ИТМО

Технологии для оптимизации сетевого стека в системах хранения данных

Автор Михаил Кудряшов
Студент группы К4211с

Зачем нужно оптимизировать сетевой стек?

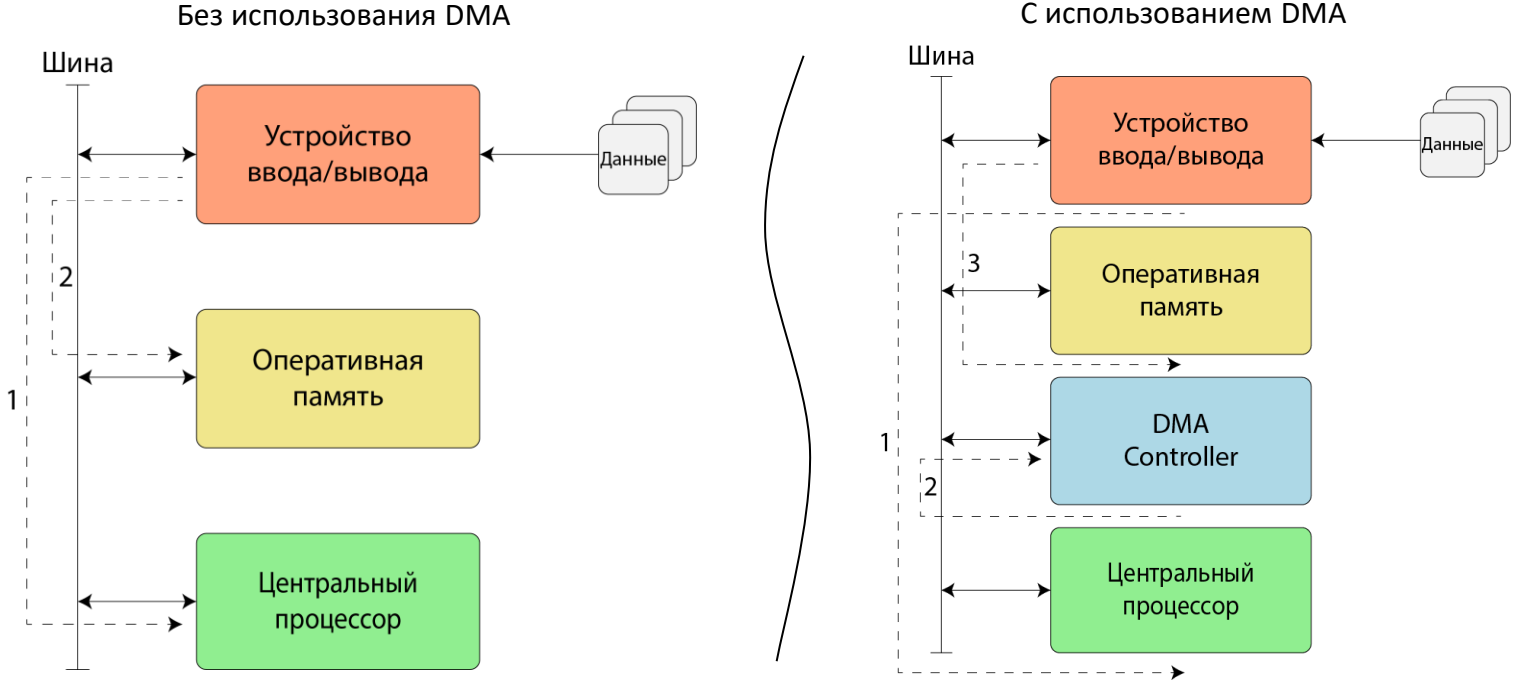
- Увеличить пропускную способность;
- Уменьшить задержки;
- Уменьшить нагрузку на CPU.

Основные технологии

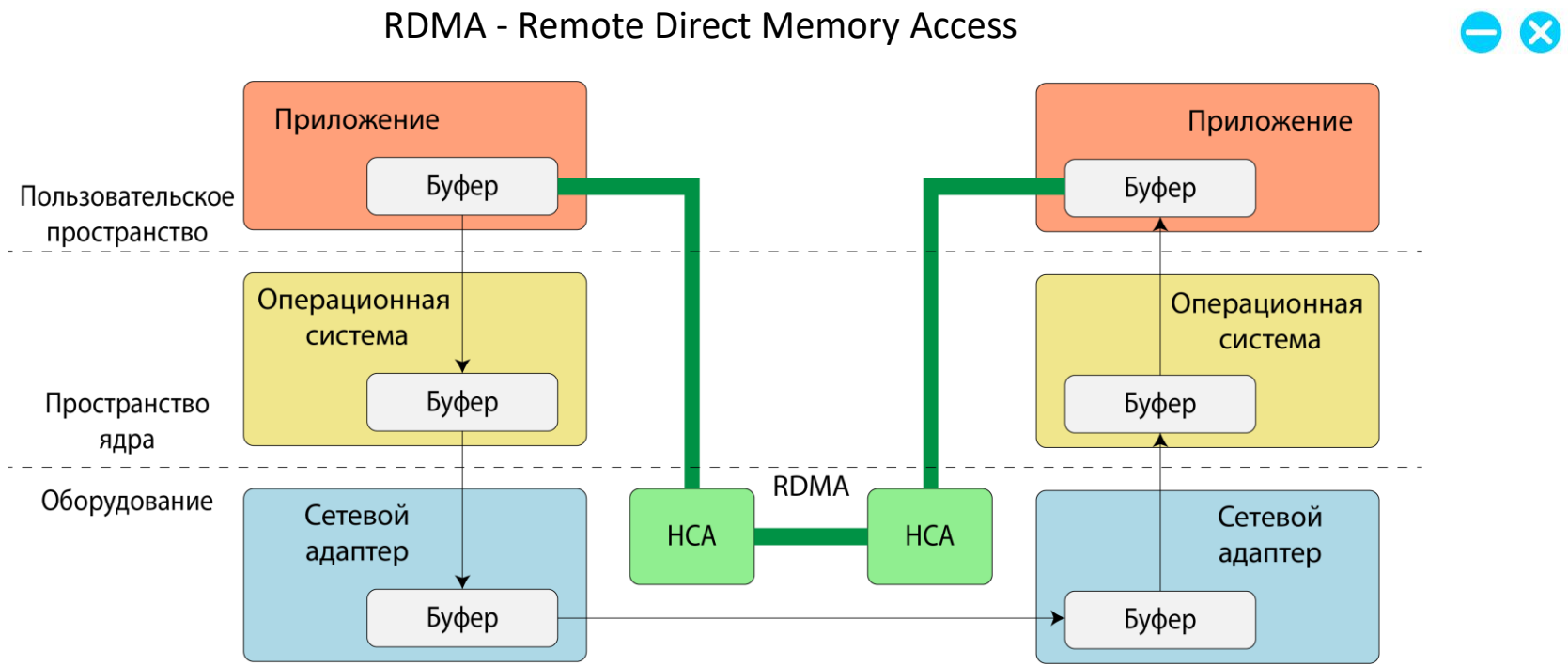
- RDMA;
- DPDK;
- SPDK;
- TCP Offload Engine;
- SmartNIC/DPU.

RDMA

DMA – Direct Memory Access



RDMA


RDMA - Remote Direct Memory Access
Приложение
Буфер
Пользовательское пространство
Операционная система
Буфер
Пространство ядра
Оборудование
Сетевой адаптер
Буфер
HCA
RDMA
HCA
Приложение
Буфер
Операционная система
Буфер
Сетевой адаптер
Буфер

RDMA

InfiniBand

- Требует проприетарных коммутаторов, кабелей, сетевых карт;
- Очень большая пропускная способность (до 800 Гбит/с);
- Очень низкая задержка (1-2 мкс).

RoCE

- 1 версия работает на стандартных коммутаторах, 2 версия требует поддержки Lossles Ethernet (не все коммутаторы её поддерживают);
- Большая пропускная способность (до 400 Гбит/с);
- Низкая задержка (2-5 мкс).

iWARP

- Работает на стандартных коммутаторах;
- Хорошая пропускная способность (до 100 Гбит/с);
- Небольшие задержки (10-30 мкс).

DPDK и SPDK

DPDK - Data Plane Development Kit - набор драйверов и библиотек.

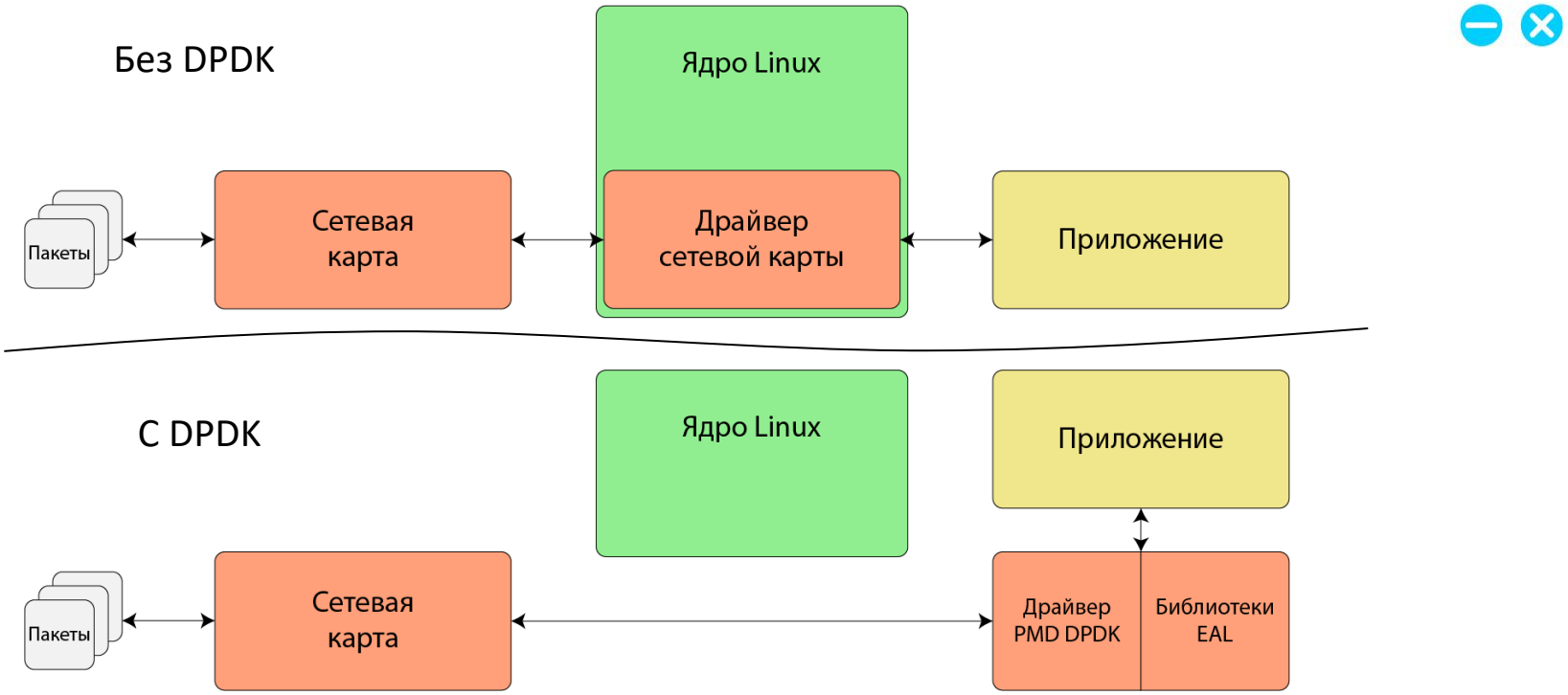
- Увеличивает пропускную способность;
- Уменьшает задержки.

SPDK - Storage Performance Development Kit - аналог DPDK для систем хранения данных.

Обе технологии:

- Используют обход ядра (kernel bypass);
- Используют PMD (Poll Mode Driver).

DPDK и SPDK


Без DPDK
Ядро Linux
Пакеты
Сетевая карта
Драйвер сетевой карты
Приложение
С DPDK
Ядро Linux
Приложение
Пакеты
Сетевая карта
Драйвер PMD DPDK
Библиотеки EAL

Другие технологии

TCP Offload Engine (TOE)

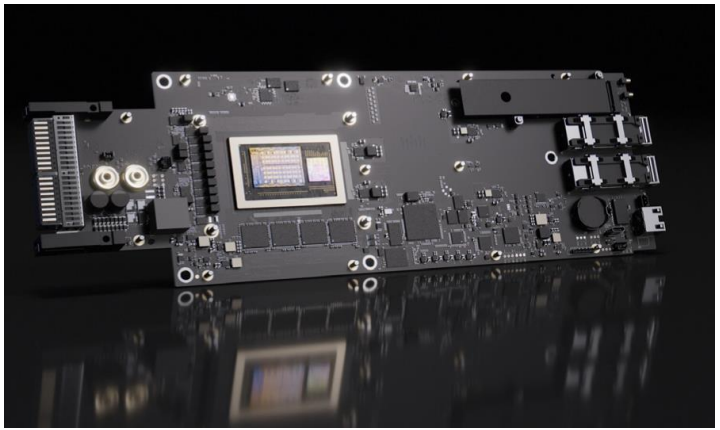
- Разгружает CPU;
- Применяются как правило в высокоскоростных адаптерах.

SmartNIC/DPU

- «Умные» сетевые карты;
- Разгружают CPU;
- Data Processing Unit.

Примеры SmartNIC/DPU

Cisco Nexus SmartNIC



NVIDIA BlueField 4
(DPU)

Практическое применение технологий

1. Виртуальные коммутаторы.

В облачных провайдерах сетевой трафик между виртуальными машинами внутри одного физического сервера проходит через виртуальный коммутатор (vSwitch). Обработка такого трафика потребляет определённый процент вычислительных мощностей.

Для уменьшения этого процента (или полного исключения) мы можем использовать **SmartNIC/DPU** и **DPDK**. Благодаря этому, снизится нагрузка на процессор и освободившиеся вычислительные мощности могут быть проданы клиентам.

2. Высокопроизводительные системы хранения, т.е. доступ к удалённым SSD-накопителям.

На стороне сервера хранения – можно использовать **SPDK**.

А для передачи самих данных по сети – можно использовать технологию NVMe-oF (NVMe over Fabric) поверх RDMA: **RoCE** или **iWARP**.

Благодаря этим технологиям можно достичь пропускной способности и задержки на уровне с локальным SSD.

Практическое применение технологий

3. Высокопроизводительные кластеры, например, кластеры ИИ.

Такие кластеры требуют очень низких задержек между виртуальными машинами. Чтобы это обеспечить, можно использовать одну из технологий RDMA – **InfiniBand**.

А также **SmartNIC/DPU** для снижения нагрузки на процессор при передаче трафика между разными хостами с виртуальными машинами в рамках одного кластера.

4. Высокопроизводительные распределенные системы хранения (Ceph) и файловые системы.

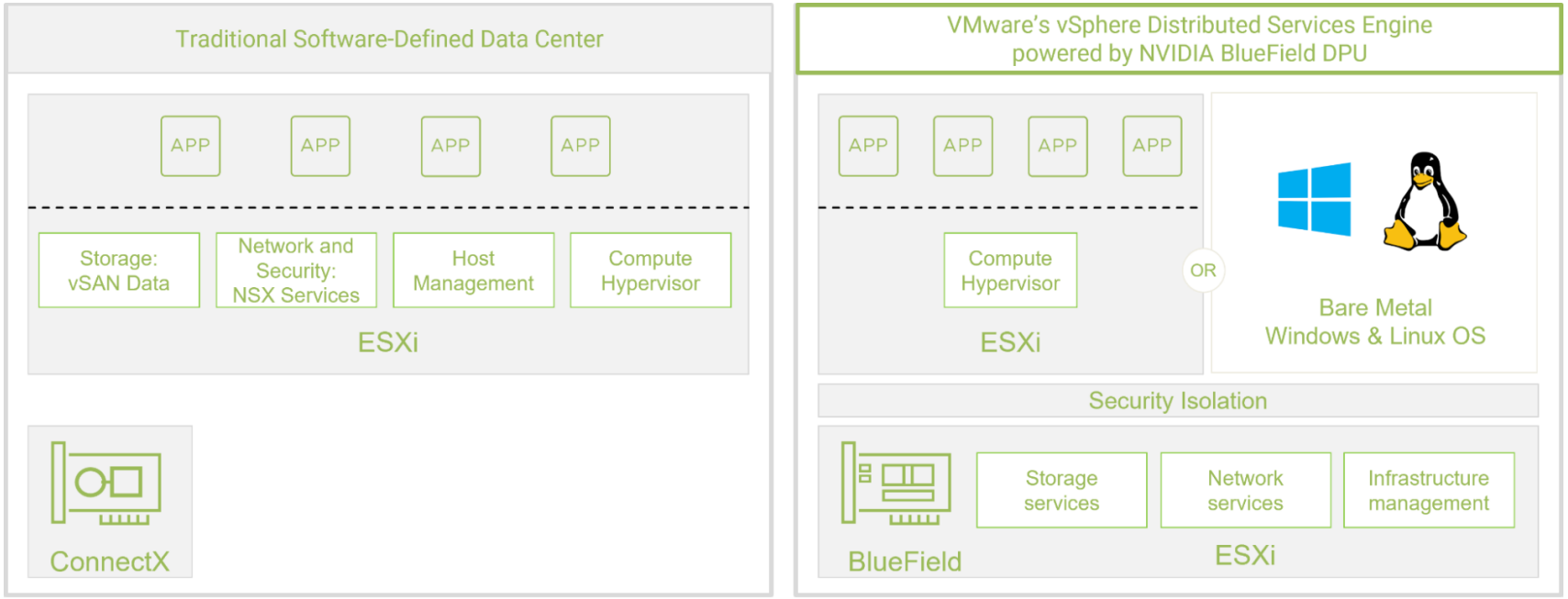
Для снижения накладных расходов сетевого стека и стека хранения на примере Ceph можно использовать:

- Для ускорения Bluestore – **SPDK**;
- Для сетевого взаимодействия между OSD (Object Storage Daemon) – **RDMA** и **SmartNIC/DPU**.

Реальные кейсы использования

Ссылка на блог NVIDIA

Использование NVIDIA BlueField DPU в **VMWare vSphere**



Реальные кейсы использования

Использование SPDK в CEPH.

CEPH в официальной документации поддерживает использование SPDK.

Ссылка на официальную документацию CEPH

MWS Cloud Platform используют в своей облачной инфраструктуре SPDK для ускорения работы с CEPH-хранилищами и не только.

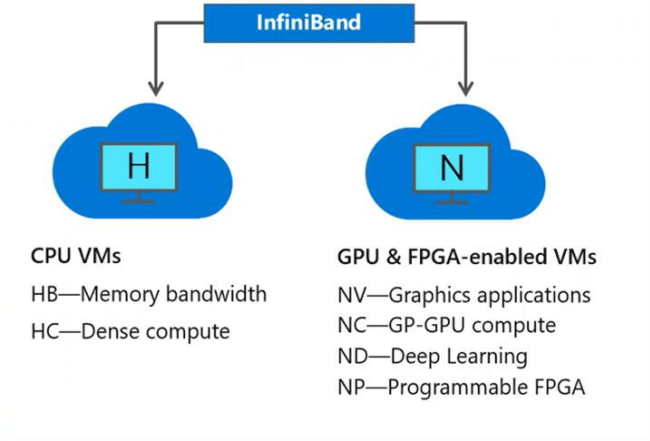
Ссылка на статью на habr, где MWS Cloud Platform рассказывают о внедрении SPDK

Реальные кейсы использования

Ссылкана блог NVIDIA

У Microsoft Azure есть облачные инстансы, предназначенные специально для высокопроизводительных вычислений – линейки HB, HC и HX.

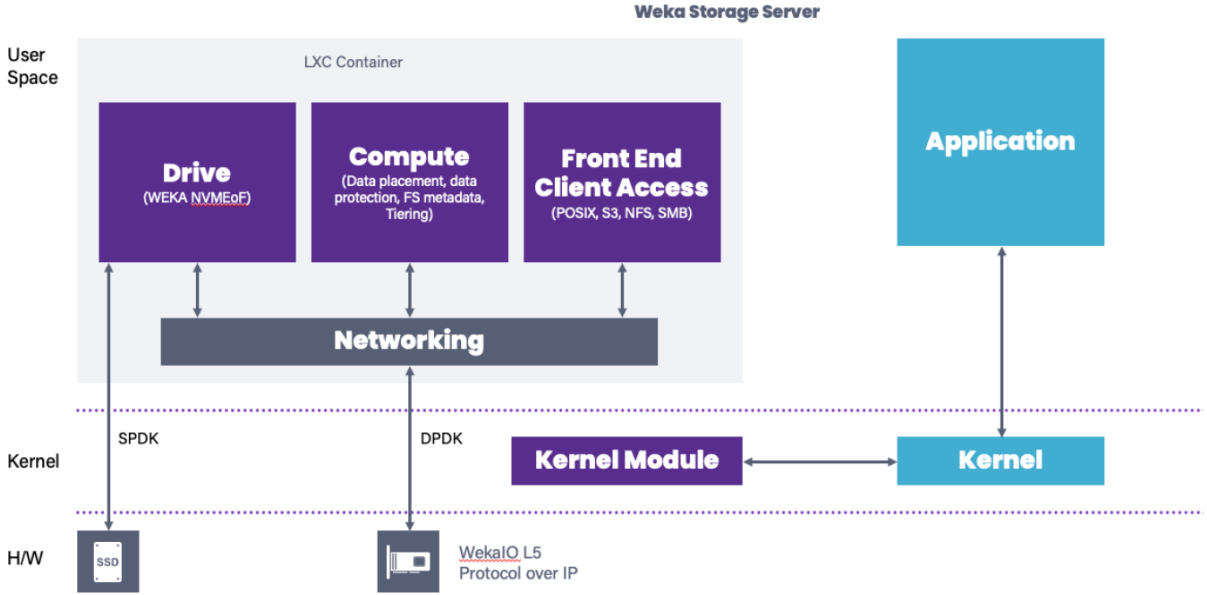
Они поддерживают интерфейс, который позволяет настроить соединение по RDMA (а именно InfiniBand) между несколькими виртуальными машинами.



Ссылка на документацию Microsoft о включении InfiniBand

Реальные кейсы использования

SPDK и DPDK используются в файловой системе **WekaFS** – распределённой файловой системе.



Ссылка на «white paper» архитектуры WEKA

Спасибо
за внимание!

ITMO more than a UNIVERSITY