Открытый подход к построению высокоскоростных и масштабируемых Ethernet фабрик для облачных ЦОД

Александр Петровский, системный инженер Mellanox

Октябрь 2016

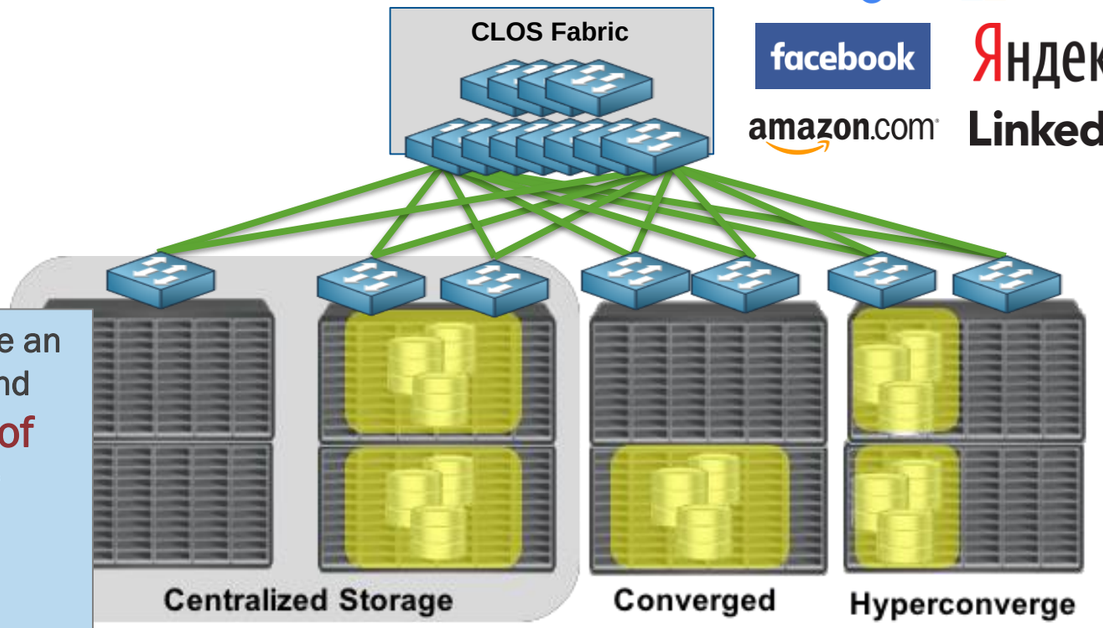
Mellanox TECHNOLOGIES

Connect. Accelerate. Outperform.™

Новый подход к архитектуре - переход к масштабируемой парадигме Web-scale IT

"By 2017, Web-scale IT will be an architectural approach found operating in 50 percent of global enterprises..."

Gartner, March 5, 2014



Google Microsoft facebook Яндекс amazon.com LinkedIn

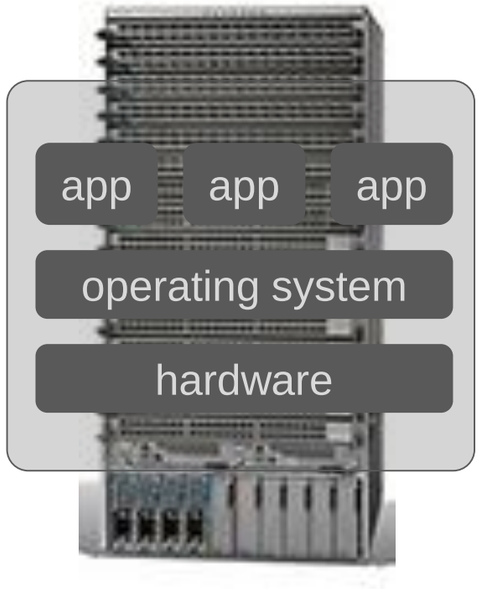
Wall street IT

- Scale-up
- Centralized
- Традиционное управление
- Проприетарное ПО
- Виртуализация
- Hardware-defined

Web-scale IT

- Scale-out
- Distributed
- Автоматизация, DevOps
- Open source
- Гиперконвергенция
- Software-defined

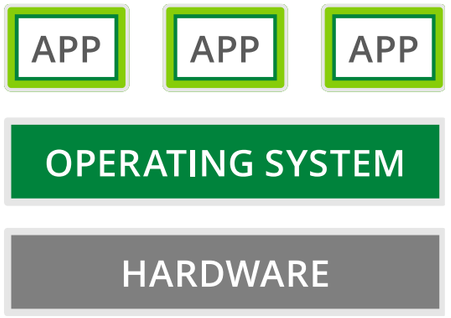
Новое видение платформы - дезагрегация и открытие аппаратных компонентов - Open Ethernet



Закрытая платформа
- Привязка к одному вендору
- Дорого!
- Медленный цикл разработки

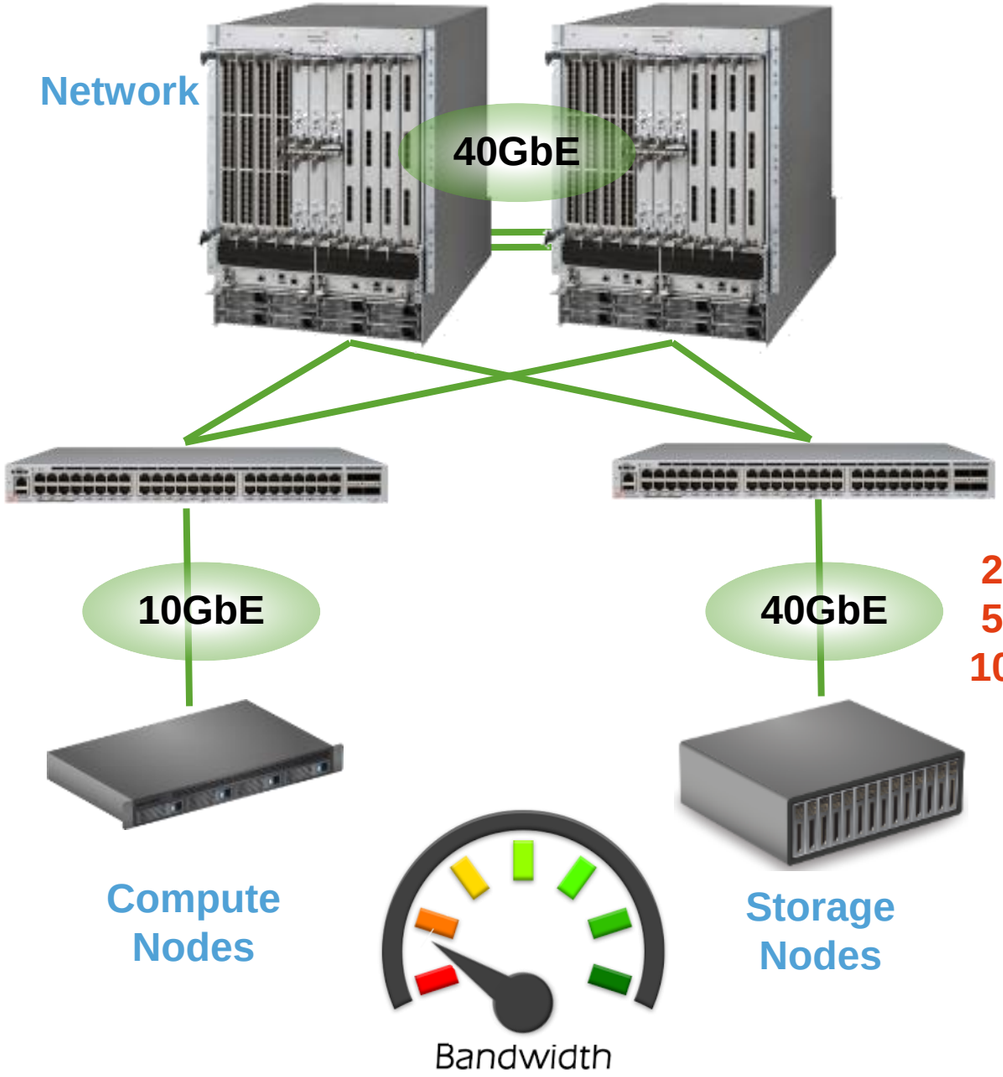
Возможность выбора
- Лучшего железа
- Лучшего ПО
- Быстрое внедрение

Дезагрегация инфраструктуры:
- ONIE, SDK API, SAI



Новый взгляд на производительность - скорости 25/50/100GbE

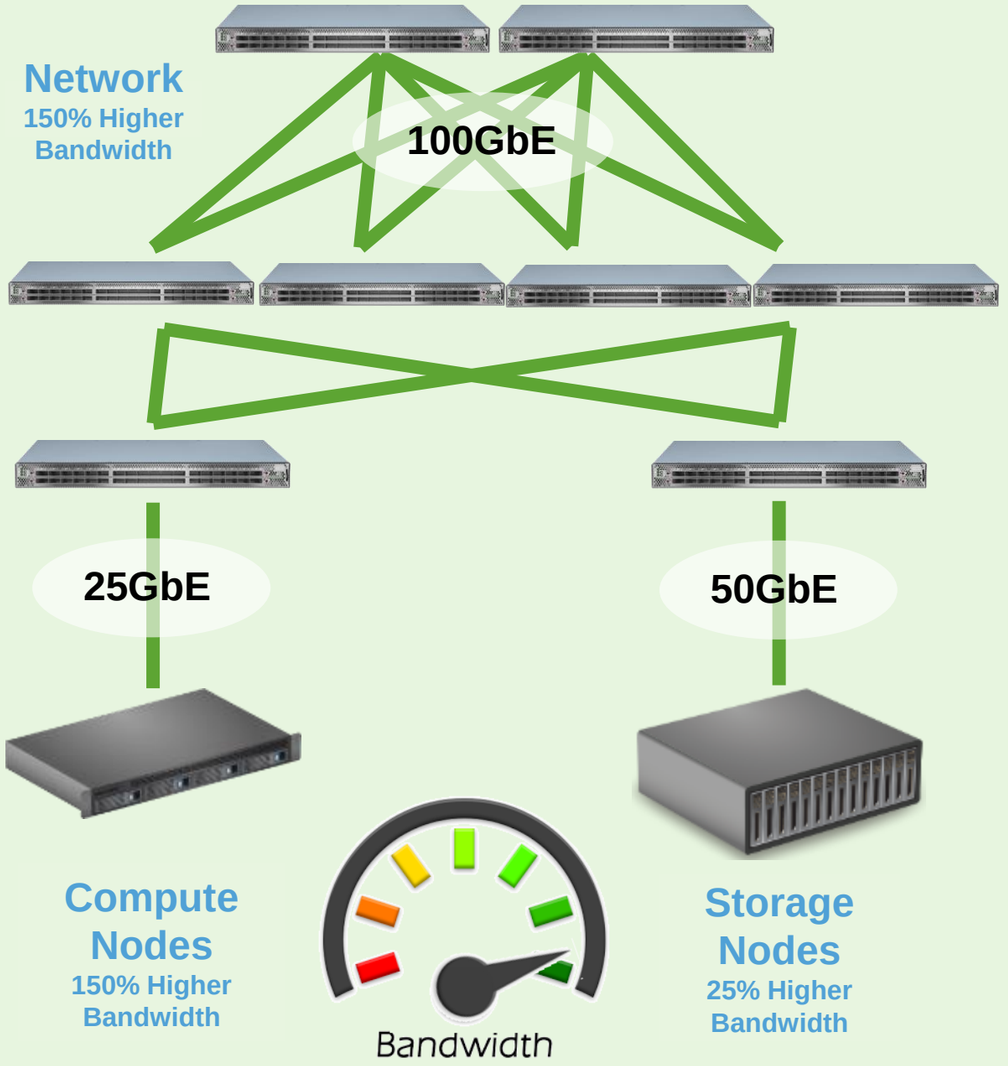
Традиционный дизайн сети ЦОД
проприетарные решения



Тот же Ethernet
Тот же СКС
Ниже цена и энергопотребление

25GbE – новый 10GbE
50GbE – новый 40GbE
100GbE – будущее ЦОД

Масштабируемый дизайн Ethernet фабрики
основан на открытых стандартах



Новая реальность интеграции и управления - виртуализация и SDN

Compute Virtualization

Windows Hyper-V
KVM
vmware ESXi
docker
kubernetes
OpenVZ Virtuozzo Containers

Storage Virtualization

RAIDIX
vmware Virtual SAN
ceph
Microsoft
ScaleIO Elastic Converged Storage

SDN Virtualization Overlay

vmware NSX
nuagenetworks
midokura

SDN

OpenFlow
OPEN DAYLIGHT
ONOS
Brain4Net Программируемые Сети
ЦЕНТР ПРИКЛАДНЫХ ИССЛЕДОВАНИЙ КОМПЬЮТЕРНЫХ СЕТЕЙ

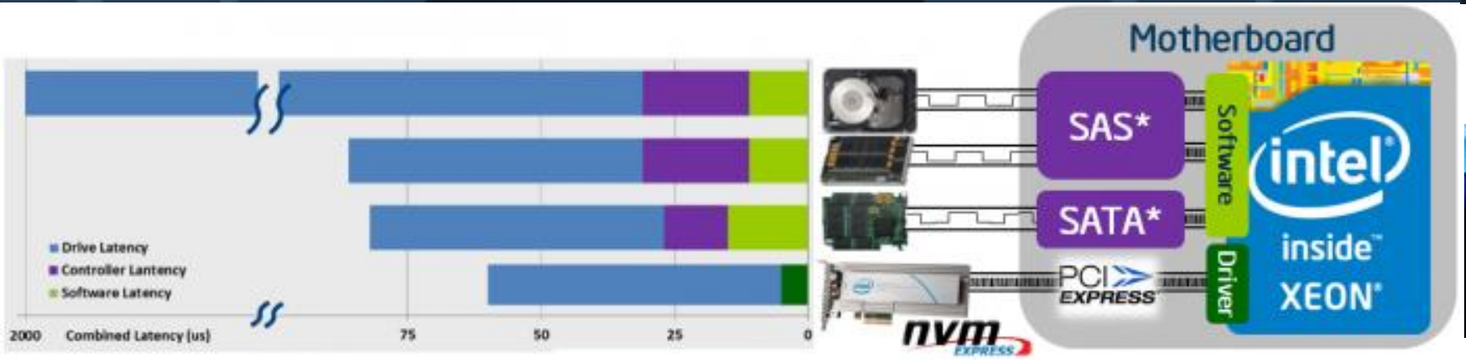
Converged / Hyper-converged

Windows Server 2016
NUTANIX
EMC Converged Platforms VCE
СКАЛА-Р IBS
vmware EVO SDDC
SCALE COMPUTING
simplivity

Orchestration

Windows Azure Pack
vmware vSphere
openstack CLOUD SOFTWARE

Скорости 25/50/100G – зачем?

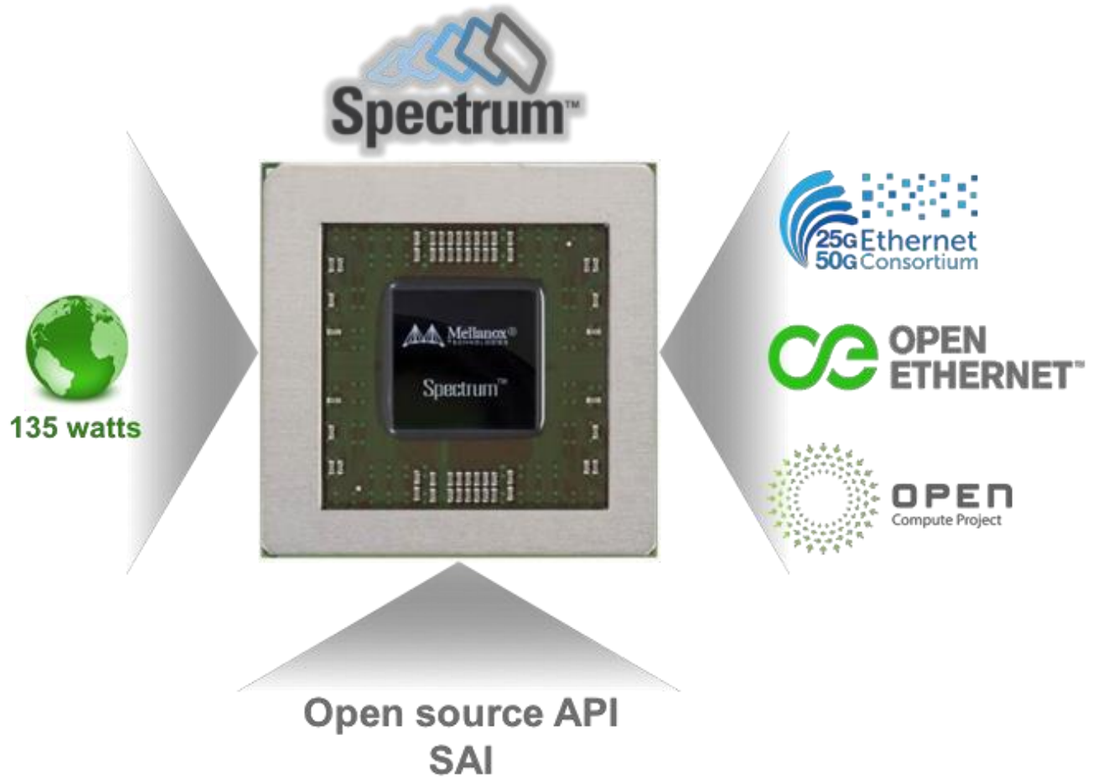


Live Migration

- 2xNVMe >100Gb/s ⟶ 10GbE? FC16/FC32? – Really?
- Миграция «живых+тяжелых» виртуальных машин:
  - Сколько по времени будет переезжать база данных размером 128GB поверх 1gb/s или 10gb/s сети?
  - **Ответ: она скорее никогда не переедет.**
  - А что происходит с сетью когда переезжает виртуальная машина (или несколько)?
  - **Ответ: Трафик между ВМ и от них в Интернет, крайне ограничен**
- Распределённые приложения: HPC, Big-Data, etc…

Spectrum 100G Ethernet ASIC – платформа Open Ethernet

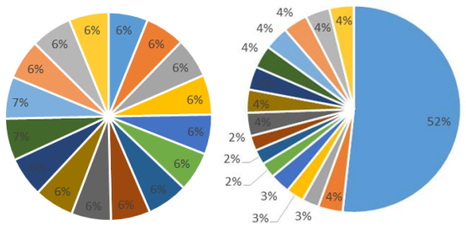
- **Лидер по производительности**
  - Неблокирующая коммутация 6.4Tb/s
  - <300ns задержки на L2/L3 от 64b до 9Kb
  - Zero Packet Loss
- **Масштабируемость для облаков**
  - Поддержка виртуализации
  - Оптимизация пропускной способности
  - Гибкие SDN возможности
- **Функциональность**
  - 32 порта по 100 / 56 / 40GbE
  - 64 порта по 50 / 25 / 10GbE
  - RDMA over Converged Ethernet
  - Программируемость для SDN и поддержка Overlay (VXLAN, NVGRE, Geneve) и MPLS



Ключевые требования к коммутаторам в ЦОД

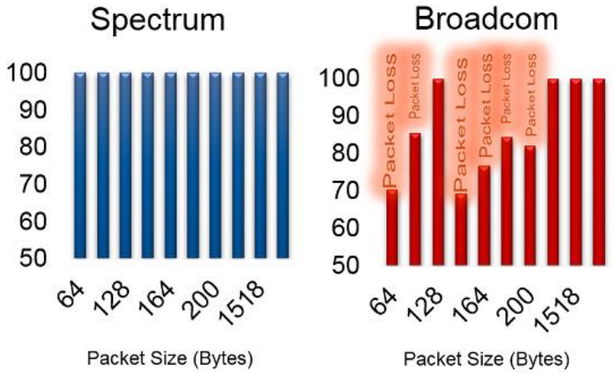
Fairness

Spectrum Broadcom



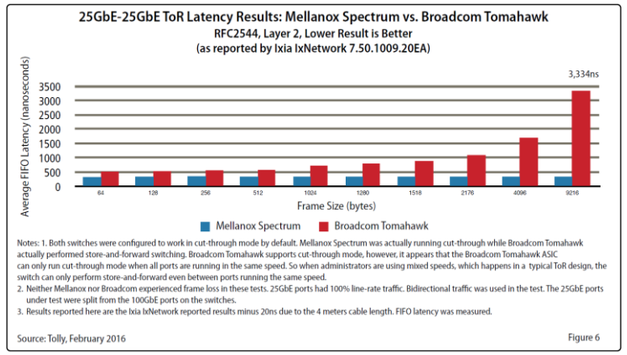
Равномерное распределение полосы

Zero Packet Loss



Отсутствие потерь пакетов любых размеров при любой нагрузке

Latency



Стабильно низкая задержка для любого типа трафика и любых размеров пакетов

Spectrum™

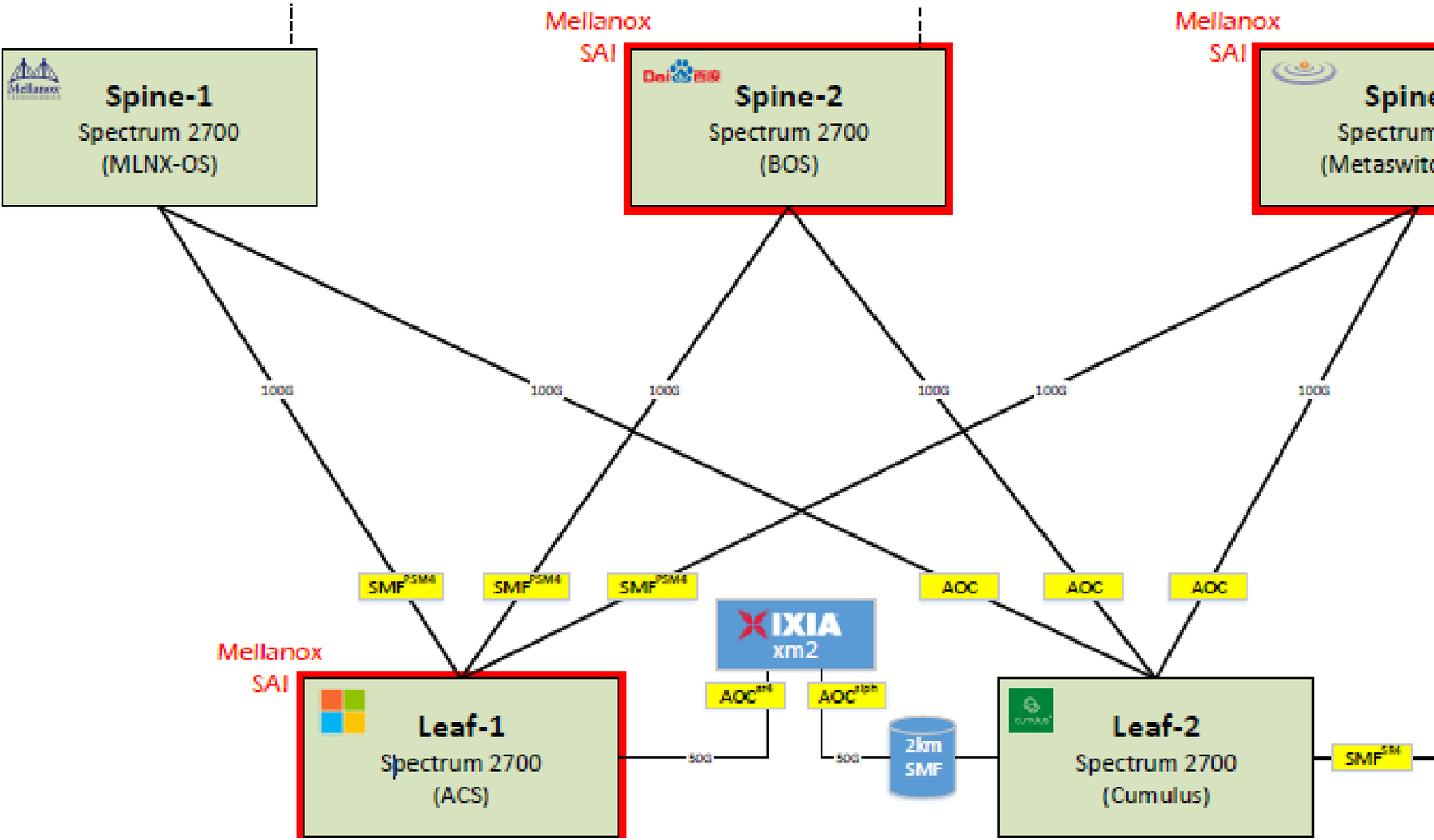
- Наивысшая производительность
- Наивысшая масштабируемость + отказоустойчивость на всех уровнях
- Интеграция с «оркестраторами» (SDN)
- Лучшее соотношение $/Гбит

Tolly.

www.zeropacketloss.com
www.Mellanox.com/tolly

Выбор сетевой ОС на Spectrum – уже реальность

OCP Summit March 2016 – Live Demo



<u>6 коммутаторов Spectrum</u> :

- Cumulus Linux
- Mellanox OS
- Microsoft ACS (SONiC)
- HP OpenSwitch (future)
- MetaSwitch (future)
- BaiduOS (*future*)

Linux на Spectrum - максимальная открытость платформы

User Space

iproute2 utilities
(tc, bridge, ip, etc.)

3rd party applications / NOS
(Quagga, OpenFlow, etc.)

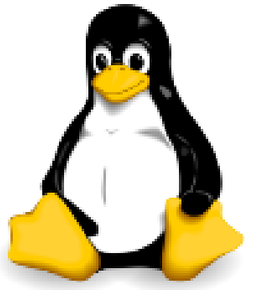
User applications

redhat fedora ubuntu debian

Kernel

Linux Network Stack

Linux Network Drivers
mlxsw
(Mellanox Switch Drivers)



Hardware

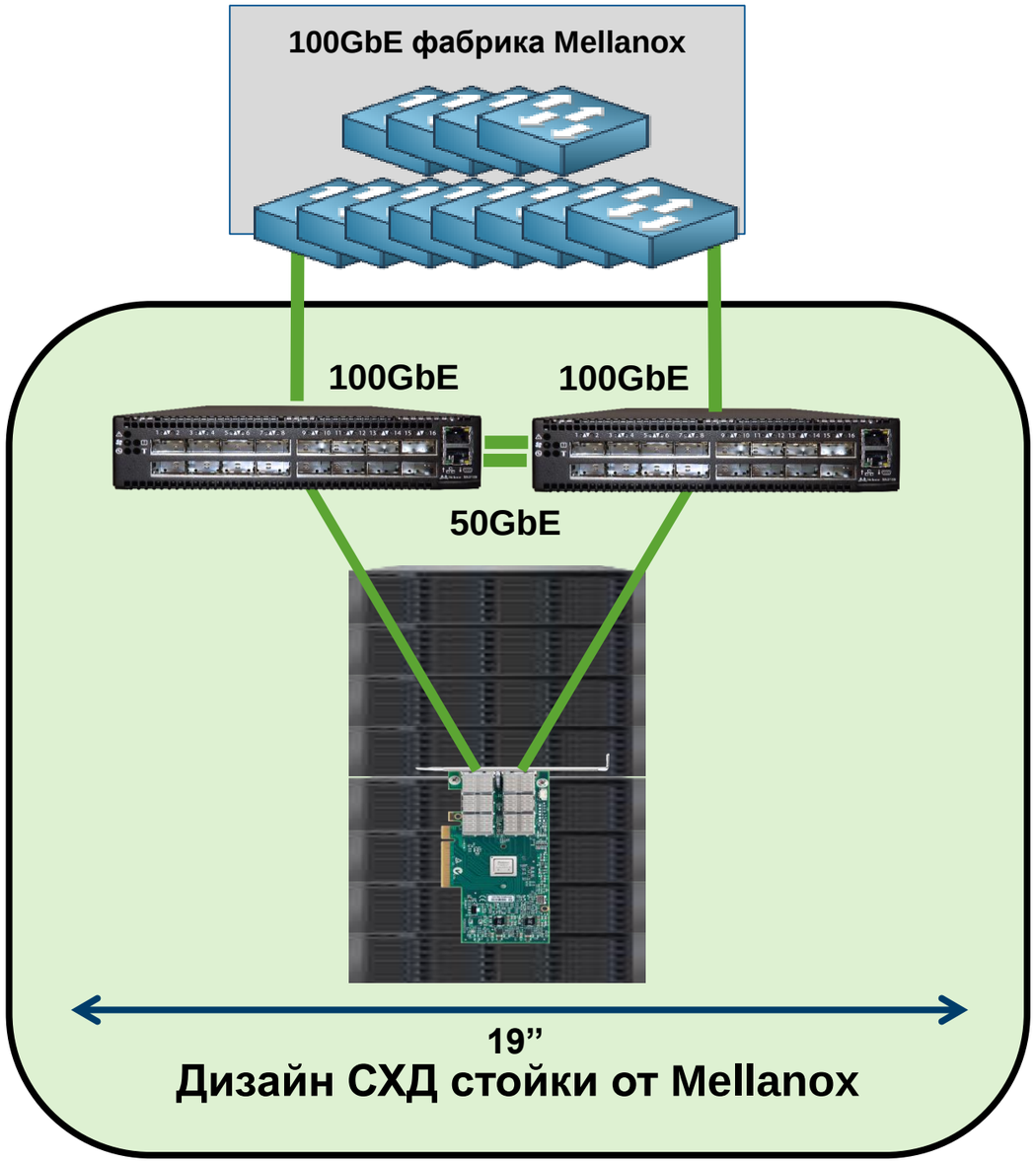
Mellanox Spectrum ASIC

Spectrum™

Mellanox - лучший интерконнект для облачных платформ

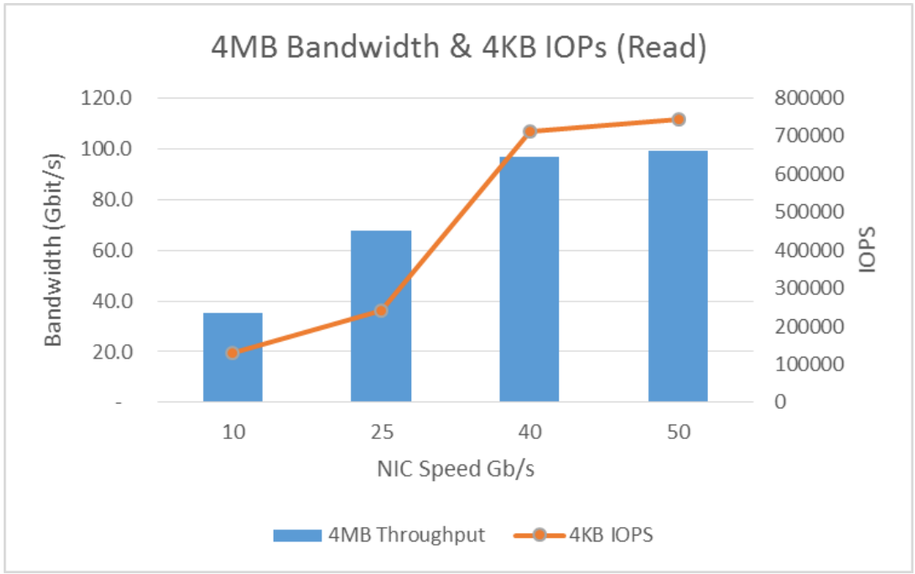
- Законченное решение
  - Проверенная и поддерживаемая архитектура для наиболее популярных систем виртуализации
- Простота управления
  - Интеграция в системы управления виртуализацией
  - Снижение OPEX
- Повышение эффективности серверных платформ
  - Поддержка RDMA и Offloads на всех трех платформах.
  - Снижение CAPEX
- Гибкие возможности масштабирования
  - Единая архитектура для разных размеров

Ethernet для СХД и гиперконвергенции - несравненная производительность



Зачем 50GbE?

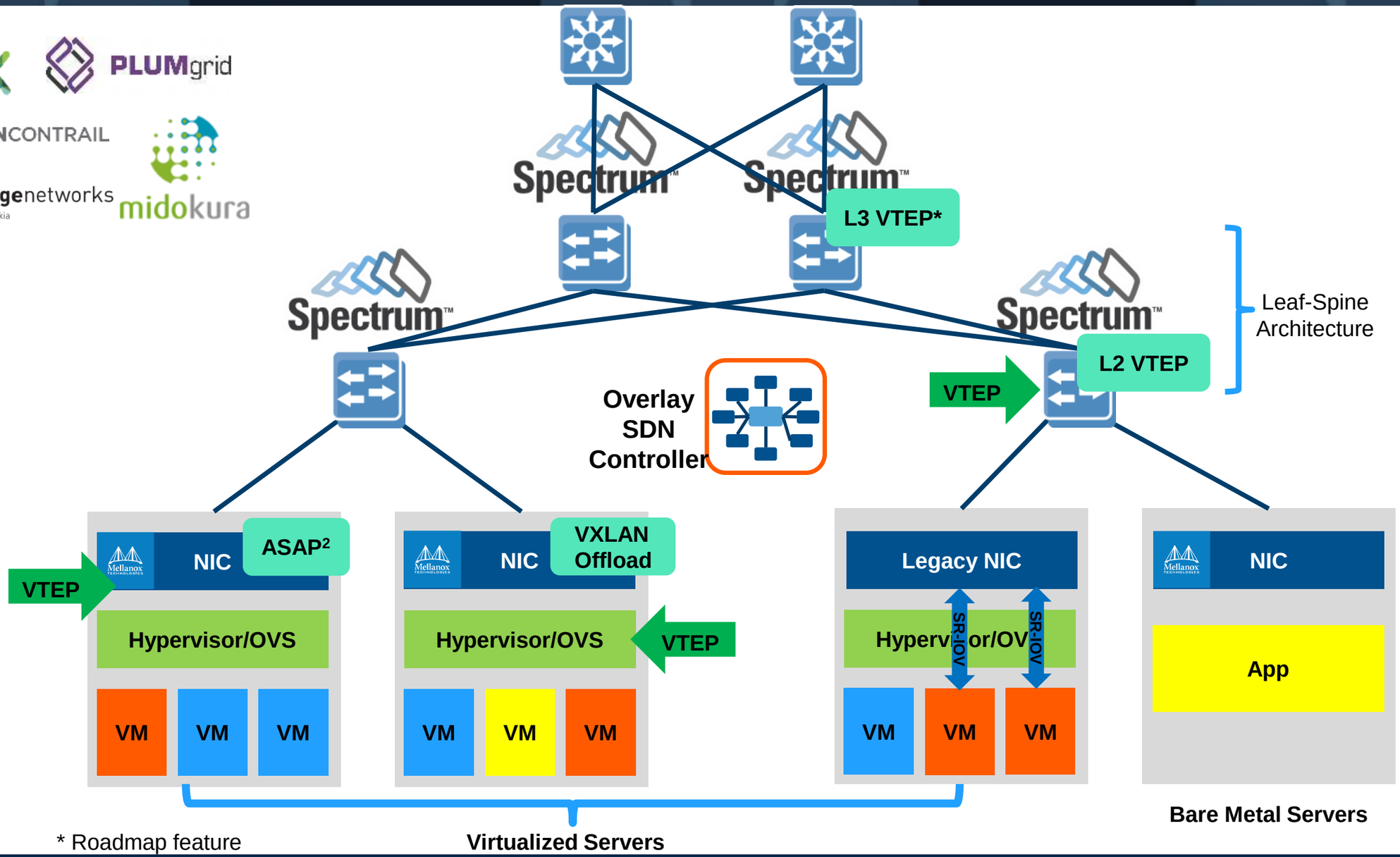
График производительности Ceph кластера:



- **Отказоустойчивость в 1RU**
- **64 порта 50GbE на 1RU**
- **Поддержка RDMA/RoCE для CPU offload**

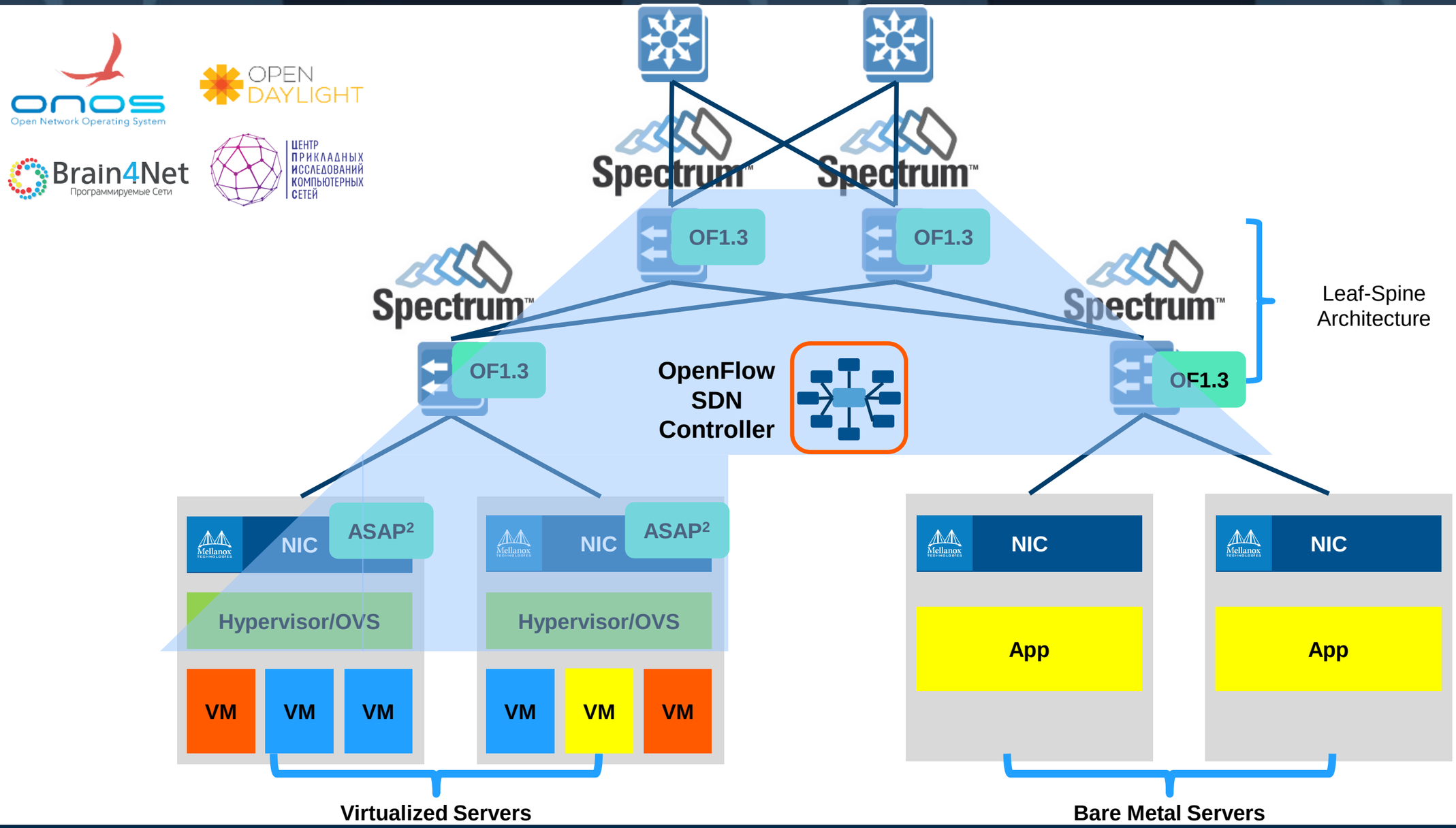
RoCE iSER 25G 50G Ethernet Consortium nvm EXPRESS ceph Open vStorage Storage designed for Cloud Computing

Унифицированная 100GbE SDN фабрика - Overlay



* Roadmap feature

Унифицированная 100GbE SDN фабрика - OpenFlow



А каким должен быть Ethernet в ЦОДе?

- Масштабируемым:CLOS L3 фабрика, ECMP
- Широкополосным: 10, **25**, 40, **50** , **100** Gb/s
- Открытым: на базе открытой платформы, выбор OS
- С низкими задержками: 1-2us сервер-сервер
- Без потерь пакетов: Lossless, с поддержкой PFC, ECN
- Справедливый scheduling: каждый клиент должен получать свой BW вне зависимости от нагрузки
- Программно-управляемым: OpenFlow, Overlay
- Интегрируемым с приложениями: – SDS, Hyperconverged, RDMA

В заключение об открытом Ethernet от Mellanox

1. Выбирайте любые программные компоненты в сети (ОС, ПО, Стеки протоколов)

2. Выбирайте самую лучшую аппаратную платформу

Спасибо!

Mellanox TECHNOLOGIES

Connect. Accelerate. Outperform.™