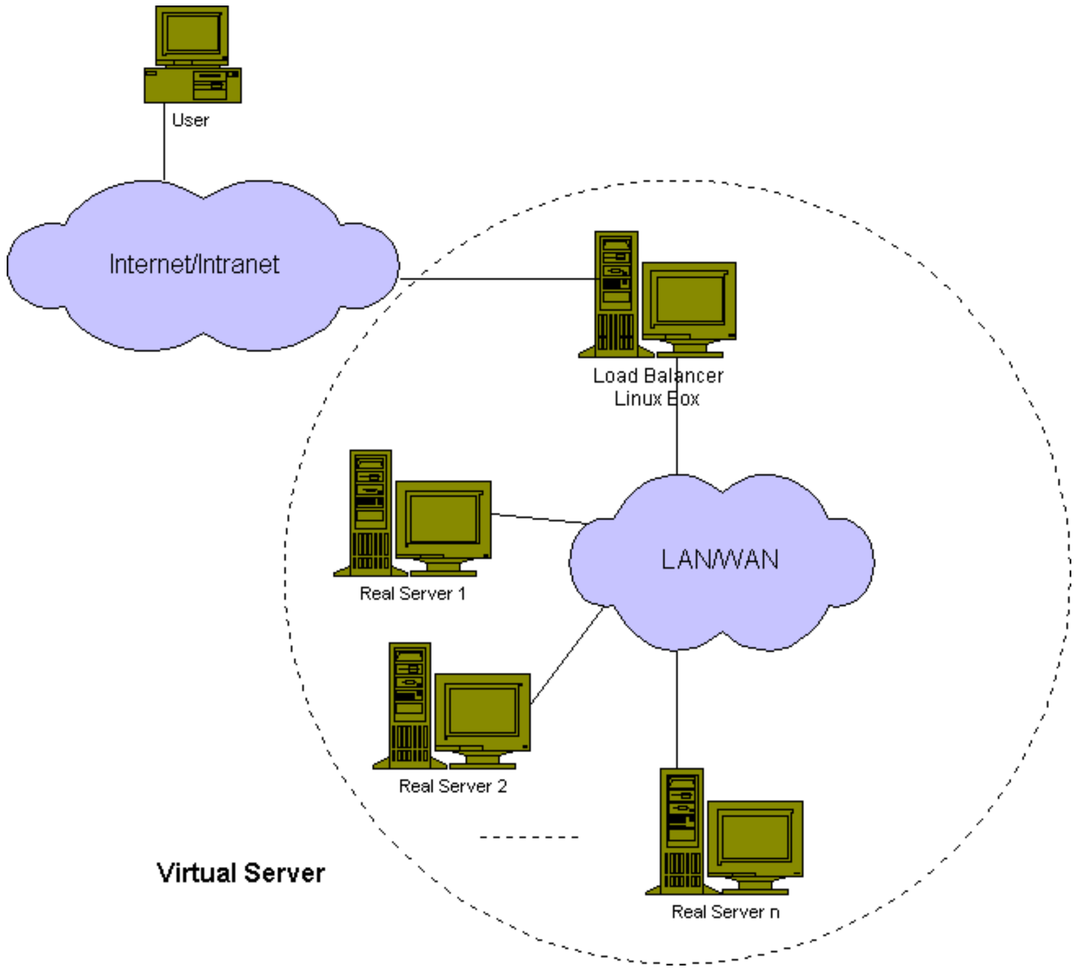
A Pure Go LVS Controller

HUANG HUA(黄 华) 2021.08

What's LVS ?

http://linuxvirtualserver.org/whatis.html



from http://linuxvirtualserver.org/whatis.html

Solve Problems?

Scalability、High Availability

How to ?

Three ways: DR、TUN、NAT

★ ip_vs

[vagrant@lvs ~]$ **modinfo ip_vs**

```
filename:       /lib/modules/3.10.0-327.4.5.el7.x86_64/kernel/net/netfilter/ipvs/ip_vs.ko
license:        GPL
rhelversion:    7.2
srcversion:     E06AC544DA352A4EDFBC73D
depends:        nf_conntrack,libcrc32c
intree:         Y
vermagic:       3.10.0-327.4.5.el7.x86_64 SMP mod_unload modversions
signer:         CentOS Linux kernel signing key
sig_key:        10:5D:A1:3D:CA:AA:74:AE:50:00:17:E7:D5:2C:DA:9B:7C:C5:10:93
sig_hashalgo:   sha256
parm:           conn_tab_bits:Set connections' hash size (int)
```

★ ipvsadm

[vagrant@lvs ~]$ **sudo ipvsadm -v**

ipvsadm v1.27 2008/5/15 (compiled with popt and IPVS v1.2.1)

repo https://git.kernel.org/pub/scm/utils/kernel/ipvsadm/ipvsadm.git

```
ipvsadm -A -t 207.175.44.110:80 -s rr
ipvsadm -a -t 207.175.44.110:80 -r 192.168.10.1:80 -m
ipvsadm -a -t 207.175.44.110:80 -r 192.168.10.2:80 -m
ipvsadm -a -t 207.175.44.110:80 -r 192.168.10.3:80 -m
```

[vagrant@lvs ~]$ **sudo ipvsadm -ln**

```
IP Virtual Server version 1.2.1 (size=4096)   <—— always found it, why??
Prot LocalAddress:Port Scheduler Flags
  -> RemoteAddress:Port           Forward Weight ActiveConn InActConn
TCP  207.175.44.110:80 rr
  -> 192.168.10.1:80              Masq    1      0          0
  -> 192.168.10.2:80              Masq    1      0          0
  -> 192.168.10.3:80              Masq    1      0          0
```

How they communicate?

How to communicate with kernel module?

▲ procfs

- vfs mapping kernel’s memory
- report kernel’s state to user space
- kernel-user space half -duplex communication mode

[vagrant@lvs ~]$ **cat /proc/net/ip_vs**

```
IP Virtual Server version 1.2.1 (size=4096)
Prot LocalAddress:Port Scheduler Flags
  -> RemoteAddress:Port Forward Weight ActiveConn InActConn
TCP  CFAF2C6E:0050 rr
  -> C0A80A05:0050      Masq    1      0          0
  -> C0A80A04:0050      Masq    1      0          0
  -> C0A80A03:0050      Masq    1      0          0
```

[vagrant@lvs ~]$ **sudo ipvsadm -ln**

```
IP Virtual Server version 1.2.1 (size=4096)
Prot LocalAddress:Port Scheduler Flags
  -> RemoteAddress:Port           Forward Weight ActiveConn InActConn
TCP  207.175.44.110:80 rr
  -> 192.168.10.1:80              Masq    1      0          0
  -> 192.168.10.2:80              Masq    1      0          0
  -> 192.168.10.3:80              Masq    1      0          0
```

```
[vagrant@lvs ~]$ cat /proc/sys/net/ipv4/ip_forward
0
[vagrant@lvs ~]$ sudo sysctl -w net.ipv4.ip_forward=1
net.ipv4.ip_forward = 1
[vagrant@lvs ~]$ cat /proc/sys/net/ipv4/ip_forward
1
[vagrant@lvs ~]$
```

▲ ioctl

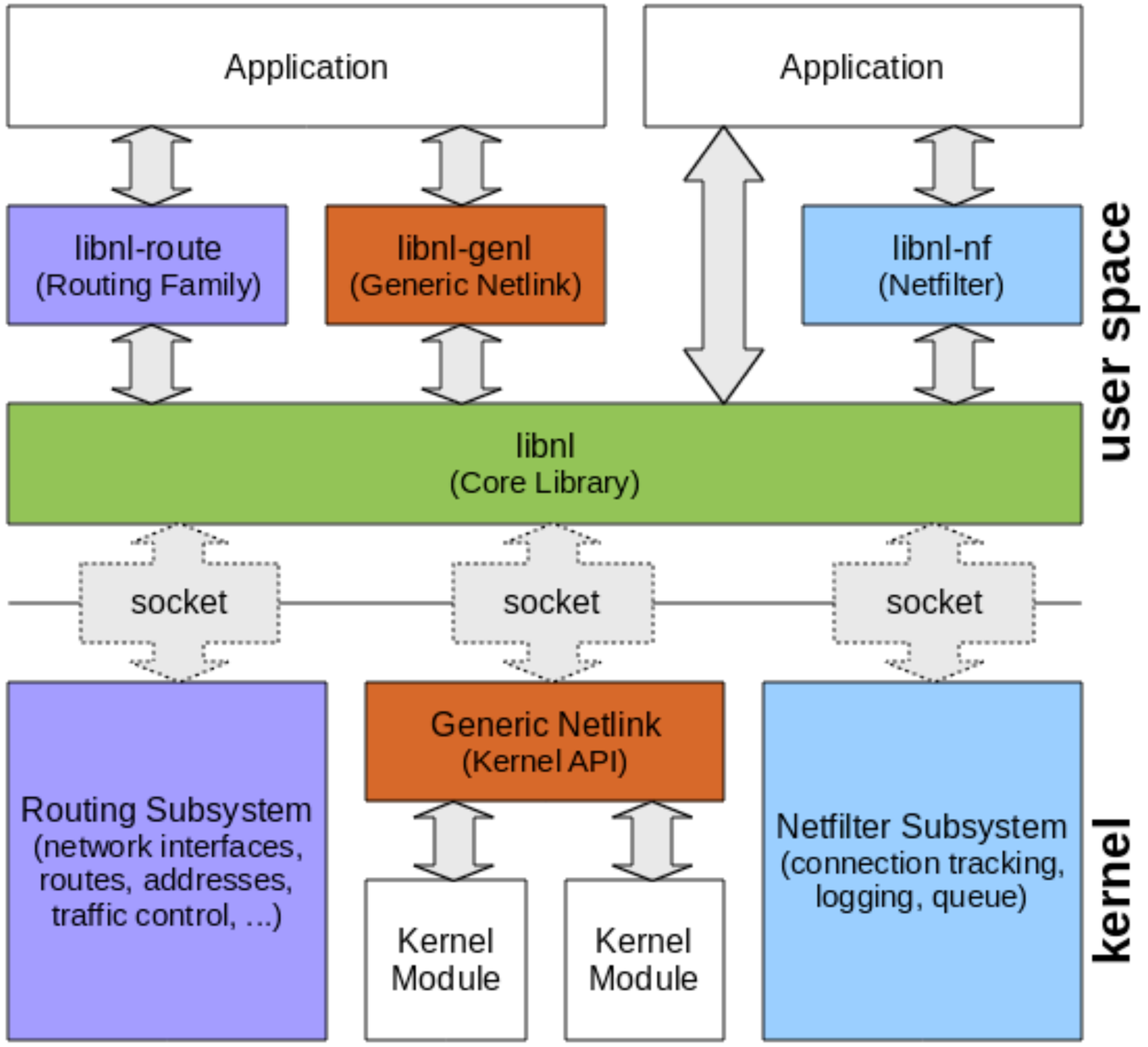
- control device
- system call
- operations like: open read write seek etc.

▲netlink

- transfer information between kernel and user-space processes
- standard sockets-based interface
- internel kernel API for kernel modules
- full -duplex communication mode
- https://tools.ietf.org/html/rfc3549

Netlink Protocol Library Suite(libnl)

- a IPC mechanism
- mainly networking related kernel configuration and monitoring interfaces



from http://www.infradead.org/~tgr/libnl/

ipvsadm communicate with ip_vs

```
➜  ipvsadm-1.27 tree libipvs
libipvs
├── Makefile
├── ip_vs.h
├── ip_vs_nl_policy.c
├── libipvs.c
└── libipvs.h
```

repo https://git.kernel.org/pub/scm/utils/kernel/ipvsadm/ipvsadm.git

```
int ipvs_init(void)
{
	socklen_t len;

	ipvs_func = ipvs_init;

#ifdef LIBIPVS_USE_NL
	try_nl = 1;

	if (ipvs_nl_send_message(NULL, NULL, NULL) == 0) {
		try_nl = 1;
		return ipvs_getinfo();
	}

	try_nl = 0;
#endif

	len = sizeof(ipvs_info);
	if ((sockfd = socket(AF_INET, SOCK_RAW, IPPROTO_RAW)) == -1)
		return -1;

	if (getsockopt(sockfd, IPPROTO_IP, IP_VS_SO_GET_INFO,
		       (char *)&ipvs_info, &len))
		return -1;

	return 0;
}
```

go communicate with ip_vs by netlink

repo1 https://github.com/google/seesaw/ipvs

repo2 https://github.com/moby/ipvs

go communicate with ip_vs by netlink

demo https://github.com/kwanhur/ipvsctl


NETLINK MSG
0
1
2
3
CMD ID
VER
RESERVED
==> General Message Header represented by genlMsgHdr
ATTR LEN
ATTR TYPE
VALUE
[]byte Array of IPVS MSG
PADDED BY 4 BYTES
==> Attribute Message represented by syscall.NetlinkRouteAttr

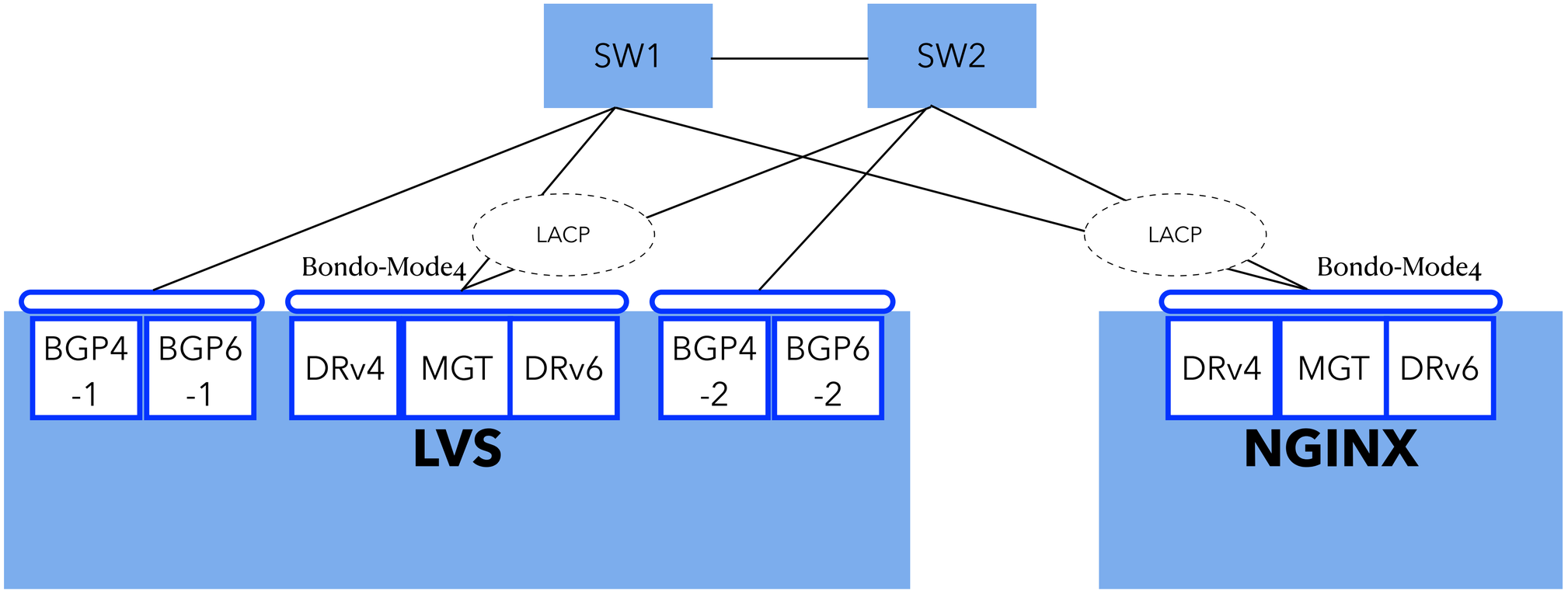

IPVS MSG
0
1
2
3
ATTR LEN
ATTR TYPE
[]byte IPVS ATTRIBUTE BY 4 BYTES
NEXT ATTRIBUTE
ATTR LEN
ATTR TYPE
[]byte IPVS ATTRIBUTE BY 4 BYTES
NEXT ATTRIBUTE
ATTR LEN
ATTR TYPE
[]byte IPVS ATTRIBUTE BY 4 BYTES

LVS in PACloud

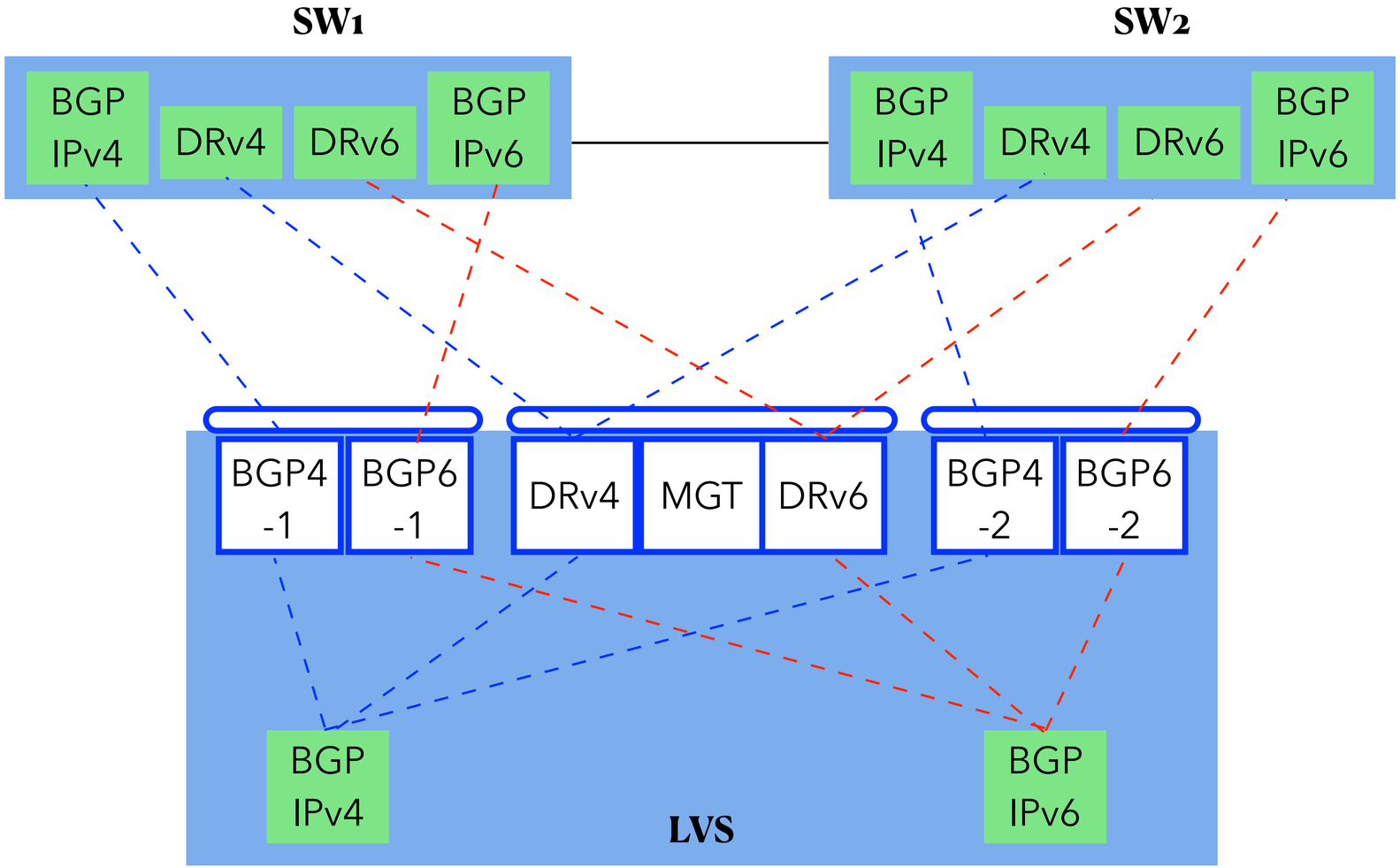
ELB Product Matrix

	LVS HA Mode	Core Capability	Networking Mode	IPv6
Public	ECMP	L4/L7	internet access (dr)	√
Private	Master-Backup	L4/L7	inner VPC (dr)	×
Partner	Master-Backup	L4	intranet tranparent (fullnat)	√

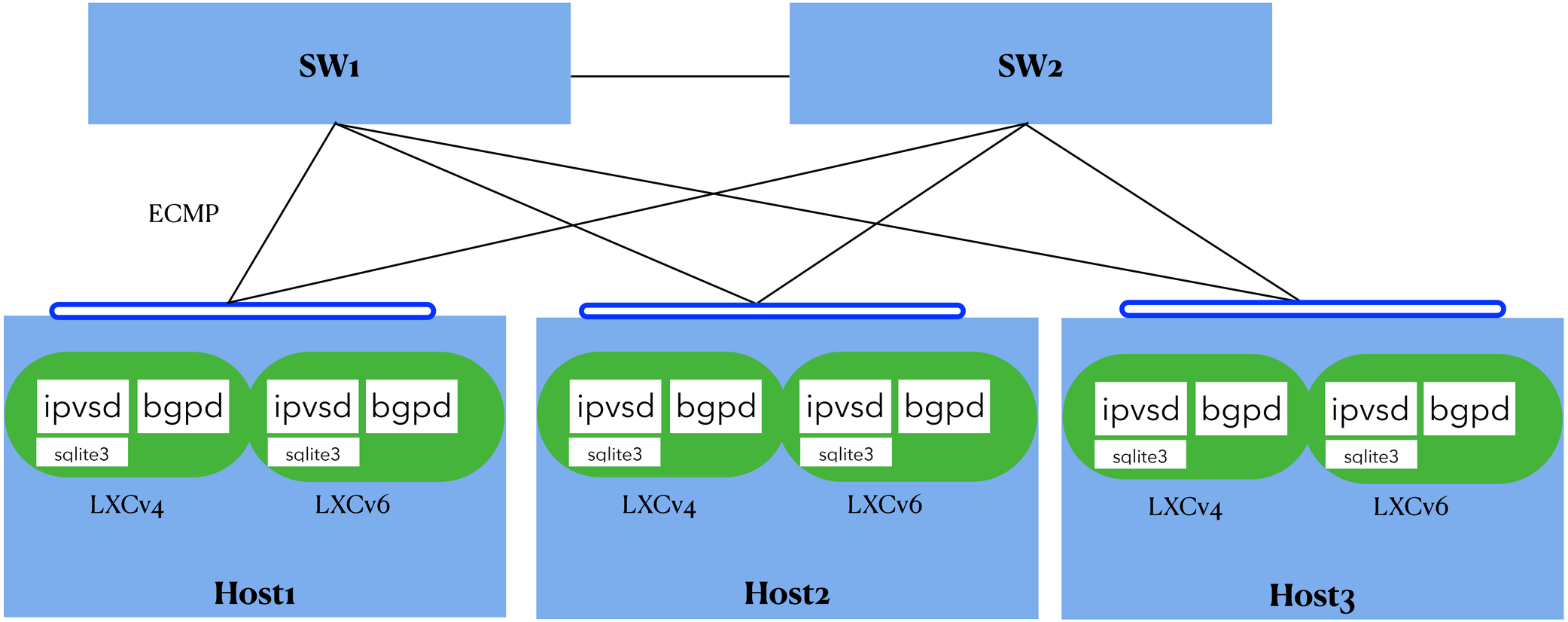
ELB(Public) Physical Topology



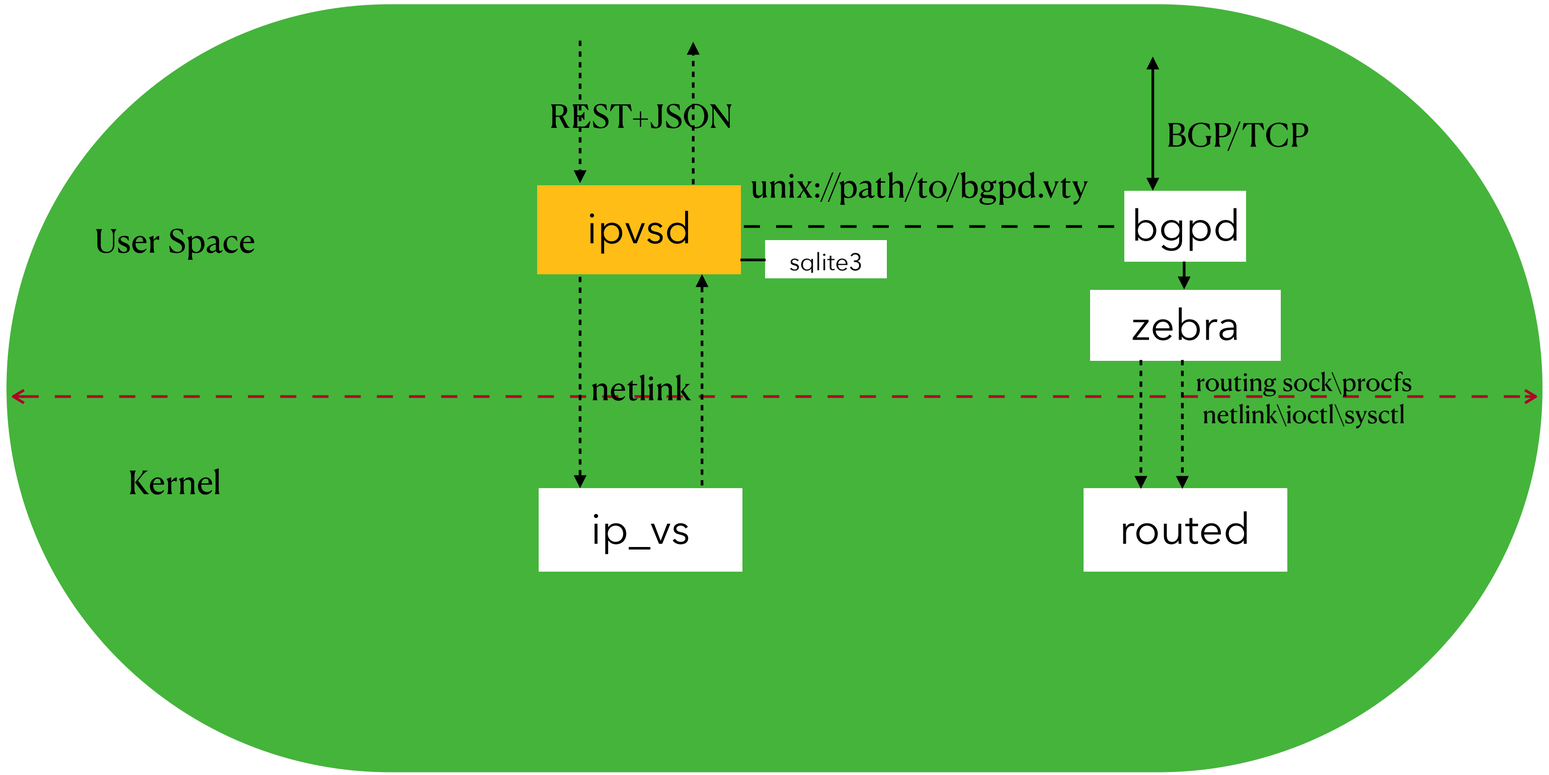
ELB(Public) LVS Logic Topology



ELB(Public) LVS Infra Arch



ELB(Public) LVS Inside LXC



LXC on Host

LVS Controller Capability Matrix

VIP	CRUD	Network	
VS	CRUD	Scheduler	Monitor
RS	CRUD	FWMark	HealthCheck

- GetServices()
- GetService(s *Service)
- NewService(s *Service)
- IsServicePresent(s *Service)
- UpdateService(s *Service)
- DelService(s *Service)
- Flush()

- GetDestinations(s *Service)
- NewDestination(s *Service, d *Destination)
- UpdateDestination(s *Service, d *Destination)
- DelDestination(s *Service, d *Destination)

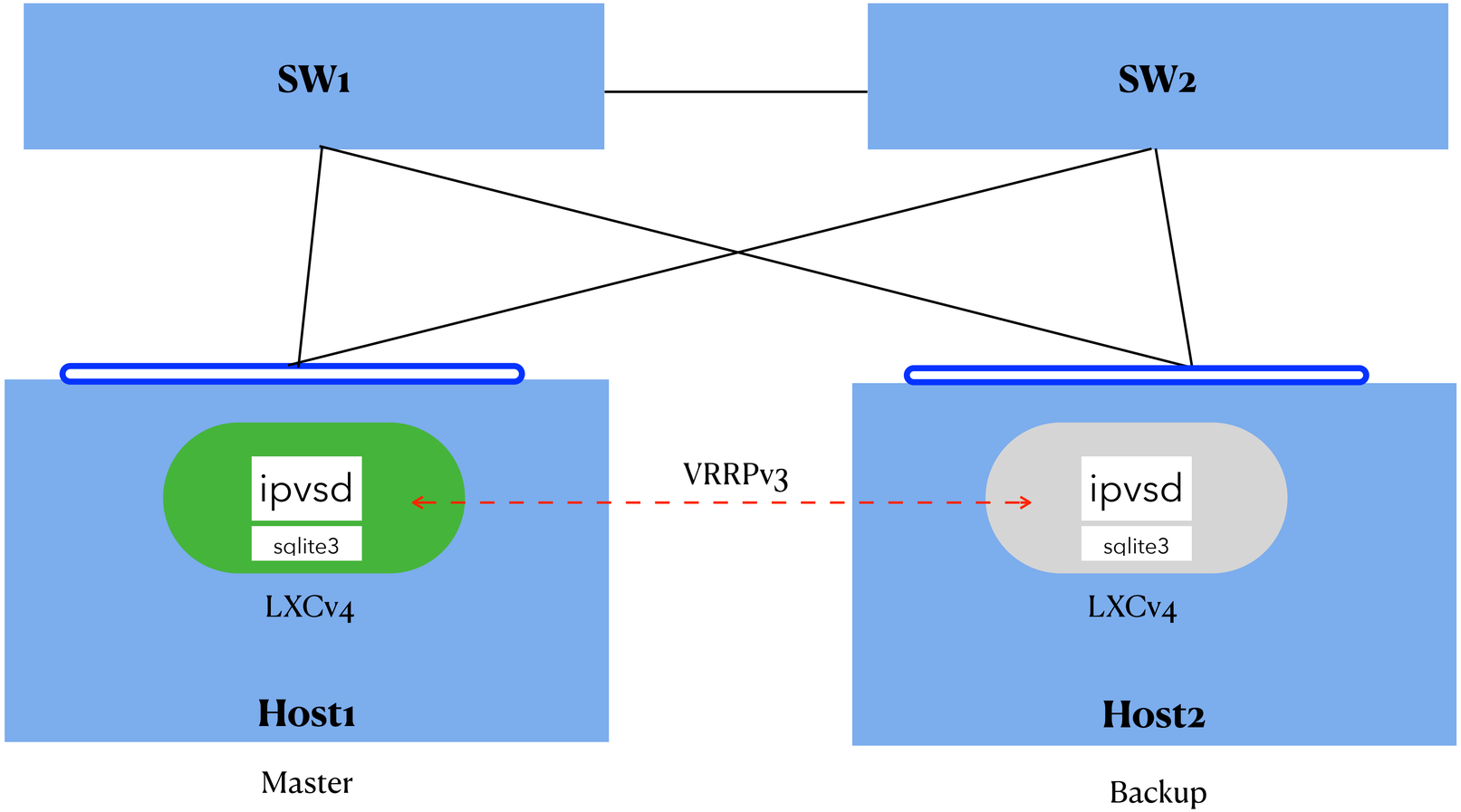
from https://github.com/moby/ipvs/blob/master/ipvs.go

- NewBGP(socket string, asn uint32) *BGP
- Configuration(string, error)
- Neighbor() ([]*Neighbor, error)
- Advertise(n *net.IPNet) error
- Withdraw(n *net.IPNet) error

- NewVTY(socket string) *VTY
- Dial() error
- Close() error
- Commands(cmds []string) error
- Command(cmd string) error

from https://github.com/google/seesaw/blob/master/quagga/bgp.go
https://github.com/google/seesaw/blob/master/quagga/vty.go

ELB(Private) LVS Infra Arch



- https://tools.ietf.org/html/rfc5798

- NewNode(cfg NodeConfig, conn HAConn, engine HAEngine) *Node
- Run() error
- Shutdown() error
- becomeMaster()
- becomeBackup()

```
// Node represents one member of  a high availability cluster
type Node struct {
NodeConfig
conn HAConn //IPHAConn implement IP multicast
engine HAEngine
haStatus seesaw.HAStatus //Master\Backup\Disable
...
}
```

from https://github.com/google/seesaw/blob/master/ha/core.go

```
// send translates an advertisement into a []byte and passes it to the IP layer for delivery.
func (c *IPHAConn) send(advert *advertisement, timeout time.Duration) error {
        deadline := time.Now().Add(timeout)
        if err := c.sendConn.SetWriteDeadline(deadline); err != nil {
            return err
        }
        // fill checksum
        // ....
        buf := new(bytes.Buffer)
        if err := binary.Write(buf, binary.BigEndian, advert); err != nil {
            return err
        }
        // multicast IPv4 address 224.0.0.18
        if _, err := c.sendConn.WriteToIP(buf.Bytes(), &net.IPAddr{IP: c.raddr}); err != nil {
            return err
        }
        return nil
}
```

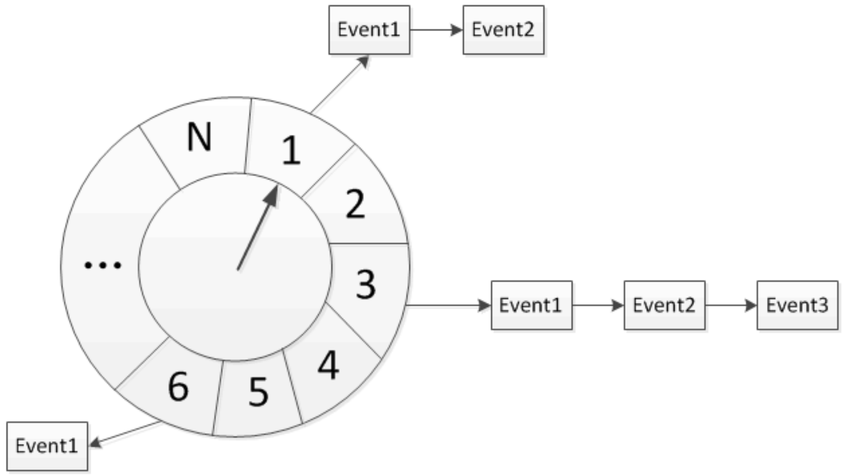
from https://github.com/google/seesaw/blob/master/ha/core.go
https://github.com/google/seesaw/blob/master/ha/net.go

```
// advertisement represents a VRRPv3 advertisement packet
type advertisement struct {
        VersionType  uint8
        VRID         uint8
        Priority     uint8
        CountIPAddrs uint8
        AdvertInt    uint16
        Checksum     uint16
}
```

We need GratuitousARP?

HealthCheck

- TimeWheel Accuracy
- Batch Events
- Mark RS health\unhealthy
- How to check:HTTP\TCP\UDP\ICMP…



from https://www.lanindex.com

Any Questions?