load balancing & clustering in nft

Netfilter Workshop, July 2019
Málaga

Intro

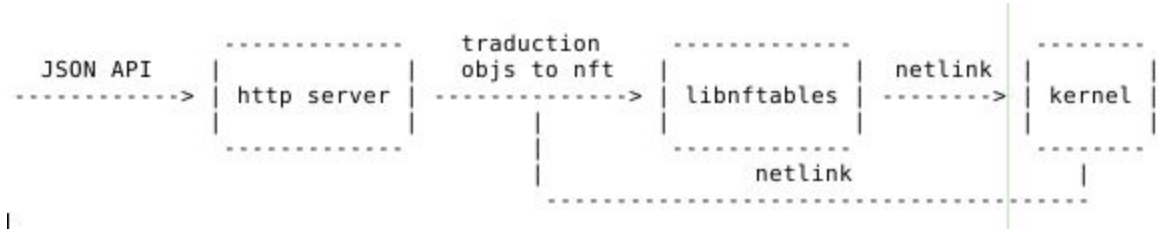
nftlb: user space daemon that manages nftables rules

for load balancing and security policies

https://github.com/zevenet/nftlb

v0.5

Intro


JSON API
http server
traduction
objs to nft
libnftables
netlink
kernel
netlink

Load Balancing

- ★ Topologies supported: Destination NAT, Source NAT, Direct Server Return
- ★ Support for both IPv4 and IPv6 families (Inet not yet supported)
- ★ Multiport support for ranges and lists of ports
- ★ Multiple virtual services (or farms) support
- ★ Schedulers available: weight, round robin, configurable hash (per IP, port, MAC or combination of them) and symmetric hash.
- ★ Priority support per backend

Load Balancing since 0.2

- ★ New topology: Stateless DNAT
- ★ L7 helpers support: sip, ftp, etc
- ★ logging support for input connections
- ★ mark masking flows per virtual service or backend
- ★ add custom source IP address per virtual service and per backend
- ★ Support of configurable persistence or client-backend affinity with a timeout (per IP, port, MAC or combination of them)
- ★ configurable logging messages using tokens

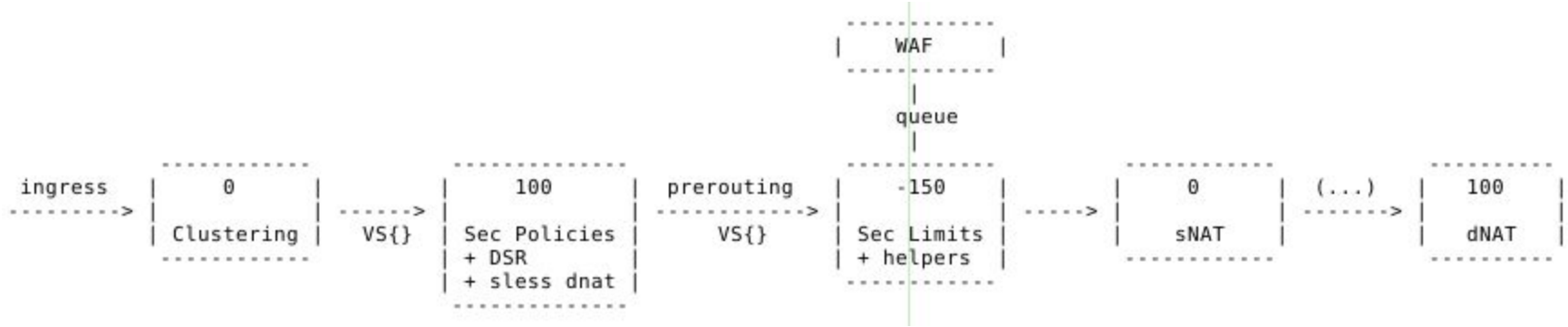
Security Policies

- ★ Support of security policies per service: white and blacklists (from ingress)
- ★ queuing to user space filter
- ★ filtering of bogus TCP frames
- ★ maximum number of established connections per virtual service
- ★ limit TCP RST per second
- ★ limit new connections per second
- ★ maximum number of established connections per backend
- ★ Support for local services as well

Clustering

- ARP filtering from ingress previous to the security policies and the DSR services
- Floating IPs with configurable source address per backend, conntrackd will do the rest.
- User space replication maps.
  - nft-sync
  - ssyncd (https://github.com/zevenet/ssyncd/): L7 proxy ctl socket, xt_recent and now nft maps + json (SEGV during cache_update() listing elements!)

Hooks


WAF
queue
ingress
0
Clustering
VS{}
100
Sec Policies
+ DSR
+ sless dnat
prerouting
VS{}
-150
Sec Limits
+ helpers
0
sNAT
(...)
100
dNAT

Integration

- ★ 1+½ years of nftlb development
- ★ 1 year to replace *tables to nftables (all features that we use)
- ★ 2 major releases with nftables running since January
- ★ Latest LTS 4.19 + new features patches from 5.0
- ★ nftables support in Kubernetes proposal
  https://github.com/kubernetes/kubernetes/issues/62720

Kubernetes prototype with kube-nftlb
https://github.com/zevenet/kube-nftlb

Todo list

- ★ nft-sync selection of rules to replicate (use of nft object flags? similar to handle)
- ★ Conntrack offload
- ★ Conntrack events management
- ★ Service metrics (conntrack + counters in ingress?)
- ★ Usage of inet
- ★ Avoid the usage of marks for persistence and connection limits per bck (user data thing between chains?)
- ★ nat with IP:port for maps
- ★ global maps to create fast paths for security
- ★ clustering multi-node
- ★ support for NAT64/NAT46