TC-PBR (WIP)

Roopa Prabhu
Cumulus Networks

What is PBR ?

- Policy based routing
- Example:
  - On match redirect to
    - Interface
    - Nexthop
    - Ecmp nexthop
    - vrf

Current Linux PBR solution

- Ip rules
- All routing redirects can be achieved by redirecting to
  - A separate table with the new routing override
- Issues:
  - A separate route table for every route policy change (Fixable and I have patch to avoid that by introducing new actions).
  - Scale might be a problem…but I do not have numbers here to say anything about it
  - No offload API yet (but fixable if needed)

Motivation for PBR with tc

- When it comes to offload, PBR are just acl rules
  - Would be nice to deploy them along with other ACL rules to maintain priority/precedence between rules
  - Use same offload API

Motivation for PBR with tc (continued)

- Same reasons to why there will be a tc conntrack action:
  - Netfilter has all the right routing hooks in the right places
  - Can we leverage them from Tc when using classifiers and filters from tc ?

Possible options

TC PBR action to route packet directly at ingress hook

- Tc <match> action <redirect-via-route-table>
- Action routes the packet directly with a route lookup in the appropriate table
- Works at ingress, egress hook too late

TC PBR action to attach route policy to skb

- TC match at ingress, use dst_metadata to attach routing policy to packets at ingress:
  - Have some WIP patch. Hit a problem with dst_metadata being dropped early
  - Can be made to work at ingress, tc egress hook too late

New TC hook at routing layer

- Add a new hook at the routing layer
- OR
- leverage an existing netfilter hook

Other Challenges

- PBR rules are global rules not tied to an interface
- We don't even need per interface stats for this
- Netfilter can do this, tc to some extent with shared blocks ?
    - Open Questions:
        - Will we still need the shared block applied to all interfaces ?
        - Will stats still show up per interface ?
        - (doing this on a system with hundreds of interfaces might not scale well)

Thank you