标注为版本号,**确保软件的版本,软件版本时间一致。**

```
GateA # get sys status
Version: Fortigate-310B v4.0,build0328,110718 (MR2 Patch 8)
Virus-DB: 11.00782(2010-05-07 00:42)
Extended DB: 1.00001(2010-05-21 13:37)
IPS-DB: 3.00032(2011-07-14 13:11)
FortiClient application signature package: 1.439(2011-11-15 16:59)
Serial-Number: FG300B3909600118
BIOS version: 04000011
Log hard disk: Available
Hostname: GateA
Operation Mode: NAT
Current virtual domain: root
Max number of virtual domains: 10
Virtual domains status: 1 in NAT mode, 0 in TP mode
Virtual domain configuration: disable
FIPS-CC mode: disable
Current HA mode: a-p, master
Distribution: International
Branch point: 328
Release Version Information: MR2 Patch 8
System time: Tue Nov 15 17:10:50 2011
```

2.如果设备包含硬盘,则集群中所有设备硬盘必须在大小,格式及分区保证一致;

通过 get sys status 查看硬盘的具体状态:

```
GateA # get sys status
Version: Fortigate-310B v4.0,build0328,110718 (MR2 Patch 8)
Virus-DB: 14.00000(2011-08-24 17:17)
Extended DB: 14.00000(2011-08-24 17:09)
IPS-DB: 3.00032(2011-07-14 13:11)
FortiClient application signature package: 1.444(2011-12-02 00:36)
```

Serial-Number: FG300B3908600981
BIOS version: 04000011
Log hard disk: Not available
Hostname: FG300B3908600981
Operation Mode: NAT
Current virtual domain: root
Max number of virtual domains: 10
Virtual domains status: 1 in NAT mode, 0 in TP mode
Virtual domain configuration: disable
FIPS-CC mode: disable
Current HA mode: standalone
Distribution: International
Branch point: 328
Release Version Information: MR2 Patch 8
System time: Sun Dec 4 21:51:10 2011

Not available 表示磁盘不存在或系统未能识别该磁盘。

Need format 表示磁盘需要格式化,可以使用 execute formatlogdisk 用于格式化硬盘。

如果分区不一致可以通过以下命令来修正

execute scsi-dev delete
execute formatlogdisk

3.HA 不支持作为 PPP 或者 DHCP 客户端,所以当接口启动 PPPOE 或者 DHCP 时,将无法启用 HA 模式。但是 HA 可以作为 DHCP 服务器分配 IP 地址

6.2HA 配置

进入系统管理-配置-高可靠性

1.选择模式;

2.设定设备优先级;

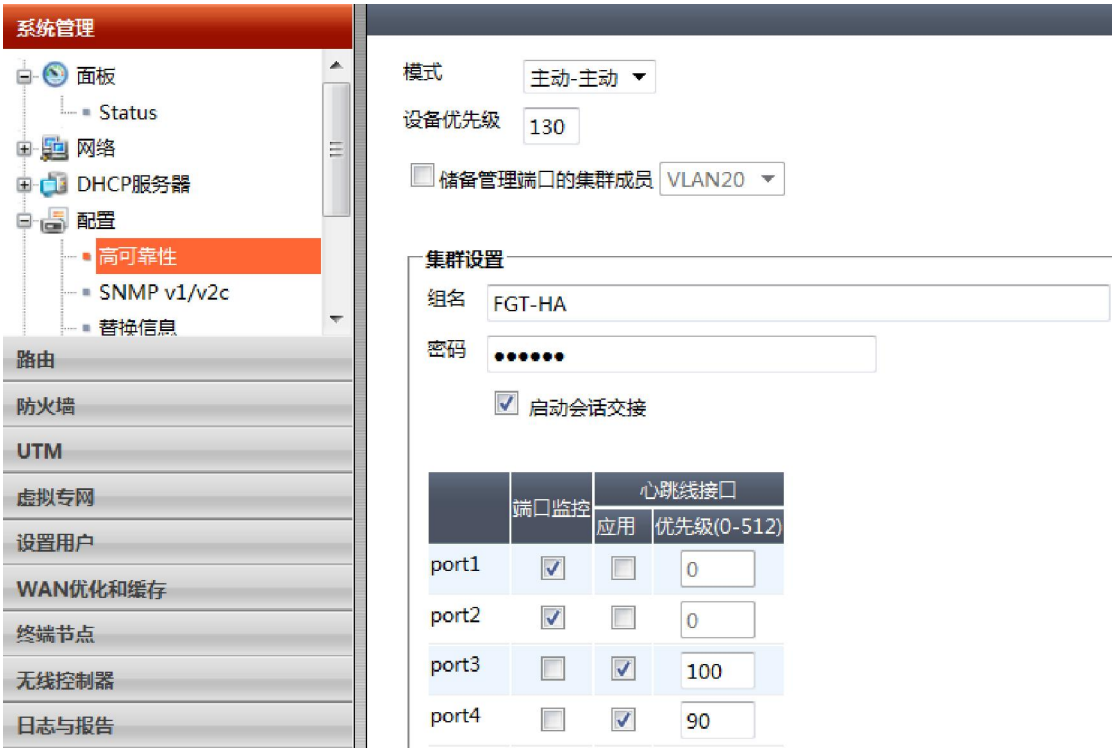
3.选择管理接口,方便管理从设备使用;

4.集群设置,及是否启用会话交接;

5.选择端口监控,以及心跳线接口,及心跳线接口的优先级,可以使用一个或多个物理接口用于心跳接口(hbdev),若同时存在多个心跳接口,以优先级最高的作为同步接口,其他作为备份心跳接口。2 台进行 HA 时,将心跳线将设备直接连接即可,集群中的多台设备可以划分 VLAN 连接至交换机用于会话同步及心跳。

4.3 版本新特性允许用户指定一个或多个接口用于会话同步和 session pickup(多个接口将负载均衡会话同步)。会话同步包使用 Ethertype 0x8892,指定会话同步口失效情况下,会话同步仍会交由心跳口处理

```
config system ha
        set session-sync-dev port10 port12
end
```



HA 集群的其他设备也可以据此进行相应设置。

7.故障恢复

HA Failover 故障恢复是一种备份机制用于降低系统风险及减小不可预期的

宕机时间,尤其在关键业务的生产环境中,HA Failover 至关重要。

FGCP 提供以下三种故障恢复机制:

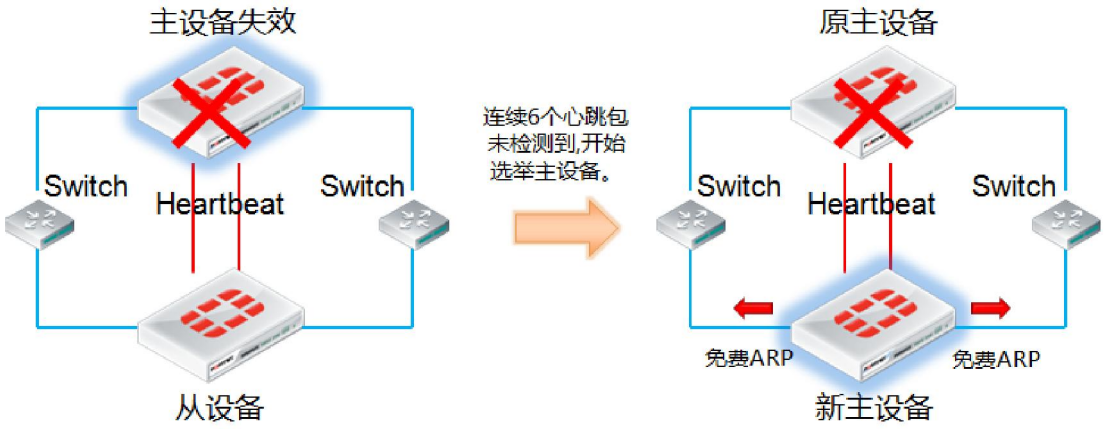
1.设备故障恢复

2.端口故障恢复

3.会话交接

7.1 设备故障恢复

一旦主设备遭遇故障,集群中的其他设备将协商选举新的主设备,拥有同样的 IP 和 MAC 地址,并在所有接口发送免费 ARP。



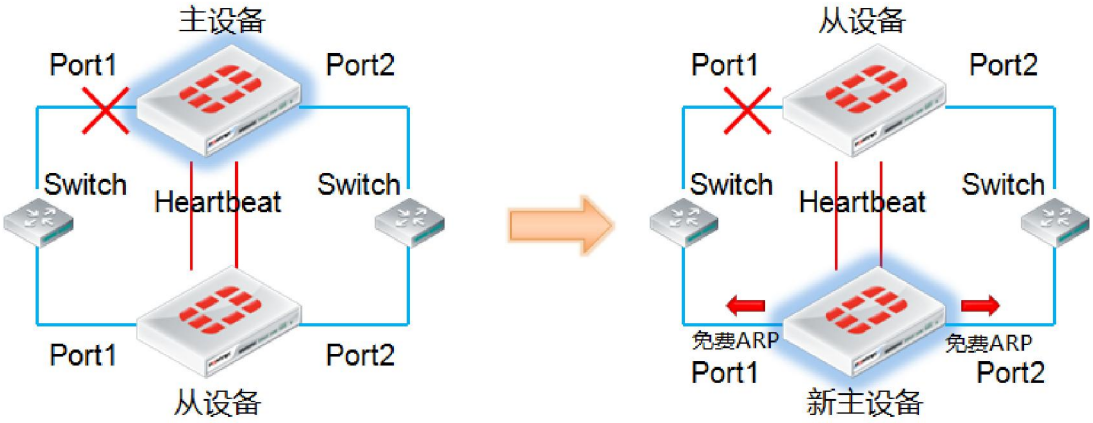
默认心跳丢失阀值为 6 个,每次心跳间隔 200 ms;

```
config system ha
  set hb-interval 2              #心跳 hello 包间隔,200ms
  set hb-lost-threshold 6        #心跳丢失阀值,6 个
end
```

建议在配置心跳线时,使用 2 个心跳接口冗余使用。

7.2 端口故障恢复

设备遭遇端口故障或失去连接,集群将重新选举,但由于原主设备仍在工作,故也将参加选举,由于有效监控端口减少且 HA 运行时间被重置,将在选举中失败,以从设备身份加入集群。



7.3 会话交接

Session Failover 会话故障恢复,在正常 HA 状态下,如果开启 Session pick-up(即会话交接),从设备报告自己的状态并接收存储会话连接与状态表更新。

一旦 HA 出现故障,集群中的从设备被选举为主设备,根据自己同步的会话连接与状态表,继续处理流量与会话。

命令行下开启会话交接

```
config system ha
    set session-pickup enable
end
```

会话交接支持和不支持的类型

会话交接支持 TCP 和 IPsec VPN 会话;

不支持 multicast,ICMP,SSL VPN 会话;

不支持 UDP 会话;

不支持被特定 UTM 处理的会话,如 AV,Web 过滤,反垃圾邮件过滤,内容归档;

支持被 IPS 处理的会话

8.HA 状态

HA 的状态可以通过 web 界面及命令行 2 种方式查看。

ha-b-118 # get sys ha status

Model: 300 #HA 设备型号

Mode: a-a #HA 工作模式

Group: 14 #HA 组 id

Debug: 0

ses_pickup: enable #是否启用会话交接

load_balance: enable #负载均衡启用

schedule: round robin #负载均衡方式

Master:130 ha-b-118 FG300B3909600118 0 #当前主设备

Slave :120 ha-a-981 FG300B3908600981 1 #当前从设备

number of vcluster: 1 #虚拟集群数

vcluster 1: work 169.254.0.1

Master:0 FG300B3909600118

Slave :1 FG300B3908600981

9.HA 下的软件更新

升级 HA 的系统同单机模式的操作方法一致,但是中间的过程却是不一样,此过程对用户及网络都是透明的,期间网络通讯并不会中断。具体步骤如下

1.通过 Web 界面上传新的 Firmware 软件版本;

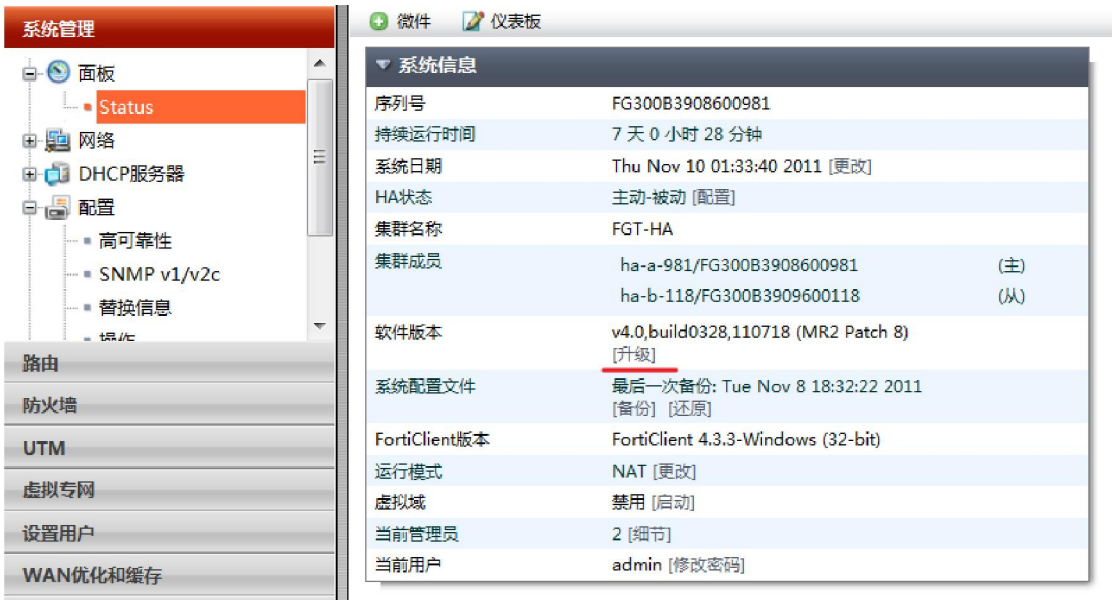
2.如果系统工作在 AA 模式下,升级时系统将关闭负载均衡;

3.系统首先升级所有的从设备;

4.一旦从设备升级完毕,将选举新的主设备,主设备将运行新的 Firmware 版本;

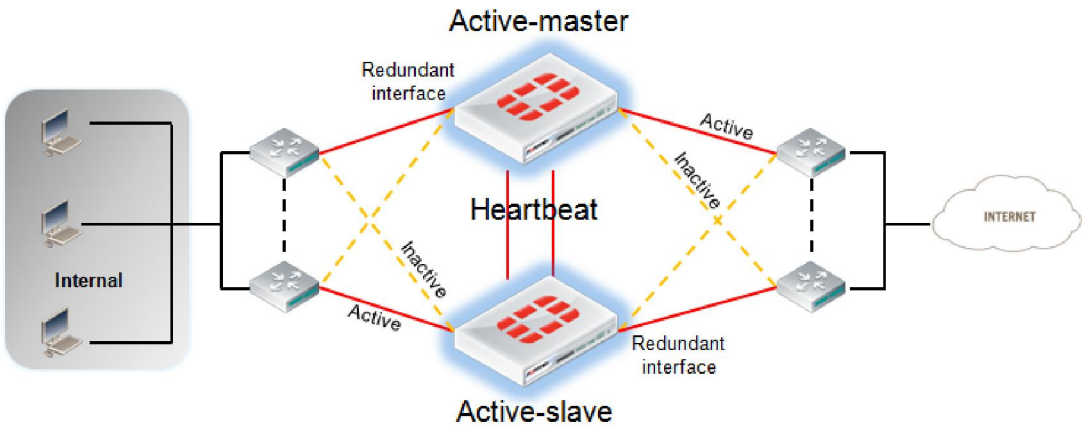
5.系统最后升级原主设备。

6.如果系统工作在 AA 模式下,升级后将开启负载均衡。



10.Full Mesh 结构 HA

HA Full mesh 是为了解决交换机出现单点或者与交换机相连的链路出现的故障提出的解决方案



Full mesh 需要冗余接口或聚会链路支持,故只有中高档以上型号支持该特性,在每台 FortiGate 上使用 redundant 接口,分别将主接口及冗余接口分别接至 2 台交换机。那么任意交换机或者任意线路失效,备用交换机及备用链路将生效,冗余接口的设置方法请参考 [FortiGate 冗余接口]

11.参考

High Availability (HA)

Technical Note : Restoring HA master role after a failover using "diag ha reset uptime"

Technical Note: FortiGate HA A-A TCP Packet Flow when a Protection Profile is enabled

HA Cluster virtual MAC addresses

Configuration changes lost when HA override enabled

FortiGate HA synchronization

Updating MAC forwarding tables when an HA link failover occurs