iptables — теория

Netfilter

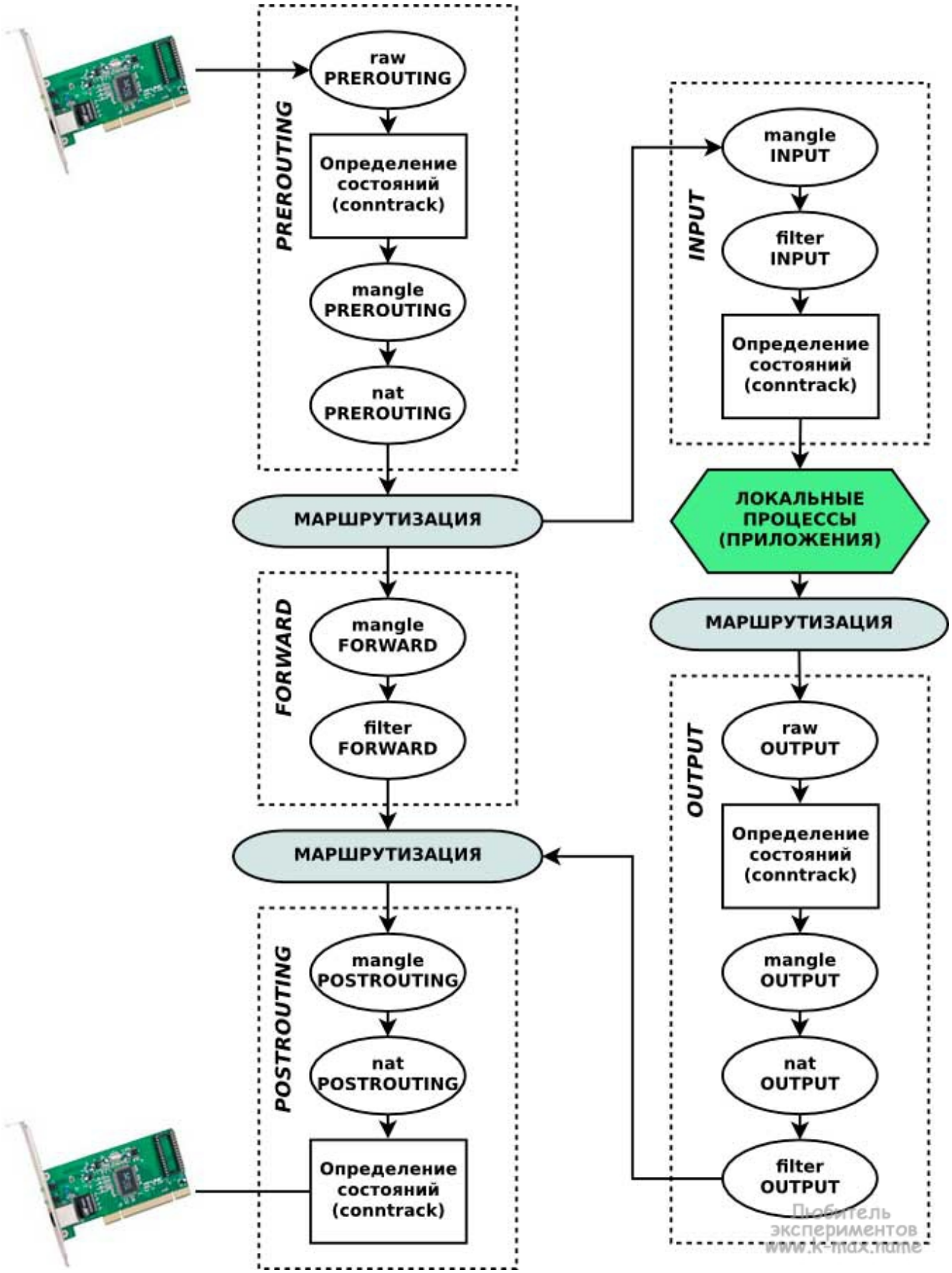


схема работы netfilter

Сетевые пакеты поступают в сетевой интерфейс, настроенный на стек TCP/IP и после некоторых простых проверок ядром (например, контрольная сумма) проходят последовательность

цепочек (chain) (обозначены пунктиром). Пакет обязательно проходит первоначальную цепочку PREROUTING. После цепочки PREROUTING, в соответствии с таблицей маршрутизации, проверяется кому принадлежит пакет и, в зависимости от назначения пакета, определяется куда он дальше попадет (в какую цепочку). Если пакет НЕ адресован (в TCP пакете поле адрес получателя – НЕ локальная система) локальной системе, то он направляется в цепочку FORWARD, если пакет адресован локальной системе, то направляется в цепочку INPUT и после прохождения INPUT отдается локальным демонам/процессам. После обработки локальной программой, при необходимости формируется ответ. Ответный пакет пакет отправляемый локальной системой в соответствии с правилами маршрутизации направляется на соответствующий маршрут (хост из локальной сети или адрес маршрутизатора) и направляется в цепочку OUTPUT. После цепочки OUTPUT (или FORWARD, если пакет был проходящий) пакет снова сверяется с правилами маршрутизации и отправляется в цепочку POSTROUTING. Может возникнуть резонный вопрос: почему несколько раз пакет проходит через таблицу маршрутизации? (об этом – ниже).

Каждая цепочка, которую проходит пакет состоит из набора таблиц (table) (обозначены овалами). Таблицы в разных цепочках имеют одинаковое наименование, но тем не менее никак между собой не связаны. Например таблица nat в цепочке PREROUTING никак не связана с таблицей nat в цепочке POSTROUTING. Каждая таблица состоит из упорядоченного набора (списка) правил. Каждое правило содержит условие, которому должен соответствовать проходящий пакет и действия к пакету, подходящему данному условию.

Проходя через серию цепочек пакет последовательно проходит каждую таблицу (в указанном на иллюстрации порядке) и в каждой таблице последовательно сверяется с каждым правилом (точнее сказать – с каждым набором условий/критериев в правиле), и если пакет соответствует какому-либо критерию, то выполняется заданное действие над пакетом. При этом, в каждой таблице

(кроме пользовательских) существует заданная по-умолчанию политика. Данная политика определяет действие над пакетом, в случае, если пакет не соответствует ни одному из правил в таблице. Чаще всего – это действие ACCEPT, чтобы принять пакет и передать в следующую таблицу или DROP – чтобы отбросить пакет. В случае, если пакет не был отброшен, он завершает свое путешествие по ядру системы и отправляется в сетевую карту сетевой интерфейс, которая подходит по правилам маршрутизации.

Компоненты

- xtables
  iptables
  ip6tables
  arp_tables
- nf_conntrack
- ebtables

Utils

- iptables/ip6tables/ip(6)tables-(save|restore)
- ipset
- ebtables
- arptables
- firewalld

В основном фильтрация *L3-L4 уровней*

Пакетный фильтр – это набор правил.
Каждое *правило* – это набор таблиц и цепочек.
Каждая *таблиц*а состоит из цепочек.
Каждая *цепочка* – упорядоченный набор правил, просматриваемый начиная с первого.

Каждое *правило* состоит из

- *критерий* срабатывания
- *действие*
  Каждое правило имеет *счетчик* срабатываний.

raw – изначальная обработка, до conntrack
mangle – модификация заголовков и маркировка пакетов
nat – трансляция адресов
filter – фильтры
security – работа с SELinux

min запросов в http пакете – 6. SYN –> ACK, ACK-SYN –>, SYN-ACK

Критерии срабатывания L3

-i/–in-interface Входящий интерфейс
-o/–out-interface Исходящий интерфейс
-s/–source Адрес источника
-d/–destination Адрес назначения
-p/–protocol IP-Протокол (tcp, udp, icmp…)
-f/–fragment Является ли фрагментом (2+ в серии)

Критерии срабатывания L4

TCP/UDP
–sport port[:port] порт или диапазон портов
–dport port[:port] порт или диапазон портов

TCP
–tcp-flags mask flags (SYN,ACK,RST,FIN SYN) Флаги TCP
–syn взведен SYN

ICMP
–icmp-type тип icmp-пакета

Действия

ACCEPT – Принять пакет
DROP – Отбросить пакет
REJECT – Отбросить и сообщить источнику icmp-сообщением
RETURN – Вернуться в вішестоящюю цепочку или применить правило поумолчанию
LOG
chain_name – Перейти в цепочку chain_name

DNAT – Destination NAT
SNAT – Source NAT
MASQUARADE – Source NAT для динамически конфигурируемых интерфейсов
SET – Добавление/Удаление адреса в ipset

Дополнительные модули срабатывания

conntrack/state – критерии срабатывания основанные на состоянии соединения
multiport – критерий срабатывания, позволяющий указать список портов, а не диапазон
iprange – критерий срабатывания, который позволяет указать ip-range вместо cidr-префикса
mark/connmark – основан на маркировке пакета/соединения
set – критерий основанный на ipset
u32 – гибкий критерий, позволяющий работать напрямую с заголовками пакетов и отдельными битами

conntract

Подсистема отслеживания состояний соединений. Базово соединения имеют состояния:
NEW – новое соединение. Отбираются пакеты устанавливающие соединения.
ESTABLISHED – установленное соединение. Отбираются пакеты не !syn/syn+ack, !rst/fin, которые относятся к уже отслеживаемым соединениям
RELATED – относятся к другому, уже установленному соединению (passive ftp, icmp-messages)
INVALID – пакеты принадлежность которых к отслеживаемым соединения установить не удалось

Подсистема очень удобна при невысокой нагрузке. При высокой же требует настройки, иначе может приводить к потере пакетов или связности в целом из-за переполнения таблицы conntrack или слишком большого ее размера.

В HL рекомендуется отключать, так как просмотр таблиц

добавляет время к обработке пакета повышая latency приложения.

В этой системе, размер таблицы поумолчанию, 15тыс соединений:

```
# sysctl -a | grep conntrack
net.nf_conntrack_max = 15624
```

Посмотреть отслеживаемые соединения:

```
# cat /proc/net/nf_conntrack
ipv4     2 tcp      6 431986 ESTABLISHED src=192.168.1.1
dst=192.168.1.126  sport=50838  dport=179  src=192.168.1.126
dst=192.168.1.1  sport=179  dport=50838  [ASSURED]  mark=0
secctx=system_u:object_r:unlabeled_t:s0 zone=0 use=2
ipv4     2 tcp      6 429361 ESTABLISHED src=192.168.113.1
dst=192.168.1.64  sport=58442  dport=22  src=192.168.1.64
dst=192.168.113.1  sport=22  dport=58442  [ASSURED]  mark=0
secctx=system_u:object_r:unlabeled_t:s0 zone=0 use=2
ipv4     2 tcp      6 431999 ESTABLISHED src=192.168.113.1
dst=192.168.113.63  sport=38272  dport=22  src=192.168.113.63
dst=192.168.113.1  sport=22  dport=38272  [ASSURED]  mark=0
secctx=system_u:object_r:unlabeled_t:s0 zone=0 use=2
ipv4     2 tcp      6 429362 ESTABLISHED src=192.168.113.1
dst=192.168.1.126  sport=53900  dport=22  src=192.168.1.126
dst=192.168.113.1  sport=22  dport=53900  [ASSURED]  mark=0
secctx=system_u:object_r:unlabeled_t:s0 zone=0 use=2
ipv4     2 udp      17 20 src=192.168.1.64 dst=195.78.244.34
sport=60897  dport=123  src=195.78.244.34  dst=192.168.113.63
sport=123              dport=60897              mark=0
secctx=system_u:object_r:unlabeled_t:s0 zone=0 use=2
ipv4         2  udp          17  1  src=192.168.113.11
dst=192.168.113.255  sport=64768  dport=1947  [UNREPLIED]
src=192.168.113.255 dst=192.168.113.11 sport=1947 dport=64768
mark=0 secctx=system_u:object_r:unlabeled_t:s0 zone=0 use=2
ipv4     2 tcp      6 431998 ESTABLISHED src=192.168.1.64
dst=192.168.1.1  sport=58718  dport=179  src=192.168.1.1
dst=192.168.1.64  sport=179  dport=58718  [ASSURED]  mark=0
secctx=system_u:object_r:unlabeled_t:s0 zone=0 use=2
```

Таблицы

- raw – изначальная обработка, до contrack

- mangle – модификация заголовков и марктровка пакетов
- nat – трансляция адресов
- filter – фильтры
- security – работа с SELinux

Цепочки

- PREROUTING – **до** принятия решения о маршрутизации
- POSTROUTING – **после** принятия решения о маршрутизации
- OUTPUT – пакеты сгенерированы локальными приложениями
- INPUT – пакеты предназначены локальной системе
- FORWARD – пакеты **проходящие** через систему

1. Таблица raw

Предназначена для базовой обработки пакетов до conntrack, в частности для управления conntrack в отношении некоторых пакетов

Цепочки:

- PREROUTING
- OUTPUT

Действия:

- NOTRACK – отключить conntrack для пакетов попадающих в правило
- CT – настроить работу с модулем conntrack, включает NOTRACK через –notrack
- DROP – отбросить пакет

DDOS трафик фильтровать лучше в raw, так как эта таблица ближе всего к сетевому интрефейсу.

2. Таблица mangle

Предназначена для маркировки и классификации пакетов, модификации заголовков (tos, mss, ttl)

Цепочки:

- PREROUTING
- INPUT
- FORWARD
- OUTPUT
- POSTROUTING

Действия:

- TTL – установить ttl
- MARK/CONNMARK – установить метку (fwmark) пакета/соединения
- CLASSIFY – классифицировать пакет для обработки в шейпере
- TCPMSS – установить TCP Maximum Segment Size (если не работает PMTU disco)

3. Таблица nat

Предназначена для манипуляций с адресами источника/назначения

Цепочки:

- PREROUTING
- INPUT
- OUTPUT
- POSTROUTING

Действия:

- SNAT/MASQUERADE – Source NAT
- DNAT – Destination NAT
- REDIRECT – подмена dst_ip:dst_port на свои собственные (частный случай DNAT)

DNAT – transparent ptroxy

4. Таблица filter

Основная таблица где происходит фильтрация пакетов

Цепочки:

- INPUT
- FORWARD
- OUTPUT

Действия:

- ACCEPT
- REJECT
- DROP

5. Таблица securuty

Предназначена для работы совместнос SELinux

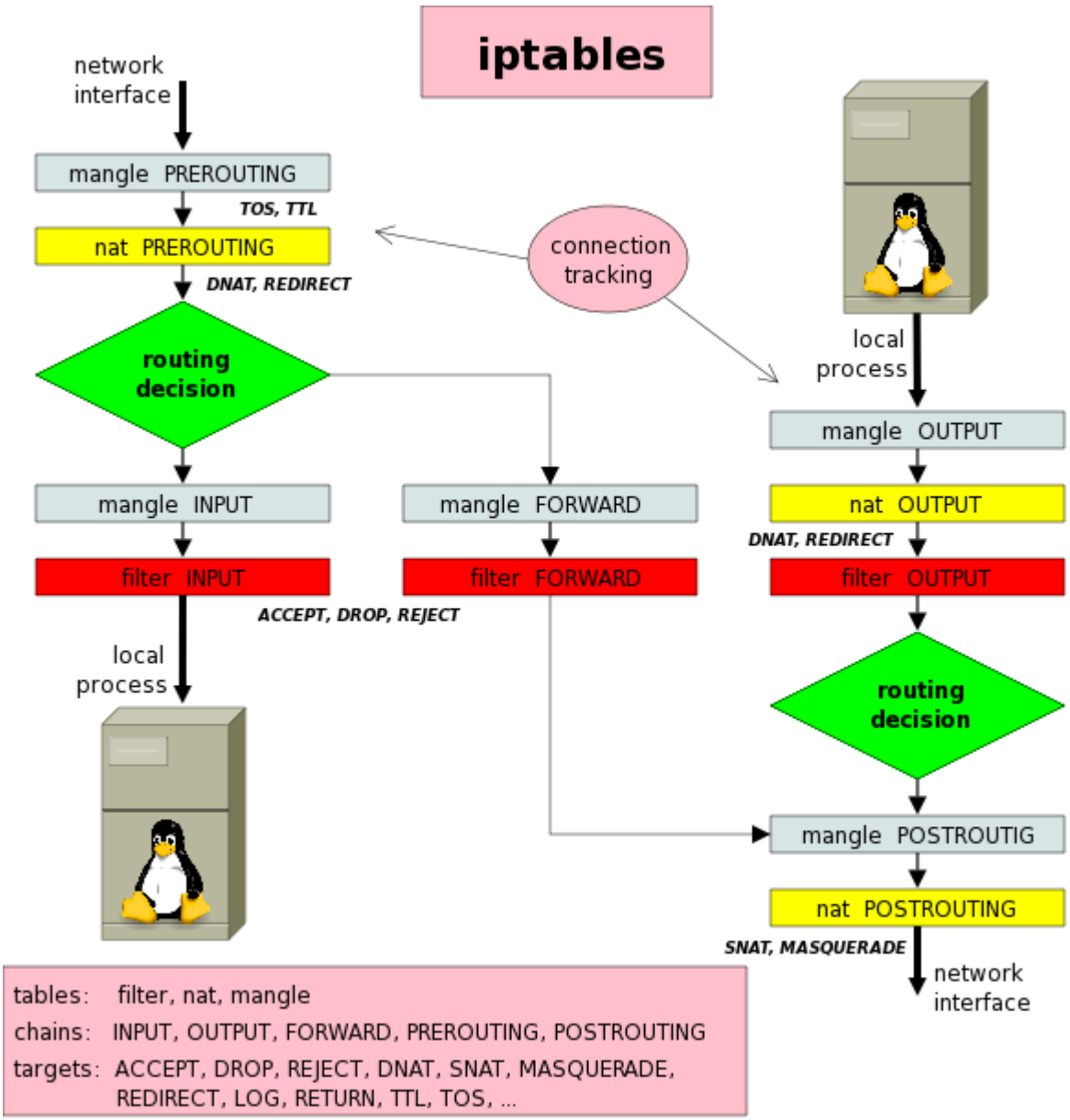
Цепочки:

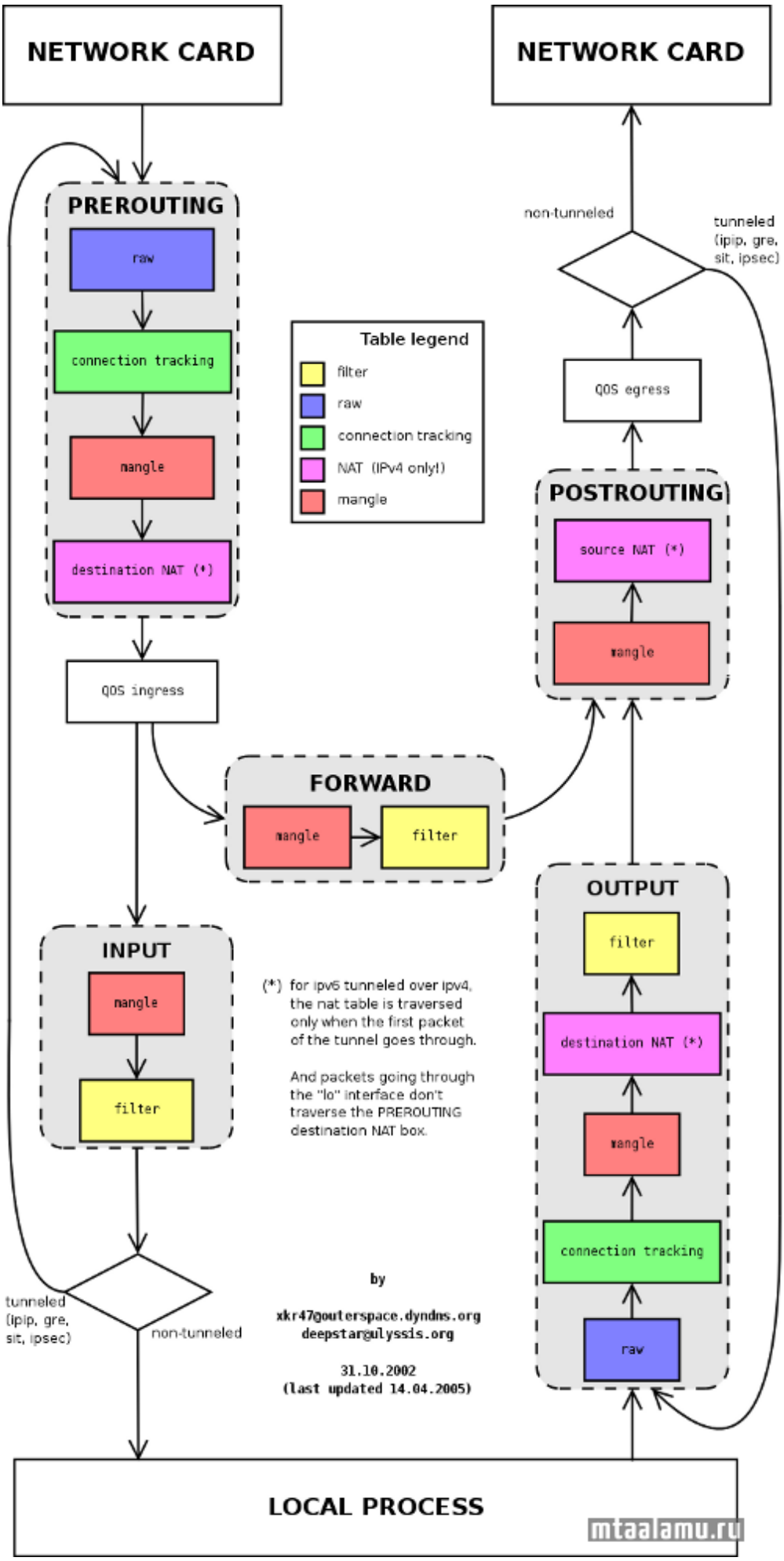
- INPUT
- FORWARD
- OUTPUT

Действия:

- SECMARK/CONNSECMARK — установить SELinux context для пакета/соединения

Еще наглядные схемы работы:


iptables
network interface
mangle PREROUTING
TOS, TTL
nat PREROUTING
DNAT, REDIRECT
routing decision
connection tracking
local process
mangle OUTPUT
nat OUTPUT
DNAT, REDIRECT
filter OUTPUT
routing decision
mangle INPUT
mangle FORWARD
filter INPUT
filter FORWARD
ACCEPT, DROP, REJECT
local process
mangle POSTROUTIG
nat POSTROUTING
SNAT, MASQUERADE
network interface
tables: filter, nat, mangle
chains: INPUT, OUTPUT, FORWARD, PREROUTING, POSTROUTING
targets: ACCEPT, DROP, REJECT, DNAT, SNAT, MASQUERADE, REDIRECT, LOG, RETURN, TTL, TOS, ...


NETWORK CARD
NETWORK CARD
PREROUTING
raw
connection tracking
mangle
destination NAT (*)
QOS ingress
non-tunneled
tunneled
(ipip, gre,
sit, ipsec)
QOS egress
POSTROUTING
source NAT (*)
mangle
Table legend
filter
raw
connection tracking
NAT (IPv4 only!)
mangle
FORWARD
mangle
filter
INPUT
mangle
filter
OUTPUT
filter
destination NAT (*)
mangle
connection tracking
raw
(*) for ipv6 tunneled over ipv4,
the nat table is traversed
only when the first packet
of the tunnel goes through.
And packets going through
the "lo" interface don't
traverse the PREROUTING
destination NAT box.
tunneled
(ipip, gre,
sit, ipsec)
non-tunneled
by
xkr47@outerspace.dyndns.org
deepstar@ulyssis.org
31.10.2002
(last updated 14.04.2005)
LOCAL PROCESS
mtaalamu.ru