Intel DPDK в решениях для противодействия DDoS-атакам от 40 Гбит/с

Дмитрий Козлюк, ведущий разработчик по сетевым решениям
АО «БИФИТ»

О чем доклад?

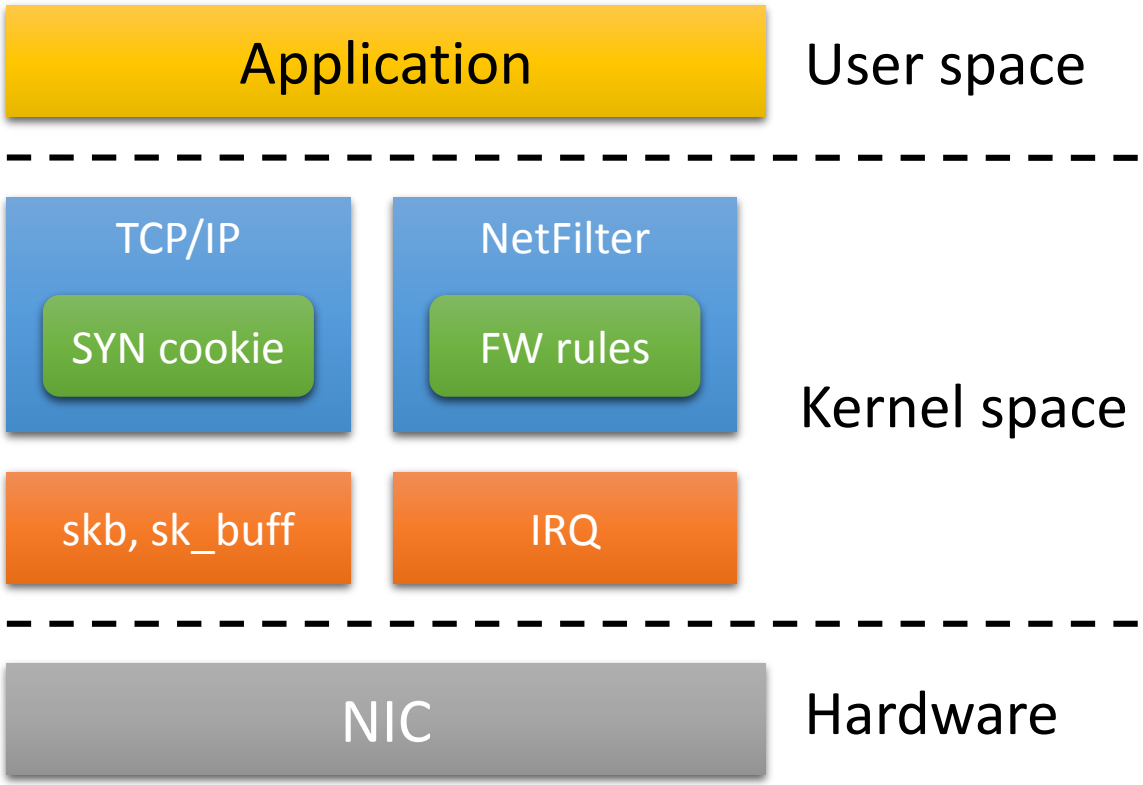
- Intel DPDK
- Нюансы из практики
  (на примере DDoS).
- DPDK и криптоускорители
- Виртуализация

Нужны программные решения на открытой платформе

- Себестоимость разработки
- Адаптация к новым реалиям
- Время обновления
- Масштабируемость
- Vendor lock
- Доверие производителю

Защита средствами Linux

- Предел — 5 Mpps. (Атака в 30 Mpps стоит 10 USD.)
- Любой ввод/вывод — это переключения контекстов и прерывания.
- Защиту в ядре трудно писать.

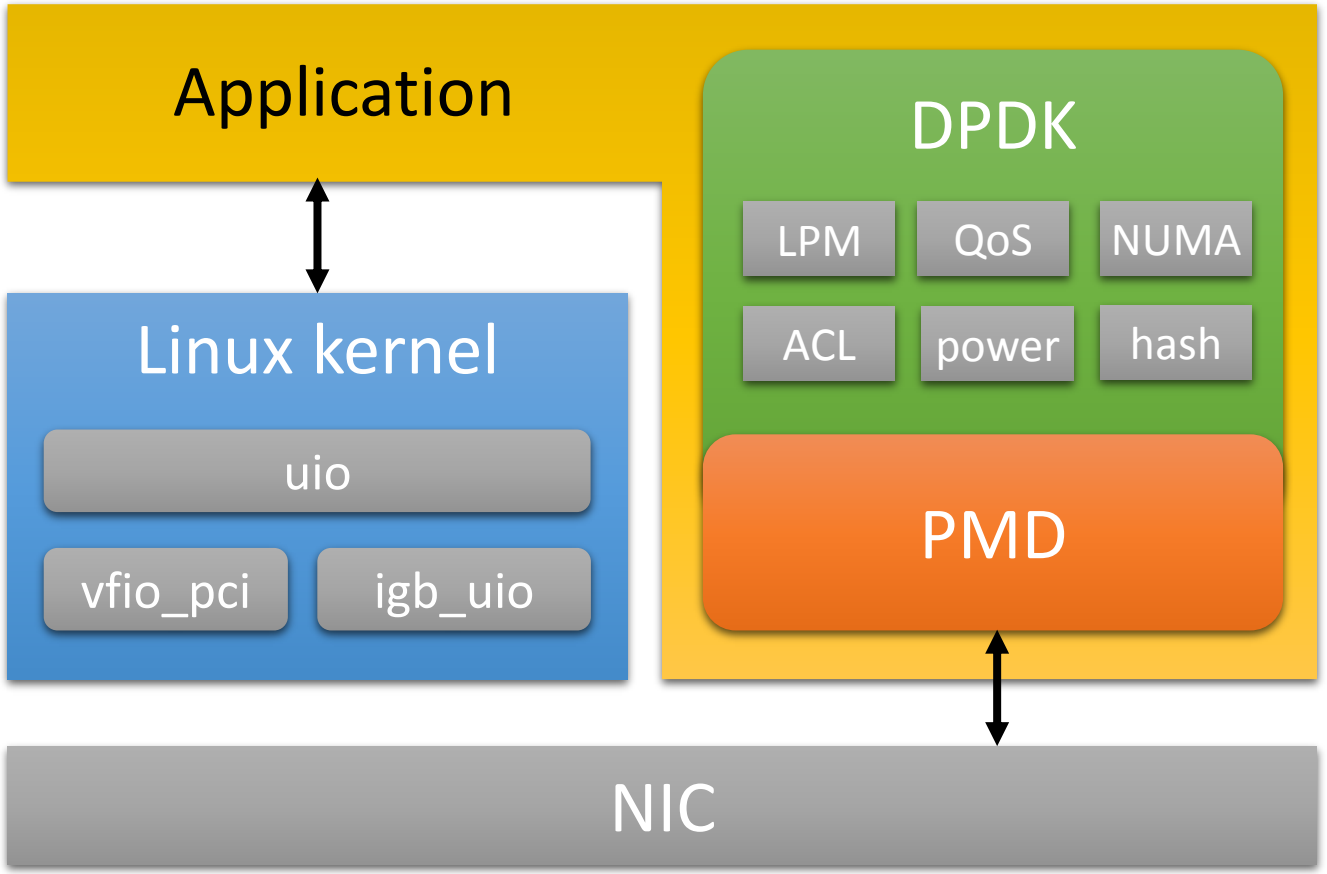


Intel DPDK

- Data Plane Development Kit — набор средств для высокопроизводительной обработки больших объёмов сетевого трафика.
- Первый релиз в 2010 (6WIND).
- Open-source с 2013 г. (покупка Intel):
  - десятки патчей в сутки;
  - общение с разработчиками в mailing list;
  - собственные правки, патчи «из будущего».
- Написано в Intel и для Intel.
  - Более 40 сетевых карт 11 производителей + виртуальные.
  - Linux/BSD, GCC/ICC/Clang.
  - x86_64, arm, power8

Intel DPDK: Poll Mode Driver

Poll mode driver (PMD)
работает с NIC
из userspace
без участия ядра
и прерываний.



Что дает DPDK существующему ПО?

Приложение	Без DPDK	DPDK	Ускорение
Nginx	6K RPS	18K RPS	3,0
Open vSwitch	1,2M pps	12M pps	10,0
Memcached	1,2M pps	4M pps	3,3

Intel DPDK: настройка NIC

Единый интерфейс для настройки из приложения:

- очередей RX/TX;
- receive side scaling (RSS);
- FlowDirector;
- offloading.

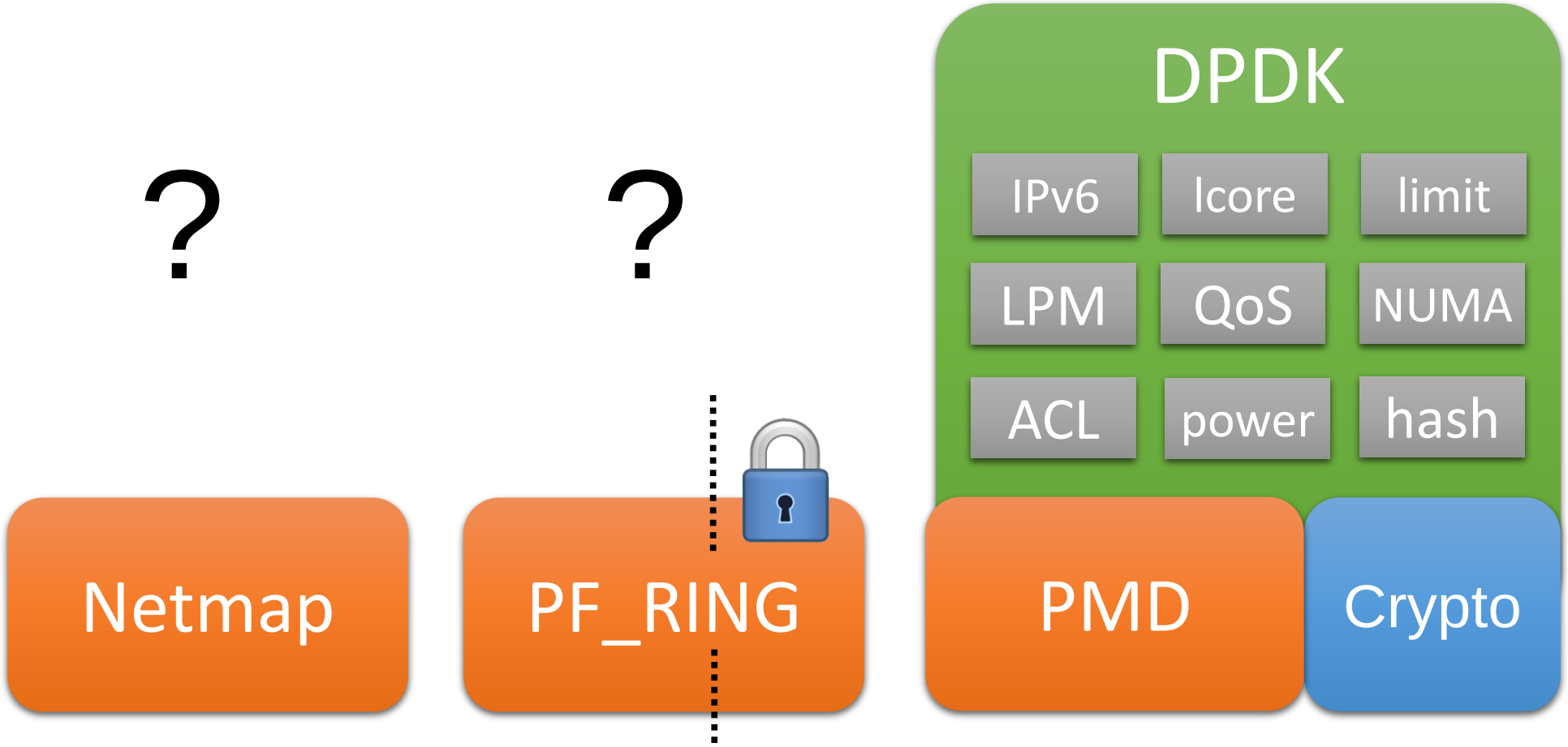
Библиотеки DPDK

- LPM (longest prefix match) для маршрутизации
  - 12 Mpps / ядро
- ACL — advanced classification library
  - FlowSpec со 100 000 правил на 40 Гбит/с
- IPv6 (LPM, фрагментация, хэши)
- Quality of Service (QoS)
  - Политики и группы политик на уровне DPDK
- Token bucket (лимиты)
- IPC
- Хэш-таблицы, деревья поиска, кольцевые буферы
  - Blacklist на всем пространстве IPv4

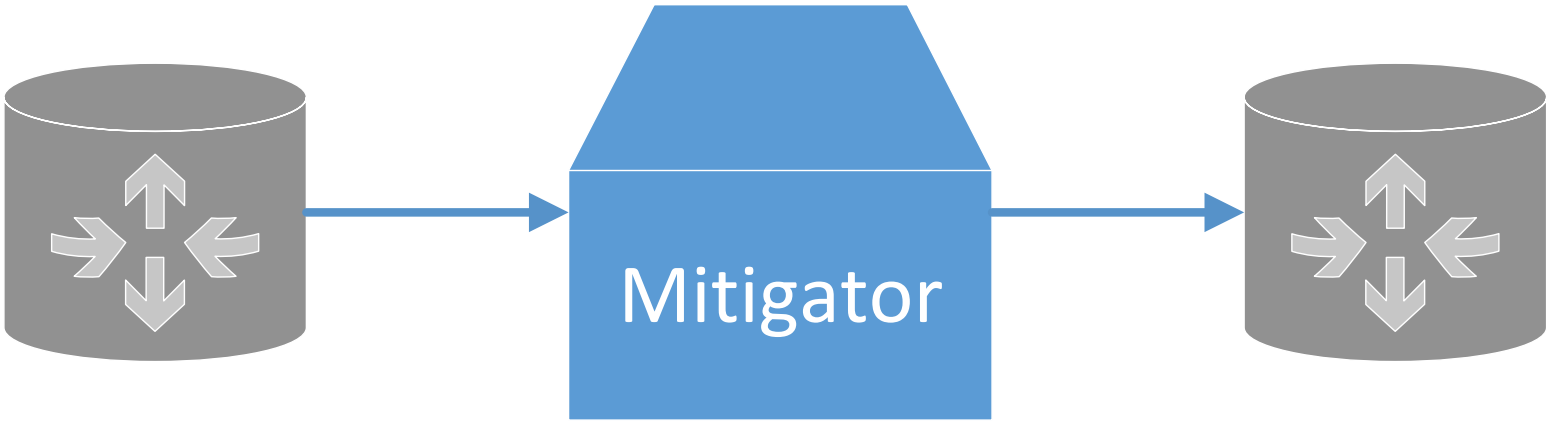
Чем DPDK не является?

- Не framework, а kit.
- Нет TCP/IP стека:
  - универсального стека не бывает;
  - есть библиотеки для написания нужного;
  - mTCP, lwIP, picotcp, libuinet, NUSE, opendp, seastar, ...
- Не распределяет нагрузку автоматически.

Альтернативы DPDK



Пример: Mitigator — защита от DDoS-атак



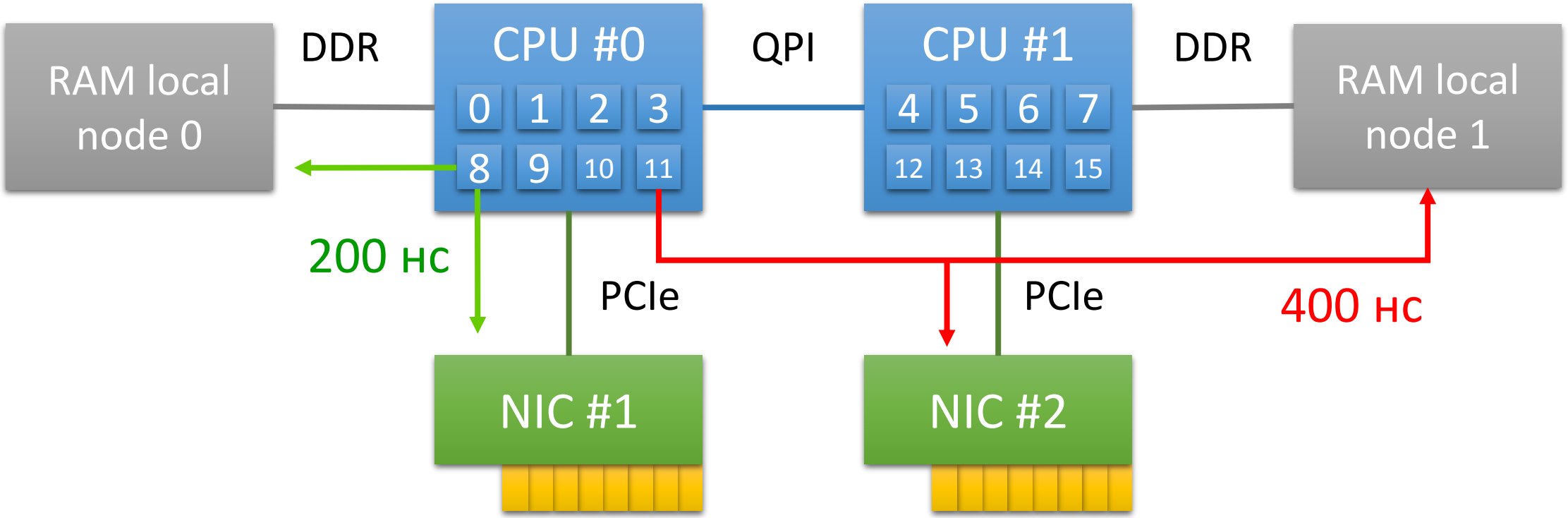
- L2 Bridge
- Поддерживает асимметрию трафика.
- REST API
- BGP FlowSpec

Атаки и контрмеры

- IP/TCP/UDP/ICMP Fragmentation Flood
- ICMP Flood
- TCP SYN Flood
- TCP SYN-ACK Flood
- TCP ACK Flood
- TCP RST/FIN Flood
- HTTP Flood
- Slow HTTP attacks
- DNS Flood
- VoIP Flood
- SSL renegotiation
- Amplification Attacks (DNS, NTP, SSDP, …)
- …

- Обработка фрагментированных IP пакетов
- Валидация пакетов
- Черный список по исходящим IPv4 адресам
- Фильтрация по странам
- Фильтрация по правилам (ACL)
- Защита от TCP Flood
- Защита от HTTP Flood
- Защита от DNS Flood
- Фильтрация по регулярным выражениям (L4 payload)
- Блокировка при превышении порогов
- Ограничение полосы (rate-limit)
- Ограничение трафика на IP адрес получателя
- Отправка BGP FlowSpec правил
- Защита игровых серверов Valve

Non-Uniform Memory Access (NUMA)



Пример распределения нагрузки

Генератор трафика Warp17, 2 сетевых порта, 32 логических ядра.

- Определяем состав узлов NUMA:
  - узел 0: порт 0 — ядра 0—7, 18—23
  - узел 1: порт 1 — ядра 8—16, 24—31
- Распределяем ядра с учетом NUMA:

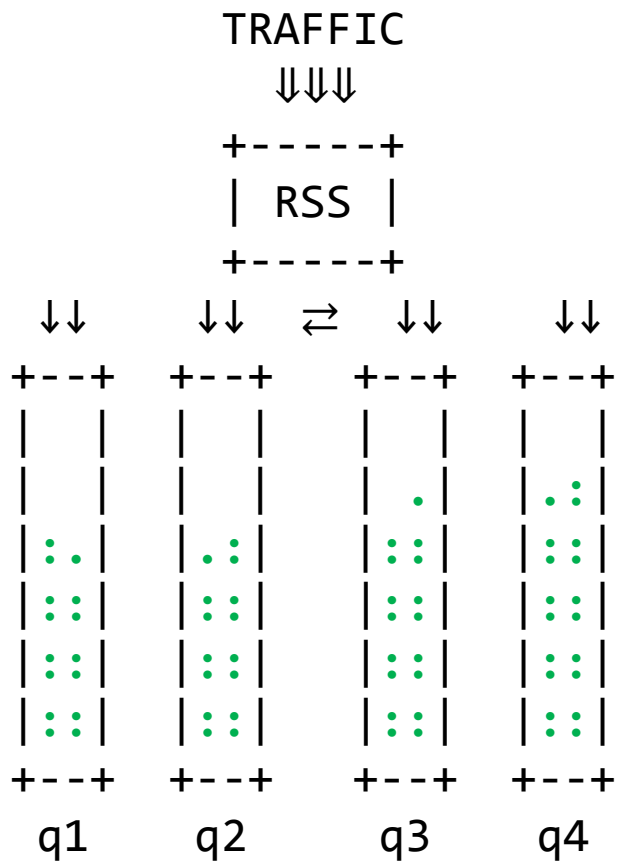
Задача	Порт	Ядра	Маска (ядро = бит)
Управление	—	0, 1	`00 00 00 03`
Порт 0	0	2—7, 18—23	`00 FC 00 FC`
Порт 1	1	8—16, 26—31	`FC 00 FF 00`

- Итого занято 28 из 32 ядер (FC FC FF FF)

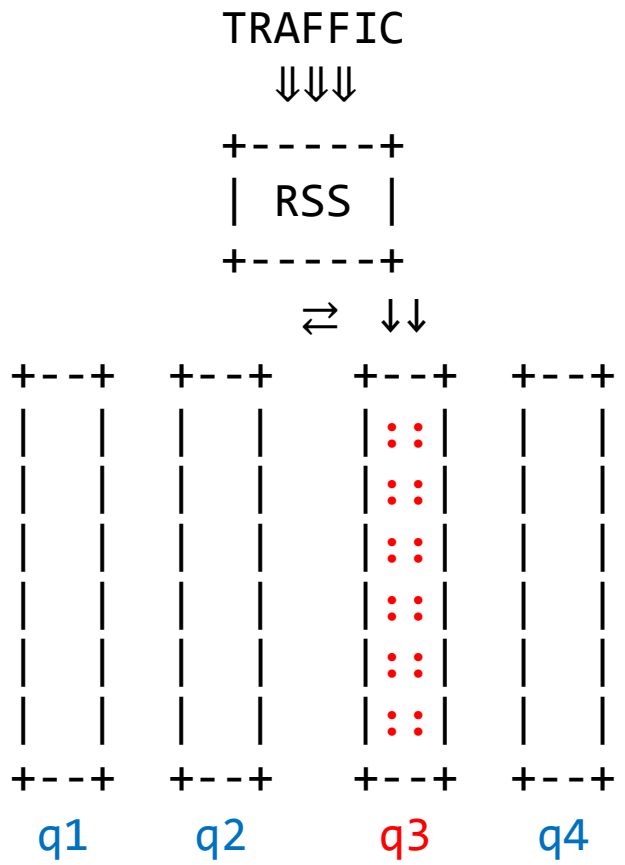
  В: почему простаивают ядра 16 и 17, 24 и 25?

  О: hyper-threading: физические ядра уже загружены логическими 0, 1, 8 и 9.

Receive Side Scaling (RSS)



Атака на RSS



Offloading

- DPDK позволяет управлять TX/RX offloading для каждого порта из приложения.
- Включённый offloading не дает использовать simple path и векторную обработку (на картах Intel).
- Генерация трафика: 1 ядро, 1 порт 10 Гбит/с (line rate 14,88 Mpps):

Path	Offloading	Скорость, Mpps
simple	no	7,00 (100%)
full	no	6,61 (94%)
full	yes	7,74 (110%)

- Не выгодно, если пакеты не меняются (редко меняются) при обработке.

Mitigator vs 40 Gbps (59,52 Mpps)

Атака	Защита	Производительность
SYN Flood	TCPFloodProtection	59 Mpps
DNS Flood	DNSFloodProtection	line-rate
IP Spoofing Flood	Global IPv4 Blacklist	line-rate
TCP Random Spoof	TCP Flood Protection	line-rate
DNS + SNMP Amp.	Advanced ACL 200K rules	line-rate
Random trash 64B	RegExFilter 20K signatures	line-rate
Random trash 128B	RegExFilter 20K signatures	line-rate

Pktgen

Генератор L3—L4 трафика с любыми адресами и флагами.

- Внутренняя разработка BIFIT для тестирования своих решений.
- ICMP/SYN/.../DNS-flood, воспроизведение атак PCAP
- Stateless (не устанавливает соединения)
- 4 × 10 Гбит/с, 10 ядер: line rate (60 Mpps)

Warp17

Генератор L4—L7 трафика с собственным стеком TCP/IP (для тестирования).

- Juniper Networks, open source, 2016 г.
- TCP и UDP со случайными данными, простой HTTP.

DPDK и криптография

- Архитектурные параллели:
  - сетевая карта — криптоускоритель;
  - отправка пакетов — запросы на выполнение операций;
  - прием пакетов — забор результатов операций.
- Та же проблема: доступ к ускорителю через ОС.
- То же решение: poll mode driver.
- То что нужно для DPI и VPN.

DPDK Crypto PMD API

- Единый асинхронный интерфейс к разным устройствам.
- Сейчас: симметричное шифрование, цифровая подпись, аутентификация.
  - Подпись за одну операцию, если поддерживает ускоритель.
  - В разработке: асимметричные алгоритмы, сжатие данных.
- Управление памятью через имеющийся DPDK API.
  - **Можно считать пакет из сети и передать на расшифровку — без копирования.**
  - Не зависит от сетевой части DPDK.

Оборудование

- Intel QuickAssist (QAT, чипы DH89xxCC).
  - Нужен специальный драйвер (отдельно от DPDK);
  - даже для сборки нужна машина с QAT.
- Расширения в процессорах Intel:
  - AES NI
  - **`libsso`**: KASUMI, SNOW 3G, ZUC, ...
- ARMv8
- **Балансировщик нагрузки на криптоускорители.**

QuickAssist & OpenSSL

На примере операции AES-128-CBC-HMAC-SHA1.

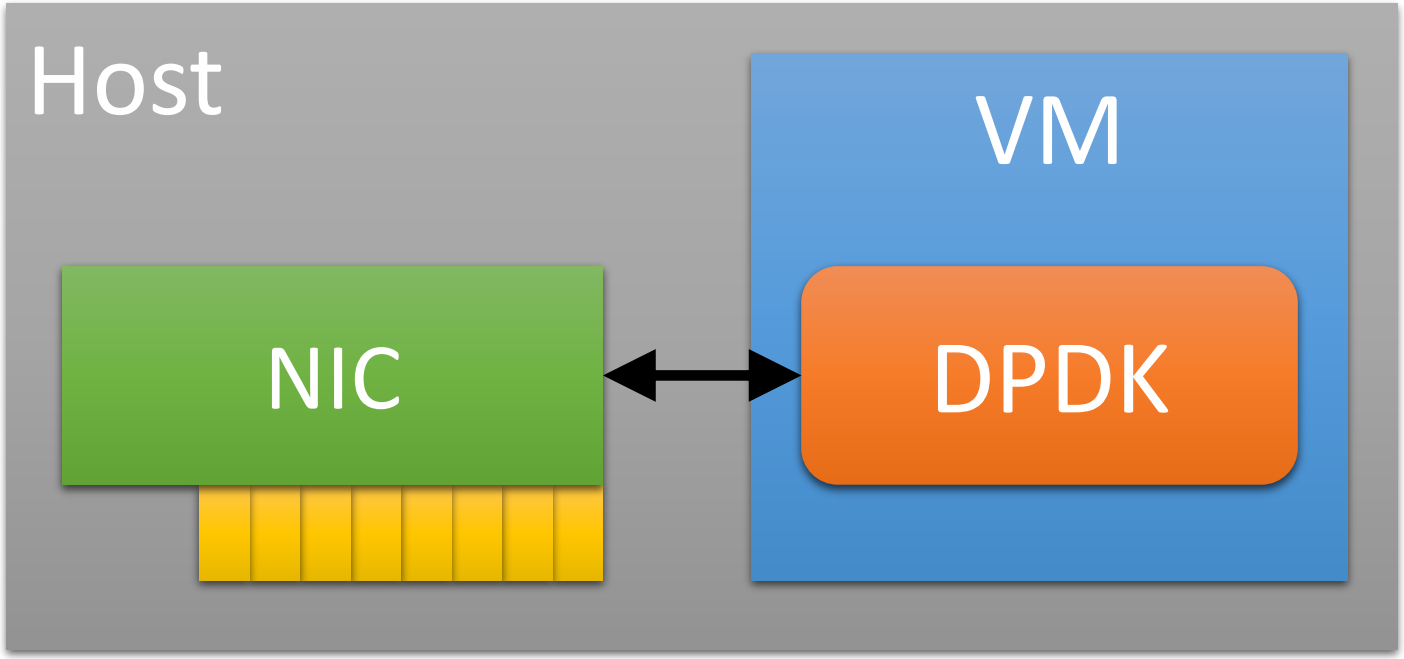
Размер запроса, байт	OpenSSL, Гбит/с	OpenSSL + QuickAssist, Гбит/с
64	0,01	0,43
1024	0,12	6,85
4096	0,49	25,03
16384	1,83	50,12

Поддержка виртуализации

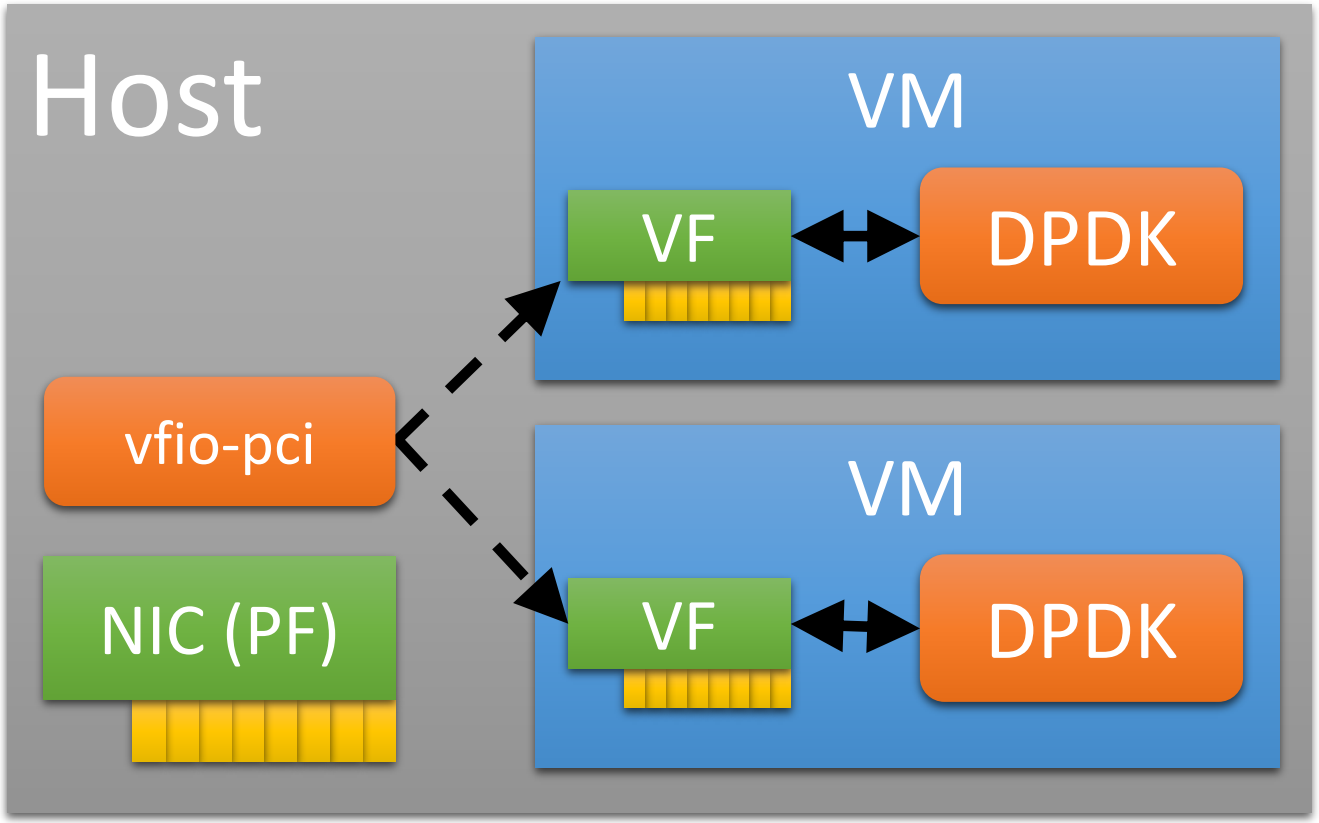
Поддерживаются «из коробки»:

- KVM: virtio
- VMware ESX: Paravirtual VMXNET3
- Xen: paravirtual NIC

Доставка трафика в ВМ: PCI Passthrough

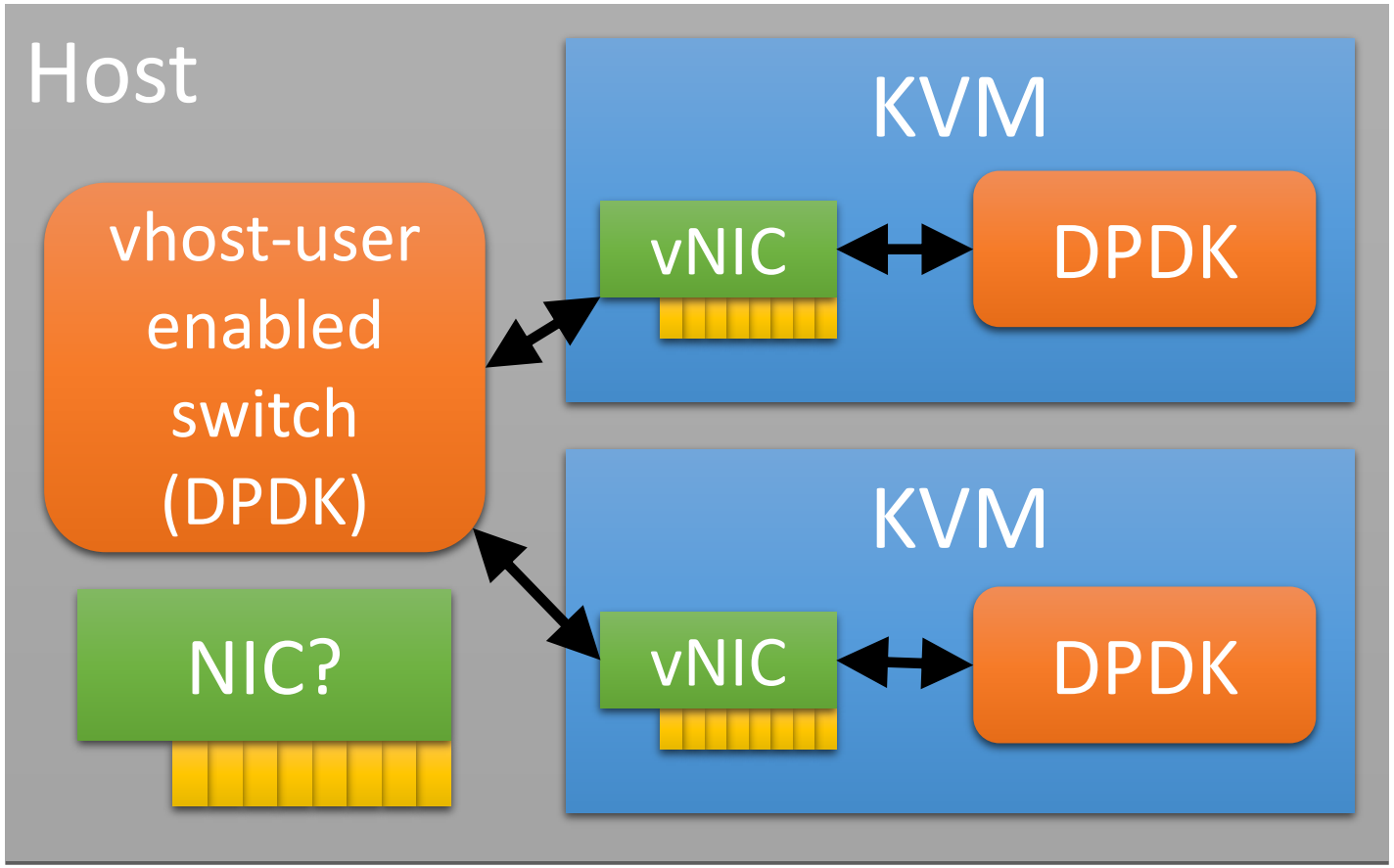


Доставка трафика в BM: SR-IOV

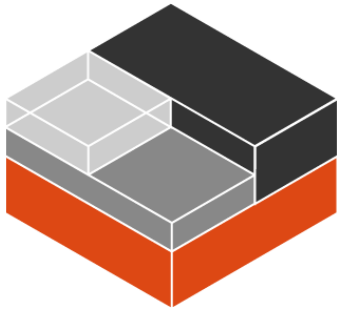


Single-Root Input/Ouput Virtualization

Доставка трафика в BM: vhost-user



LXC



- Приложение с DPDK — обычное Linux-приложение.
- RW-доступ к **`/dev/uio*`** или **`/dev/vfio/*`**:

```
lxc.cgroup.devices.allow = c 247:0 rwm
lxc.cgroup.devices.allow = c 247:1 rwm
```

- Явное указание приложению количества доступной памяти:
  - **`lxc.cgroup.hugetlb.1G.limit_in_bytes = 8589934592`**

```
$ dpdk-app [...] -m 8192
```

- Unprivileged containers.

Выводы

- DPDK позволяет строить быстрые, чисто программные защиты на оборудовании общего назначения.
- Доступ к сетевой карте перестает быть «узким местом».
  - Им становится: ЦП, шина PCI, память, задержки в сети
  - Warp17: 1,8 млн. TCP-сеансов в секунду, но только back-to-back.
- DPDK годится не только для сетевых решений.
- Open-source дает преимущество.

Что почитать?

- DPDK:
  - сайт, документация, mailing list: http://dpdk.org/
  - DPDK Summit: https://dpdksummit.com/
- TCP-стеки:
  - mTCP: https://github.com/eunyoung14/mtcp
  - lwIP: http://savannah.nongnu.org/projects/lwip
  - picotcp: http://www.picotcp.com
  - libuinet: https://github.com/pkelsey/libuinet
  - NUSE: https://github.com/libos-nuse/net-next-nuse
  - opendp: https://github.com/opendp/dpdk-ans
  - seastar: https://github.com/scylladb/seastar

Что почитать?

- Нюансы DPDK:
  - опыт Brocade vRouter: https://events.linuxfoundation.org/sites/events/files/slides/DPDK-Performance.pdf
  - рекомендации Intel: https://software.intel.com/en-us/articles/dpdk-performance-optimization-guidelines-white-paper
- Криптография:
  - OpenSSL и QuickAssist: http://www.intel.com/content/dam/www/public/us/en/documents/solution-briefs/accelerating-openssl-brief
  - DPDK PMD: https://www.slideshare.net/harryvanhaaren/symmetric-crypto-for-dpdk-declan-doherty
  - развитие DPDK PMD: https://dpdksummit.com/Archive/pdf/2016Userspace/Day01-Session06-Userspace2016.pdf

Спасибо за внимание!

Дмитрий Александрович Козлюк

kozlyuk@bifit.com

Mitigator.ru