Netfilter Workshop 2010 Summary

David S. Miller

Red Hat Inc.

Boston, USA, 2010

ATTENDEES

- Pablo Neira Ayuso (host)
- Patrick McHardy
- Harald Welte
- Jozsef Kadlecsik
- Holger Eitzenberger
- Jan Engelhardt
- Simon Horman
- Eric Leblond
- Jesper Dangaard Brouer
- Krisztian Kovács
- Balázs Scheidler
- Florian Westphal
- Eric Dumazet
- Sven Schnelle
- Ulrich Weber

IPVS Developments: Simon Horman

- DNAT support
- One Packet Scheduling
- Persistence Engine (f.e. SIP)
- Network Namespace support (in-progress)
- Connection Synchronization (also in-progress)

Ingress Bandwidth Shaping: Jesper Dangaard Brouer

- IFB vs. iptables
- IFB happens at the wrong point
- iptables requires hackish out of tree patch

RPS, RFS, AND SKB LISTS: DAVID S. MILLER

- Who cares...

XTABLES-ADDONS AND BLOBS: JAN ENGELHARDT

- Addons maintains out-of-tree netfilter modules
- For obscure yet useful to someone stuff
- Not ready for upstream, although some end up merged
- Blob layout and format improvement attempted
- But in the end, current format turns out to be optimal
- Memory usage, “jumping”, cache locality
- Jumps are expressed as offsets in ruleset

UPLINK BALANCING: ULRICH WEBER

- Handling multiple uplinks with working DNAT, etc.
- SNAT also creates problems (happens POSTROUTING)
- Use conntrack to query pre-SNAT'd address
- Only on routing cache miss
- Similar to multipath routing cache we used to have

OOFW: ERIC LEBLOND

- Completely crazy.

GPL Violations: Harald Welte

- A bit offtopic...

IPSET: JOZSEF KADLECSIK

- Optimizes matching on “sets” of addresses
- User interface rewrite to handle ipv6 and use netlink
- Different primitives: hash, bitmap, etc.
- Uses libmnl, see next slide
- Hopefully we can merge this soon
- It’s been around forever

LIBMNL: PABLO NEIRA AYUSO

- Already have 2 other netlink libraries, so why?
- libnetlink strongly tied to iproute2
- libnl abstracts I/O too heavily for some needs
- Small set of helpers, and minimal callback mechanism
- Any aspect can be done by hand by application
- Bulk of library is attribute validation/construction
- Does not obviate libnl and friends
- GIT: git://1984.lsi.us.es/libmnl

nftables: Patrick McHardy

- Userspace generates filter “codes”, and uploads to kernel
- Kernel runs the filter blob as the ruleset
- Similar to BPF but more sophisticated
- Contains high-level objects, like dictionaries
- Hash table, trie, bitmaps, etc.
- Ruleset optimizations, in userspace
- Need to retain existing netfilter bits “forever”
- Unification of filters (socket filter, pktsched, netfilter)

TPROXY: Balázs Scheidler

- IPV6 support
- Lots of bug fixes:
- Hash table lookup crashes
- Conntrack identity clashes
- All of this is in 2.6.37-RC1

IPTV: Jesper Dangaard Brouer

- Packet drops –> crappy video
- Analyzing 600mbit traffic
- Increasing backbone pipe didn't help
- MPEG2 stream analyzer xtables module
- Quickly found that bursts caused problems
- Ethernet switch buffers too small

PERF: ERIC DUMAZET

- Analysis of SMP performance problems using perf
- Test case: UDP DoS
- Routing table writes to read-heavy value
- Simple fix.
- Routing table backend now faster than routing cache

OPENVSWITCH: SIMON HORMAN

- kernel side is simple dumb flow switch
- userspace does signalling, table bulding etc.
- VLAN encapsulation types (802.1ad, 802.1ah)
- Offloading
- QoS

GIT: JAN ENGELHARDT

- Pure git, stgit, and quilt
- Interactive additions and rebasing
- Again, mostly offtopic for here