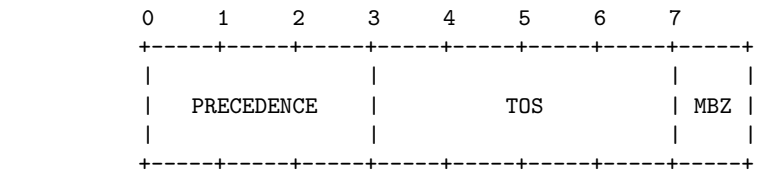
Contribution ID: **129** Type: **not specified**

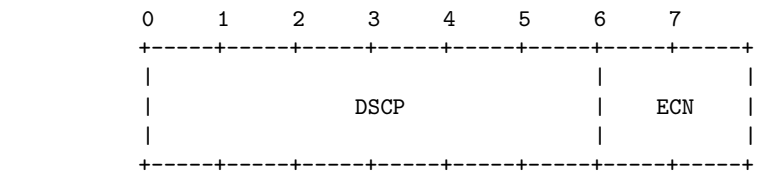
Untangling DSCP, TOS and ECN bits in the kernel

Friday, September 24, 2021 10:20 AM (40 minutes)

In Linux, the IPv4 code generally uses IPTOS_TOS_MASK (0x1e) when handling the TOS (Type of Service) of IPv4. This mask follows the definition of RFC 1349:



However RFC 1349 is only one of several contradicting RFCs that try to define how to interpret the IPv4 TOS. In the end, the IETF settled on the DSCP+ECN interpretation (RFC 2474 and RFC 3168):



That was 20 years ago, so the layout is finally stable. But as the diagrams show, RFC 1349 is incompatible with ECN as it already uses bit 6 in its TOS field.

Therefore, the IPv4 code also uses another mask, IPTOS_RT_MASK (0x1c), to clear bit 6. This mask is used almost every time the kernel does an IPv4 route lookup.

Finally, RFC 2474 and RFC 3168 (DSCP+ECN) also cover IPv6. However, the IPv6 code generally doesn't mask the ECN bits and considers them as part of the TOS for policy routing.

This situation creates several problems:

- Regressions brought by patches "fixing" places where IPTOS_TOS_MASK wasn't applied (thus breaking users that used bits 0-2).
- IPTOS_TOS_MASK is spreading to IPv6 (through RT_TOS()), where it doesn't make sense at all (IPv6 has never used the RFC 1349 layout).
- In some edge cases, IPv4 route lookups are done without masking the ECN bits (thus giving different results depending on the ECN mark). New cases are introduced every now and then.
- IPv4 and IPv6 inconsistency.
- Impossibility to use the full DSCP range in IPv4.
- Policy-routing can break ECN with IPv6 and in some IPv4 edge cases.
- Parts of the stack define their own mask to respect the DSCP+ECN layout, but without making it reusable.

The objective of this talk is to bring practical examples of user-visible inconsistencies and to discuss different ways forward for minimising them and avoiding more ECN regressions in the future.

It will be oriented towards the following goals (by decreasing order of perceived feasibility):

- Remove all uses of IPTOS_TOS_MASK for IPv6.
- Prevent IPv4 policy routing from breaking ECN.
- Remove IPTOS_TOS_MASK entirely from the kernel, so people don't mistakenly copy/paste such code (but keep the definition in include/uapi of course).
- Allow full DSCP range in IPv4.
- Prevent IPv6 policy routing from breaking ECN.
- Prevent breaking ECN again in the future (for example by defining a new type for storing TOS values, so that Sparse could warn about invalid use cases).
- Make TOS and ECN handling consistent between IPv4 and IPv6 (somewhat implied by the previous bullet points).

The main road blocks are code churn and drawing the line between bugs and established behaviours.

I agree to abide by the anti-harassment policy

I agree

Primary author: NAULT, Guillaume (Red Hat)

Presenter: NAULT, Guillaume (Red Hat)

Session Classification: BPF & Networking Summit

Track Classification: Networking & BPF Summit (Closed)