Хардуерна маршрутизация с Marvell switch chip

MikroTik Net Camp 2023, с. Емен

Петър Димитров

Съдържание

- Начини за обработка на трафика:
  - Slow Path
  - Fast Path
  - Fast Track
  - HW Offload
- HW Offload
  - конфигурация
  - особености
  - ограничения
  - наблюдение
  - демонстрация

За мен - Петър Димитров

- MikroTik Trainer: от 2013 г.
- Ubiquiti Trainer: от 2018 г.
- Предлагани обучения:
  - Въведение в компютърните мрежи
  - Мониторинг с The Dude
  - MTCNA, MTCSWE, MTCRE, MTCINE, MTCWE, MTCEWE, MTCTCE, MTCUME, MTCSE, MTCIPv6E
  - UBWS, UBWA, UBRSS, UBRSA, UNS, UEWA

Slow Path

- Трафика преминава обработка по пълната packet flow диаграма
- Могат да се използват всички възможности на RouterOS
- Няма ограничения по отношение на конфигурация/изисквания към драйвери на интерфейси
- Натоварва CPU и внася (сравнително) голяма латентност

Fast Path

- Трафика не се обработва от Linux kernel-а, препраща се на ниво драйвери на интерфейси
- Изключително ограничени възможности за обработка (на практика без контрол/филтриране на трафика)
- Изисква поддръжка от драйверите на интерфейсите
- Натоварването на CPU и внасяната латентност са малки

Fast Track

- Fasttrack = Fast Path + Connection Tracking
- Connection Tracking системата прекарва голяма част от трафика на връзките, маркирани за fasttrack, по Fast Path
- Пълна съвместимост с функционалностите, осигурявани от firewall
  - Това не означава съвместимост с например опашки или IPsec
- Натоварването на CPU и внасяната латентност са (сравнително) малки

HW Offload

- Трафика не се обработва от CPU/Linux kernel
- Възможна е маршрутизация (различните чипове могат да поемат различен брой маршрути)
- **Само за някои чипове** е възможен Offload на ограничен брой връзки, маркирани за fasttrack (ползващи или не NAT)
- Изисква Marvell switch chip

HW Offload конфигурация

- Активирането на HW Offload става чрез указване на l3-hw-offloading=yes в меню /interface/ethernet/switch
- Пакетите, за които и входящия, и изходящия порт са с включен l3-hw-offloading=yes в меню /interface/ethernet/switch/port (по пордазбиране е включен на всички портове), ще бъдат препратени само от switch chip-а (и няма да минат през CPU)
- Филтрирането на тези пакети може да се реализира само чрез /interface/ethernet/switch/rule

HW Offload особености

- За да обработите чрез CPU (например /ip/firewall/filter) определен трафик, трябва да е изключен l3-hw-offloading за входящия или изходящия порт
- Достъп по MAC адрес/чрез RoMON за някои чипове изискват ACL за пренасочване на конкретните фреймове към CPU

```
/interface ethernet switch rule
add switch=switch1 ports=ether2 protocol=udp dst-port=20561 redirect-to-cpu=yes
add switch=switch1 ports=ether2 mac-protocol=0x88BF redirect-to-cpu=yes
```

Ограничения

Модел	Switch Chip	Брой IPv4 маршрути
CRS318-1Fi-15Fr-2S	98DX224S	13312
CRS310-1G-5S-4S+	98DX226S	13312
CRS318-16P-2S+	98DX226S	13312
CRS305-1G-4S+	98DX3236	13312
CRS326-24G-2S+	98DX3236	13312
CRS328-24P-4S+	98DX3236	13312
CRS328-4C-20S-4S+	98DX3236	13312

Ограничения

Модел	Switch Chip	Брой IPv4 маршрути	FastTrack връзки	NAT записи
CRS504-4XQ	98DX4310	60K - 120K	4.5K	4K
CRS510-8XS-2XQ	98DX4310	60K - 120K	4.5K	4K
CRS354-48?-4S+2Q+	98DX3257	16K - 36K	2.25K	2.25K
CRS309-1G-8S+	98DX8208	16K - 36K	4.5K	3.9K
CRS312-4C+8XG	98DX8212	16K - 36K	2.25K	2.25K
CRS317-1G-16S+	98DX8216	120K - 240K	4.5K	4K
CRS326-24S+2Q+	98DX8332	16K - 36K	2.25K	2.25K
CRS518-16XS-2XQ	98DX8525	60K - 120K	4.5K	4K

Ограничения

Модел	Switch Chip	Брой IPv4 маршрути	FastTrack връзки	NAT записи
CCR2116-12G-4S+	98DX3255	16K - 36K	2.25K	2.25K
CCR2216-1G-12XS-2XQ	98DX8525	60K - 120K	4.5K	4K

Ограничен брой маршрути

- IPv4 и IPv6 маршрутите споделят обща (ограничена) памет в switch chip-а
- При запълване на паметта на switch chip-а, само част от трафика се маршрутизира хардуерно, останалите пакети се обработват от CPU
- По-специфичните маршрути (с по-дълъг prefix length) приоритетно се маршрутизират хардуерно
- **Важно!** Флаг H в маршрутната таблица не означава, че маршрута е с HW Offload, а само, че е възможно да бъде избран за HW Offload

Ограничен брой маршрути

- За управление кои маршрути да бъдат или да не бъдат маршрутизирани хардуерно, може да се забрани използването на HW Offload за конкретен маршрут чрез опцията suppress-hw-offload=yes
- За маршрути от динамични маршрутизиращи протоколи опцията се управлява с Routing Filters, например:

```
/routing/filter/rule/add chain=isp1-in rule=\
"if (bgp-communities includes 100:359) {accept} else {set suppress-hw-offload yes; accept}"
```

Ограничен брой връзки

- Fasttrack връзките с HW Offload и NAT записите (за тези от тях, за които е необходимо), споделят обща (ограничена) памет с ACL rules
- При недостатъчна памет в switch chip-а, HW Offload ползват с предимство връзките с повече трафик
- По оптимално е използването на ACL rules (когато е възможно) - с едно правило можете да филтрирате трафик, който иначе би породил множество връзки (заемащи много повече памет)

Наблюдение на HW Offload

- Текущото състояние и статистики могат да се наблюдават чрез /interface/ethernet/switch/l3hw-settings/monitor (или с разширения вариант на инструмента /interface/ethernet/switch/l3hw-settings/advanced/monitor)
  - ipv4-routes-total / ipv4-routes-hw / ipv4-routes-cpu дават информация за общия брой маршрути, колко от тях са с HW Offload и колко се обработват от CPU
  - ipv4-shortest-hw-prefix дава информация за prefix length на най-малко специфичния маршрут с HW Offload

Лабораторна постановка за демонстрация

Генериране/обявяване
на ~12 000 маршрута

Генериране
на трафик

CRS318-1Fi-15Fr-2S
(Marvell 98DX224S)

Резултати от тестовете

	Входящ трафик	Изходящ трафик	Натоварване CPU
Slow Path 1500 bytes	950 Mbps	950 Mbps	95%
Slow Path 64 bytes	770 Mbps*	90 Mbps	100%
Fast Path 1500 bytes	950 Mbps	950 Mbps	78%
Fast Path 64 bytes	770 Mbps*	137 Mbps	100%
HW routing 1500 bytes	950 Mbps	950 Mbps	2%
HW routing 64 bytes	770 Mbps*	770 Mbps	2%

* На машината, генерираща трафик, беше заложен 950 Mbps stream, но CPU-то се натовари до 100% и не успя да генерира повече от 770 Mbps.

Обобщение

Въпроси

Благодаря за вниманието!