nftables cluster deployment at CICA

Our experience

Arturo Borrero Gonzalez
Netfilter Workshop 2017

CICA

→ CICA: Regional NREN Andalusia, Spain
→ Part of RedIRIS NREN
→ Since 1989

Key points:
→ Own complete data center
→ big network deployment (40G / 10G backbone)
→ 300 Linux server ; including HPC clusters
→ committed to Free/Libre Open Source Software

Context

Our clients:

→ Researchers

→ Students

→ General internet users

Typical security issues:

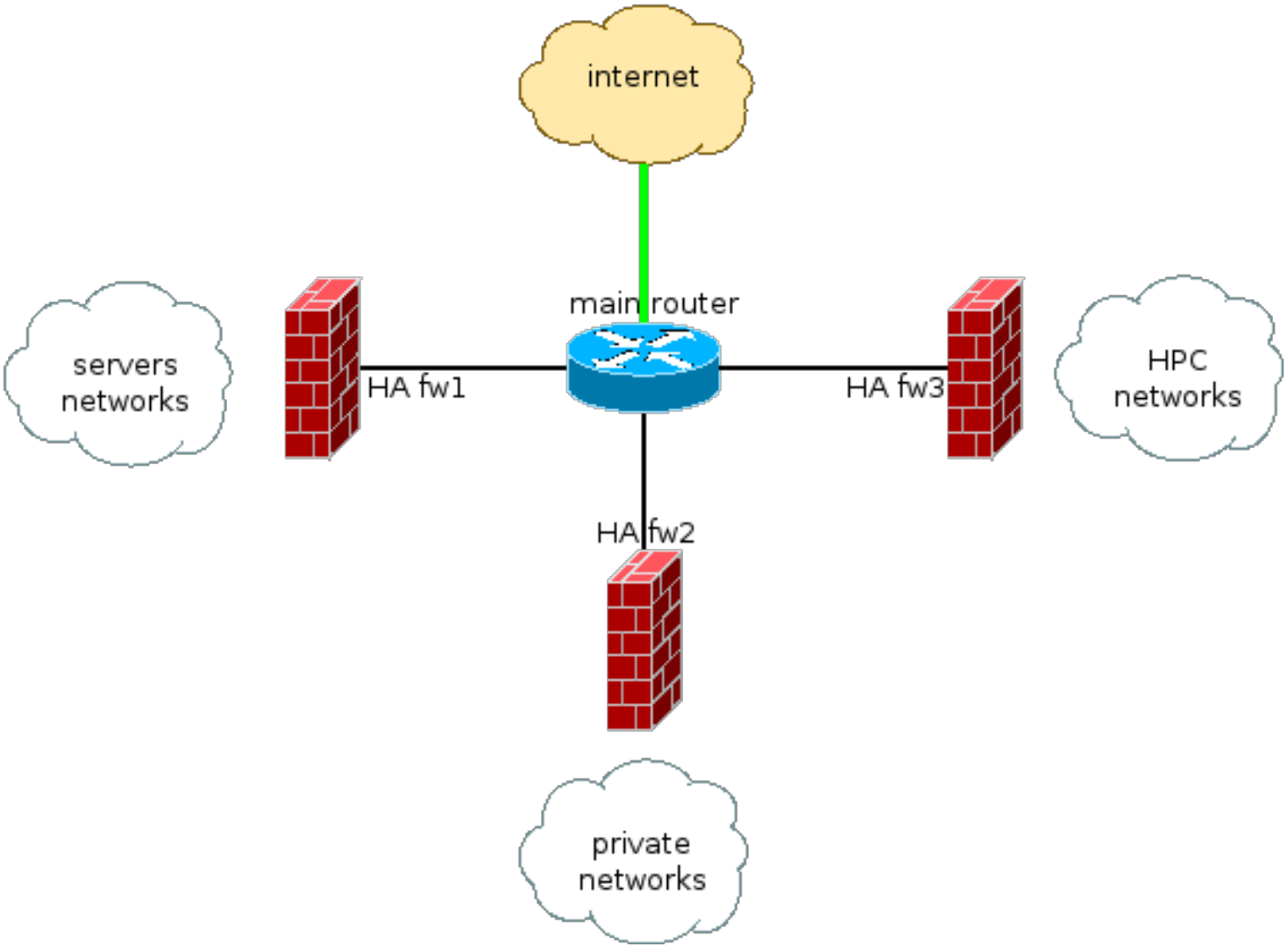
→ amplifications (NTP/DNS/etc)

→ simple leaks (f.e. printers reachable from the internet)

→ email spamming from internal servers

We need packet firewalls!

Context


internet
main router
servers networks
HA fw1
HA fw3
HPC networks
HA fw2
private networks

Context

Our requirements:

→ stateful firewall

→ HA, active-active

→ IPv4 / IPv6 dual stack

→ policy atomic updates

→ policy version control (i.e, git)

Past deployments:

→ custom scripts to ease dual stack

→ non-atomic :-(

→ no version control

→ HA, backup-slave :-(

Tech

We use: netfilter
firewalling, NAT, and packet mangling for linux

Others technologies don’t match our requirements:

→ stateless firewalling

→ not FLOSS

→ not designed for sysadmins

→ too much marketing-driven

→ over-engineered for our use case

Model

The solution:

→ active/active symmetric cluster

→ conntrackd NOTRACK mode (faster failover, no lazy backup)

→ HA: pacemaker/corosync

→ use git hooks to update policy (nft -f)

No lazy backup?

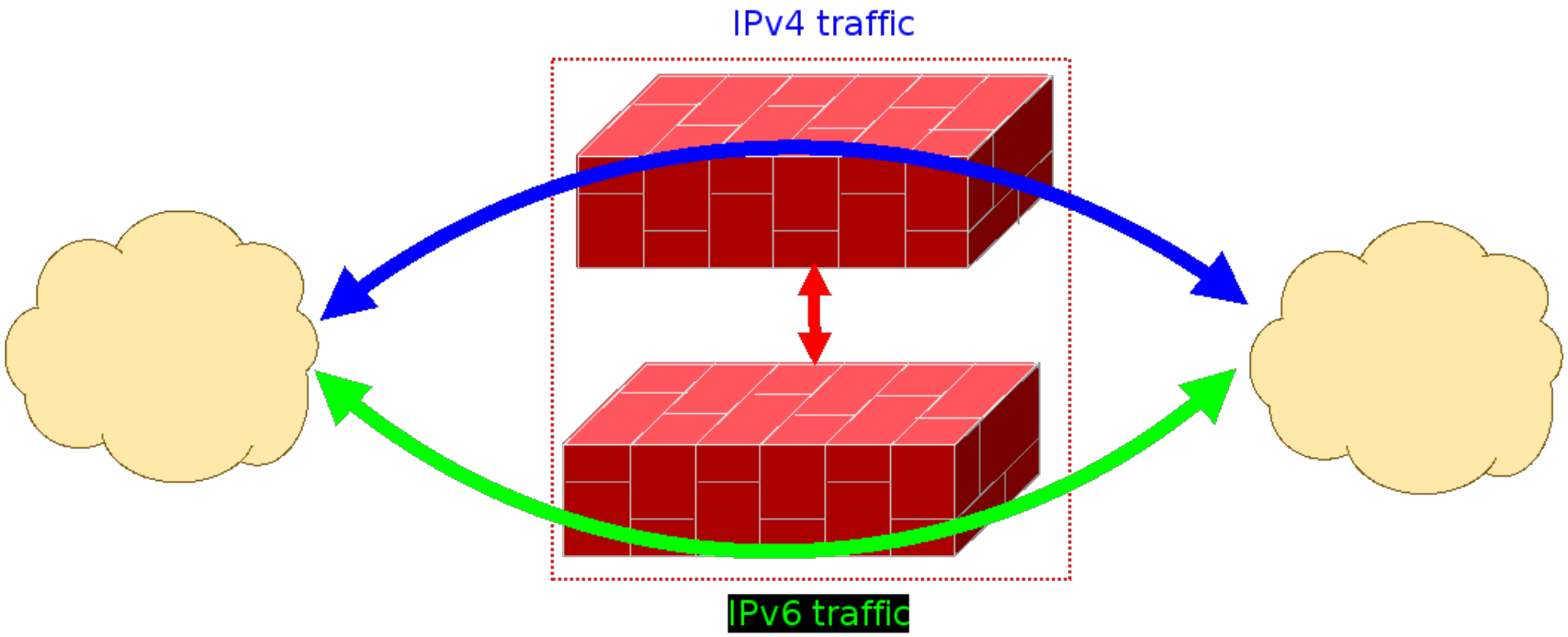
→ a node which does nothing is hard to justify

→ one node IPv4, other IPv6

→ both have same policy and are updated by git hook

Model

Each node filters a network protocol

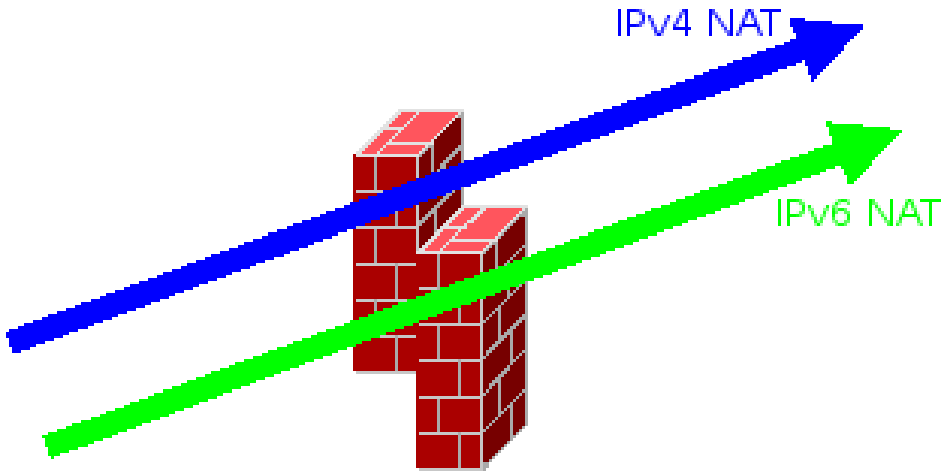


conntrackd

→ Fixed issue: lack of IPv6 NAT support

[PATCH] conntrackd: add support for NTA_(S|D)NAT_IPV6

→ Ease dual stack management by mirroring IPv4 NAT

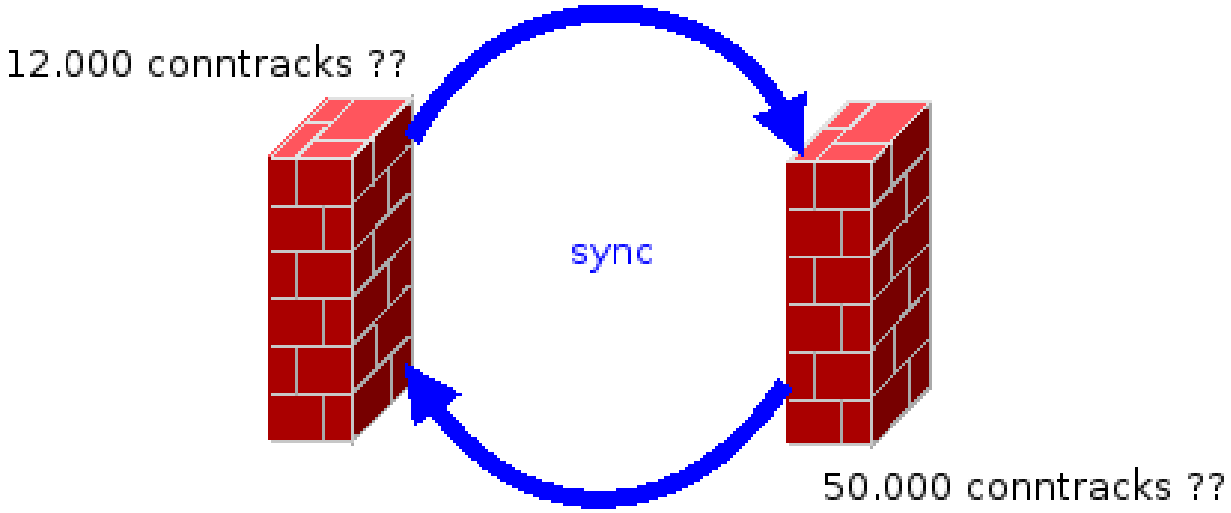


conntrackd

→ Fixed issue: no [NEW] conntrack in slave node (cosmetic)

[PATCH] netfilter: ctnetlink: using bit to represent the ct event

→ Open issue: number of conntracks diverge in each node



conntrackd

→ Feature added: systemd integration (optional)

[PATCH] conntrackd: add systemd support

→ Saw random crashes in the past. Use watchdog.

debian bug #796877 *https://bugs.debian.org/796877*

→ Better machine boot process (avoid races)

→ Small code, many benefits

→ Yes, we optionally depend on systemd

From ipt to nft

Mix of approaches:

1) Use translation tools where possible

2) Use custom scripts in some cases (ie, ipset)

3) Manual translation

→ Hard work even with the translations/compat tools

→ In some cases, native rewriting is easier/cheaper

Thanks nft!

→ inet family

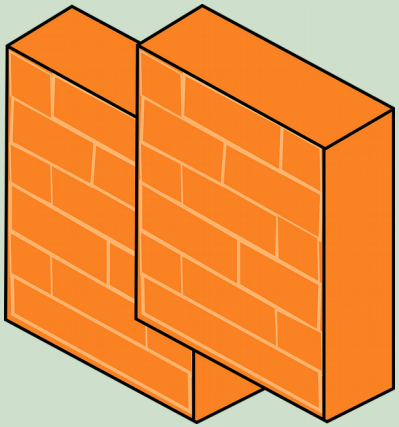
→ sets / maps / concatenations

→ defines (variables)

→ multiple actions in a single rule

→ friendly syntax

→ ruleset still plain text, use git



Issues with nft

→ error reporting in nested/included files

→ small set names

→ input/output not reentrant

→ no support for iifname in concatenations

→ sets elements printing

But so far, so good!

error reporting

→ misleading message

→ already fixed

[PATCH] src: error reporting for nested ruleset representation

```
% nft -f ruleset.nft

ruleset.nft:2:1-2: Error: Could not process rule: No such file or directory

table t {
^^
```

```
% nft -f ruleset.nft

In file included from ./ruleset2.nft:3:1-25:
                 from ruleset.nft:5:17-41:
./ruleset3.nft:1:1-7: Error: Could not process rule: No such file or directory

jump x
^^^^^^
```

small set names

→ currently limited to 32

→ even worst with bad error reporting

→ Open discussion: Make name variable sized?

```
% nft add set t abcdefghijklmnopqrstuvwxyzzzzzzz {type ipv4_addr\; }

<cmdline>:1:1-62: Error: Could not process rule: Numerical result out of range

add set t abcdefghijklmnopqrstuvwxyzzzzzzz {type ipv4_addr; }
^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
```

input/output not reentrant

→ can’t add the ruleset you obtain from ‘nft list ruleset’

→ already fixed (in detected cases)

[PATCH] payload: explicit network ctx assignment for icmp/icmp6 in special families

```
% nft add rule inet t c ip6 nexthdr icmpv6 icmpv6 type echo-request


% nft list ruleset
[…]
icmpv6 type echo-request
[…]


% nft add rule inet t c icmpv6 type echo-request
<cmdline>:1:19-29: Error: conflicting protocols specified: inet-service vs. Icmpv6
add rule inet t c icmpv6 type echo-request
                  ^^^^^^^^^^^
```

strings in concatenations

→ concats are hashes, strings are variable sized datatype

→ useful for iifname-based ruleset trees

→ open challenge

```
% nft add rule t c iifname . tcp sport {eth0 . 2}

<cmdline>:1:14-20: Error: can not use variable sized data types (string) in concat expressions
add rule t c iifname . tcp sport {eth0 . 2}
             ^^^^^^^^~~~~~~~~~~~~~
```

sets elements printing

→ unsorted, already fixed

[PATCH] src: sort set elements in netlink_get_setelems()

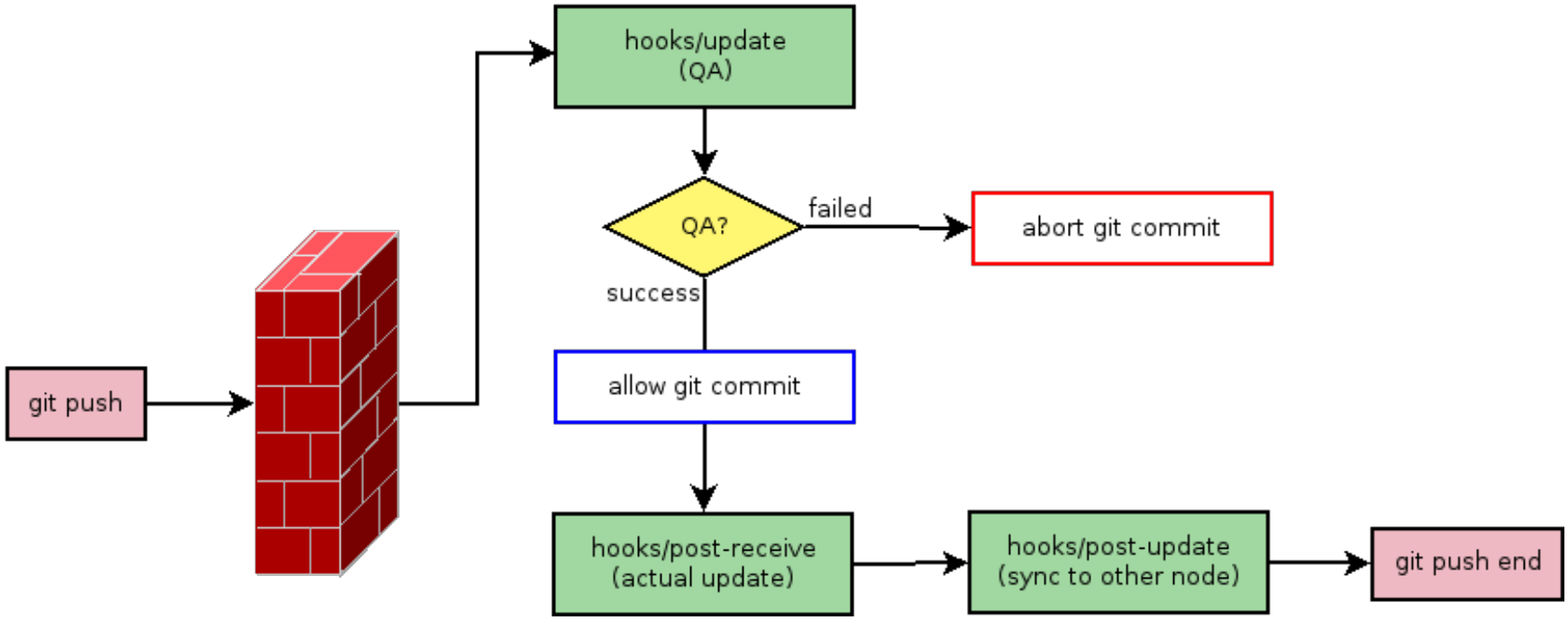
→ printed in single line, already fixed

[PATCH] expression: print sets and maps in pretty format

nftables syncing

→ git hooks custom scripts

→ could this be **nft-sync**?



Thanks!

Arturo Borrero Gonzalez
Netfilter Workshop 2017