AN INTRODUCTION TO LINUX POLICY ROUTING

Tom Eastep

SeaGL 2013

2013-10-12

Seattle, Washington

- About the presenter
- Routing
- Routing Tables
- Routing Rules
- The route cache
- Defining additional Tables
- Routing/Netfilter interaction
- Use cases

AGENDA

- Software Architect at Hewlett-Packard – High availability file systems.
- Telecommuter
- 44 Years in the computer industry
- Creator of Shorewall
- Self-taught concerning Linux and networking
  - *IP Fundamentals: What Everyone Needs to Know About Addressing & Routing*, Thomas A. Maufer, June 4, 1999, ISBN-10: 0139754830, ISBN-13: 978-0139754838, Edition: 1

ABOUT THE PRESENTER

- Process of determining what to do with a packet
  - Process on the local system
  - Send directly to the destination via a network interface
  - Forward the packet to a router
  - Return an error (ICMP) to the sender
  - Ignore

ROUTING

- Displaying the current routes with *ip*

```
[teastep@centos ~]$ ip route ls
172.20.1.0/24 dev eth1  proto kernel  scope link  src 172.20.1.136  metric 1
default via 172.20.1.254 dev eth1  proto static
[teastep@centos ~]$
```

Note: 'ip' commands are the same as its output. To create the second route:

```
ip route add default via 172.20.1.254 dev eth1 proto static
```

ROUTING – TRIVIAL CASE

```
teastep@mint14 $ ip route ls
15.192.0.142 via 172.20.1.254 dev eth1  proto static
70.90.191.120/29 via 172.20.1.254 dev eth1  proto static
10.0.2.0/24 dev eth0  proto kernel  scope link  src 10.0.2.15  metric 1
172.20.1.0/24 dev eth1  proto kernel  scope link  src 172.20.1.191  metric 1
default via 10.0.2.2 dev eth0  proto static
teastep@mint14 $
```

Note 1: The above output is sorted – 'ip route ls' output is unsorted ☹

Note 2: 'shorewall show routing' output is sorted.

ROUTING – TWO INTERFACES

- There are multiple routing tables, each one identified by a unique number
- The rt_tables file allows assigning names to the tables

```
root@mail:~# cat /etc/iproute2/rt_tables
#
# reserved values
#
255     local
254     main
253     default
0       unspec
…
root@mail:~#
```

ROUTING TABLES

- The *main* table is the default for commands

```
[teastep@centos ~]$ ip route ls
172.20.1.0/24 dev eth1  proto kernel  scope link  src 172.20.1.136  metric 1
default via 172.20.1.254 dev eth1  proto static
[teastep@centos ~]$ ip route ls table main
172.20.1.0/24 dev eth1  proto kernel  scope link  src 172.20.1.136  metric 1
default via 172.20.1.254 dev eth1  proto static
```

ROUTING – THE MAIN TABLE

- The *local* table defines addresses on the host as well as broadcasst addresses

```
root@mail:~# ip route ls table local
broadcast 70.90.191.120 dev eth0  proto kernel  scope link  src 70.90.191.124
broadcast 70.90.191.120 dev eth1  proto kernel  scope link  src 70.90.191.122
local 70.90.191.122 dev eth1  proto kernel  scope host  src 70.90.191.122
local 70.90.191.124 dev eth0  proto kernel  scope host  src 70.90.191.124
broadcast 70.90.191.127 dev eth0  proto kernel  scope link  src 70.90.191.124
broadcast 70.90.191.127 dev eth1  proto kernel  scope link  src 70.90.191.122
broadcast 127.0.0.0 dev lo  proto kernel  scope link  src 127.0.0.1
local 127.0.0.0/8 dev lo  proto kernel  scope host  src 127.0.0.1
local 127.0.0.1 dev lo  proto kernel  scope host  src 127.0.0.1
broadcast 127.255.255.255 dev lo  proto kernel  scope link  src 127.0.0.1
root@mail:~#
```

ROUTING TABLES - CONTINUED

- The *default* table is normally empty

```
root@mail:~# ip route ls table default
root@mail:~#
```

- Unused tables are also empty

```
root@mail:~# ip route ls table 100
root@mail:~#
```

ROUTING TABLES - CONTINUED

ROUTING TABLES - CONTINUED

- Routes may be added to any table

```
root@mail:~# ip route add 1.2.3.4/32 dev eth1 table 100
root@mail:~# ip route ls table 100
1.2.3.4 dev eth1  scope link
root@mail:~#
```

ROUTING RULES

- Routing rules define the order in which the tables are traversed
- Rules are processed until the packet is routed

```
root@mail:~# ip rule ls
0:      from all lookup local
32766:  from all lookup main
32767:  from all lookup default
root@mail:~#
```

- Routing table lookups are *cached*
- The cache is searched before the tables

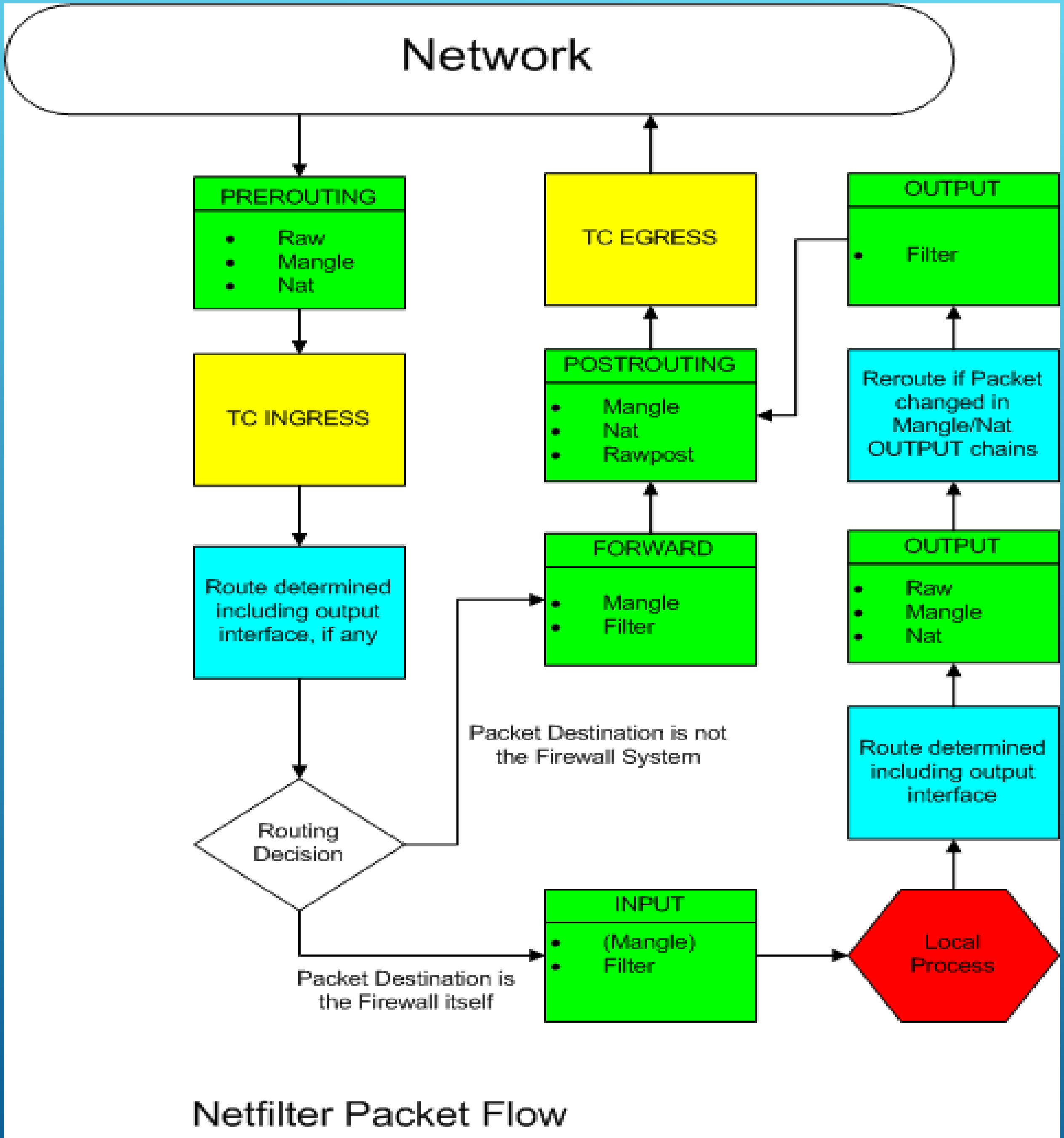
```
root@mail:~# ip route ls cache
172.20.1.145 from 70.90.191.122 via 70.90.191.121 dev eth0
    cache  ipid 0x8a81 rtt 100ms rttvar 78ms cwnd 10
66.249.74.23 from 70.90.191.124 via 70.90.191.121 dev eth0
    cache  ipid 0xbd7b
213.188.126.148 from 70.90.191.124 via 70.90.191.121 dev eth0
    cache  ipid 0x77d3
local 70.90.191.122 from 172.20.2.254 dev lo  src 70.90.191.122
    cache <local>  ipid 0x64b9 iif eth1
201.162.19.120 from 70.90.191.124 via 70.90.191.121 dev eth0
…
root@mail:~#
```

ROUTING RULES

- Routing rules have predicates

```
root@mail:~# ip rule lsroot@gateway:~# ip rule ls
0:      from all lookup local
999:    from all lookup main
1000:   from 70.90.191.121 lookup ComcastB
1000:   from 70.90.191.123 lookup ComcastB
1000:   from 70.90.191.149 lookup ComcastB
1000:   from 172.20.1.191 lookup ComcastB
1000:   from 10.0.0.4 lookup ComcastC
10000:  from all fwmark 0x10000/0x30000 lookup ComcastB
10001:  from all fwmark 0x20000/0x30000 lookup ComcastC
11000:  from all iif br0 lookup ComcastB
32765:  from all lookup balance
32767:  from all lookup default
root@gateway:~#
```

ROUTING RULES



NETFILTER/ROUTING INTERACTION

- The *PREROUTING* and OUTPUT hooks allow the packet destination and fwmarks to be altered.
- DNAT target in the *nat* table
- MARK target in the *mangle* table

- Multiple Internet Uplinks
- TPROXY
- Transparent Proxy

USE CASES

```
Routing Rules

0:      from all lookup local
999:    from all lookup main
10000:  from all fwmark 0x1/0xff lookup LAN
10001:  from all fwmark 0x2/0xff lookup WLAN
20000:  from 10.0.0.10 lookup LAN
20000:  from 172.20.1.153 lookup WLAN
32765:  from all lookup balance
32767:  from all lookup default

Table balance:

default via 10.0.0.1 dev eth0

Table default:

default via 172.20.1.254 dev eth1 src 172.20.1.153 metric 2

Table LAN:

default via 10.0.0.1 dev eth0 src 10.0.0.10
```

MULTIPLE INTERNET PROVIDERS

```
Table local:

local 172.20.1.153 dev eth1 proto kernel scope host src 172.20.1.153
local 127.0.0.1 dev lo proto kernel scope host src 127.0.0.1
local 10.0.0.10 dev eth0 proto kernel scope host src 10.0.0.10
broadcast 172.20.1.255 dev eth1 proto kernel scope link src 172.20.1.153
broadcast 172.20.1.0 dev eth1 proto kernel scope link src 172.20.1.153
broadcast 127.255.255.255 dev lo proto kernel scope link src 127.0.0.1
broadcast 127.0.0.0 dev lo proto kernel scope link src 127.0.0.1
broadcast 10.0.0.255 dev eth0 proto kernel scope link src 10.0.0.10
broadcast 10.0.0.0 dev eth0 proto kernel scope link src 10.0.0.10
local 127.0.0.0/8 dev lo proto kernel scope host src 127.0.0.1

Table main:

172.20.1.0/24 dev eth1 proto kernel scope link src 172.20.1.153
10.0.0.0/24 dev eth0 proto kernel scope link src 10.0.0.10 metric 1

Table WLAN:

default via 172.20.1.254 dev eth1 src 172.20.1.153
```

MULTIPLE INTERNET PROVIDERS -- CONTINUED

```
Chain PREROUTING (policy ACCEPT 443 packets, 37552 bytes)
 pkts bytes target     prot opt in     out     source               destination
  443 37552 CONNMARK   all  --  *      *       0.0.0.0/0            0.0.0.0/0            CONNMARK restore mask 0xff
  209 16061 routemark  all  --  eth0   *       0.0.0.0/0            0.0.0.0/0            mark match 0x0/0xff
  233 21439 routemark  all  --  eth1   *       0.0.0.0/0            0.0.0.0/0            mark match 0x0/0xff

Chain routemark (2 references)
 pkts bytes target     prot opt in     out     source               destination
  209 16061 MARK       all  --  eth0   *       0.0.0.0/0            0.0.0.0/0            MARK xset 0x1/0xff
  233 21439 MARK       all  --  eth1   *       0.0.0.0/0            0.0.0.0/0            MARK xset 0x2/0xff
  442 37500 CONNMARK   all  --  *      *       0.0.0.0/0            0.0.0.0/0            mark match !0x0/0xff CONNMARK
save mask 0xff
```

MULTIPLE INTERNET PROVIDERS – ENSURE THAT CONNECTIONS ALWAYS USE THE SAME UPLINK

```
root@gateway:~ $ ip rule ls

0:      from all lookup local
1:      from all fwmark 0x80000/0x80000 lookup TProxy
999:    from all lookup main
1000:   from 70.90.191.121 lookup ComcastB
1000:   from 70.90.191.123 lookup ComcastB
1000:   from 70.90.191.149 lookup ComcastB
1000:   from 172.20.1.191 lookup ComcastB
1000:   from 10.0.0.4 lookup ComcastC
10000:  from all fwmark 0x10000/0x30000 lookup ComcastB
10001:  from all fwmark 0x20000/0x30000 lookup ComcastC
11000:  from all iif br0 lookup ComcastB
32765:  from all lookup balance
32767:  from all lookup default

root@gateway:~ $ ip route ls table TProxy

local default dev lo scope host
```

TPROXY – ROUTING PART

```
Chain PREROUTING (policy ACCEPT 379 packets, 52077 bytes)
 pkts bytes target     prot opt in     out     source               destination
…
    0     0 divert     tcp  --  eth1   *       0.0.0.0/0            0.0.0.0/0            tcp spt:80 flags:! 0x17/0x02 socket --transparent
    0     0 divert     tcp  --  eth0   *       0.0.0.0/0            0.0.0.0/0            tcp spt:80 flags:! 0x17/0x02 socket --transparent
    0     0 TPROXY     tcp  --  eth2   *       0.0.0.0/0            0.0.0.0/0            tcp dpt:80 TPROXY redirect 172.20.1.254:3129 mark
0x80000/0x80000

Chain divert (3 references)
 pkts bytes target     prot opt in     out     source               destination
    0     0 MARK       all  --  *      *       0.0.0.0/0            0.0.0.0/0            MARK or 0x80000
    0     0 ACCEPT     all  --  *      *       0.0.0.0/0            0.0.0.0/0
```

Note: In the above configuration, eth0 and eth1 are Internet uplinks and eth2 interfaces to the local LAN.

TPROXY – NETFILTER PART

Q & A