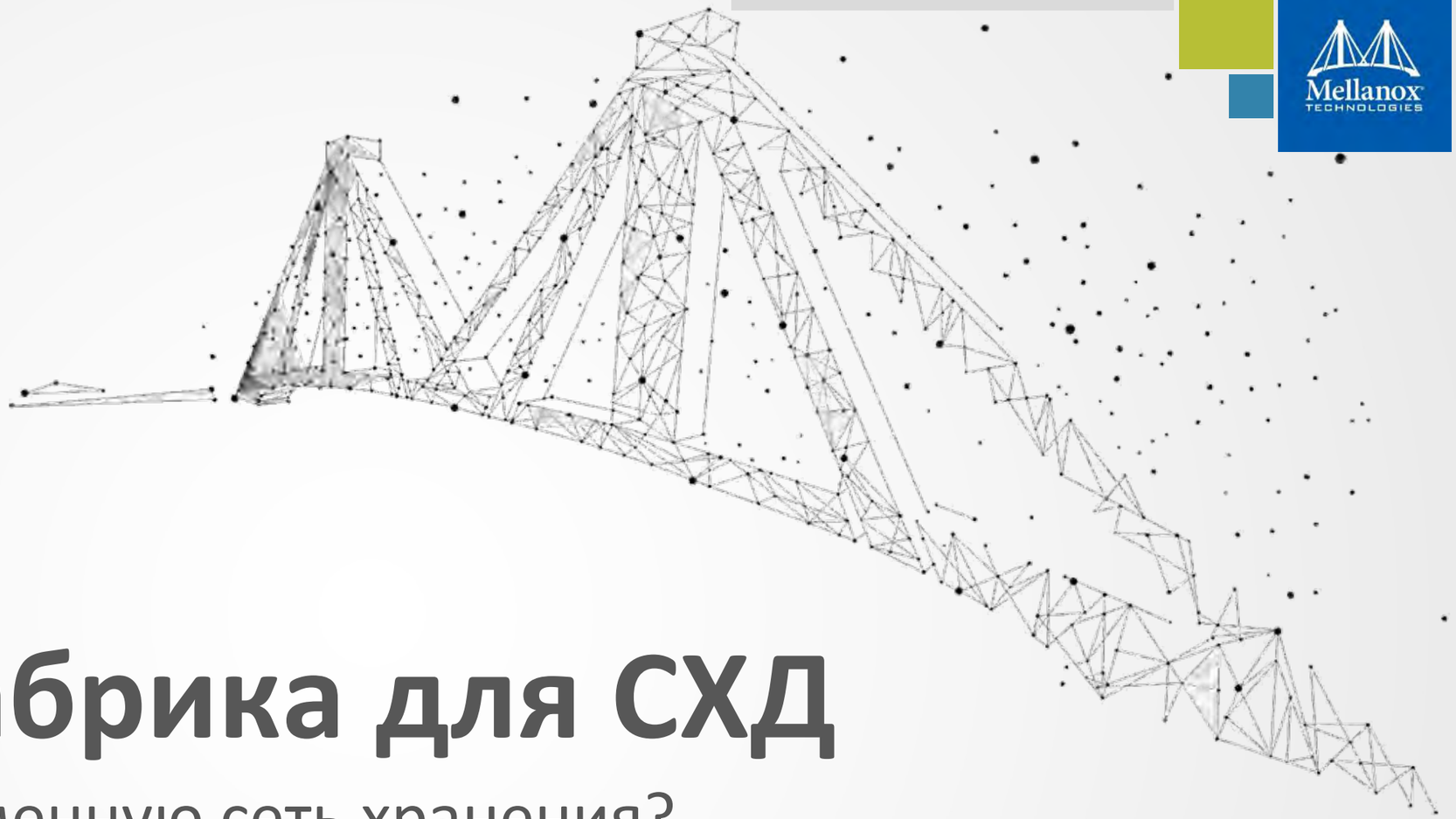


Ethernet-фабрика для СХД

Как построить современную сеть хранения?

Александр Петровский

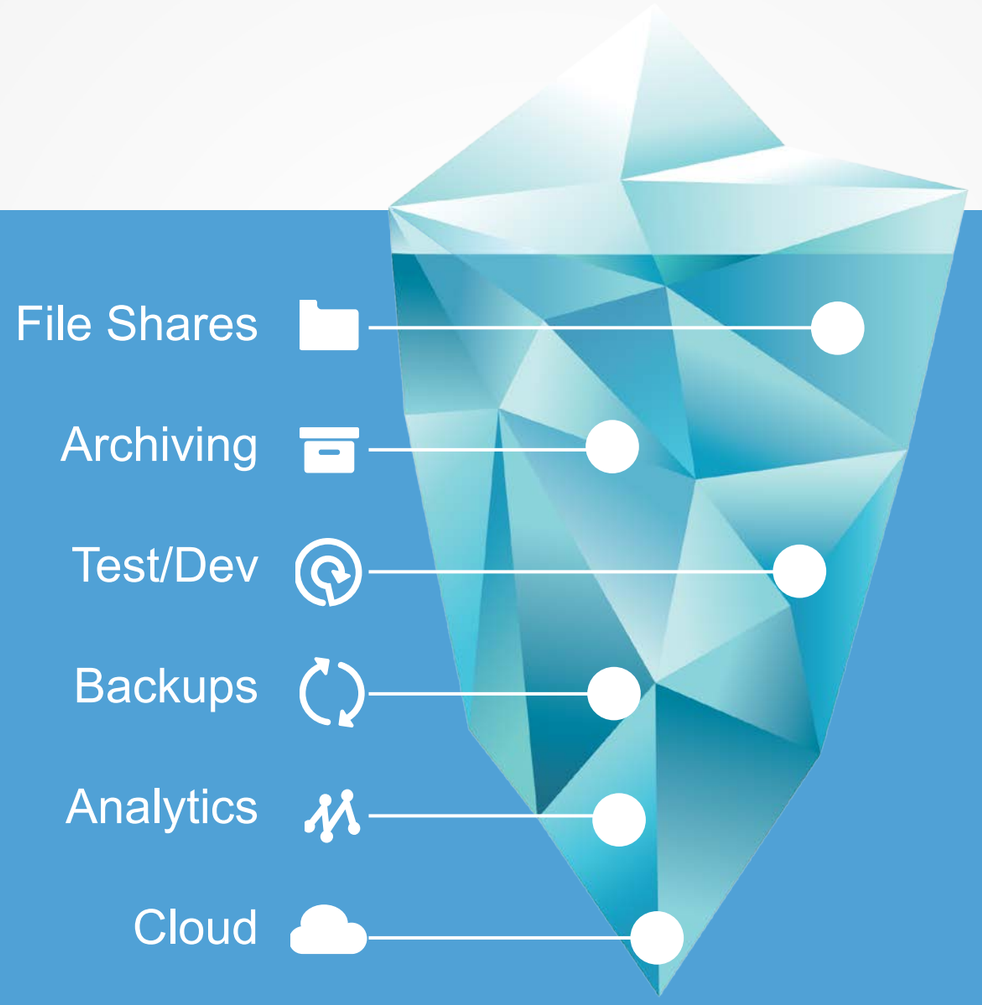
системный инженер Mellanox

Сентябрь 2018

© 2018 Mellanox Technologies

Ландшафт решений СХД сегодня

STORAGE ICEBERG



PRIMARY STORAGE

- Традиционный SAN
- 20% всех данных

SECONDARY STORAGE

- 80% всех данных
- Block, File, Object
- Tiering
- 95% Ethernet

© 2018 Mellanox Technologies

Почему рынок СХД меняется?

Изменения

Flash, быстрые серверы

Social/Mobile/Video

Гиперконвергенция

Облака

Большие данные

Распределенные приложения

Эффекты

Быстрая сеть: 10/25/40/50/100GbE

Рост данных: больше файловых и объектных данных

Распределенные кластерные СХД поверх Ethernet

Виртуализация, software-defined storage, удешевление «железа»

Файловые и распределенные хранилища

Больше east-west трафика

Итог: больше Ethernet трафика в СХД

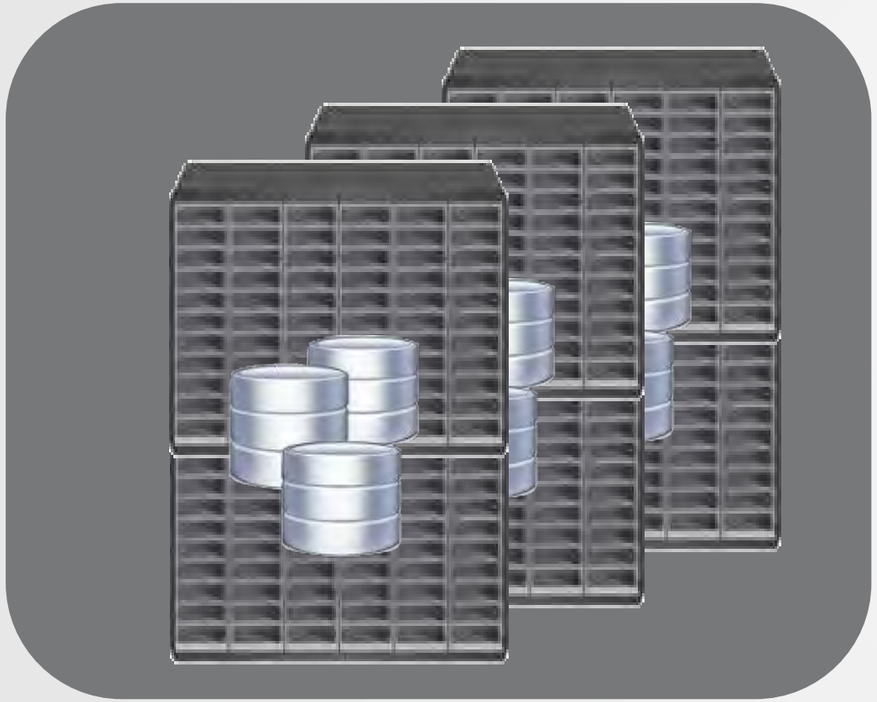
© 2018 Mellanox Technologies

Почему Ethernet для СХД?

© 2018 Mellanox Technologies

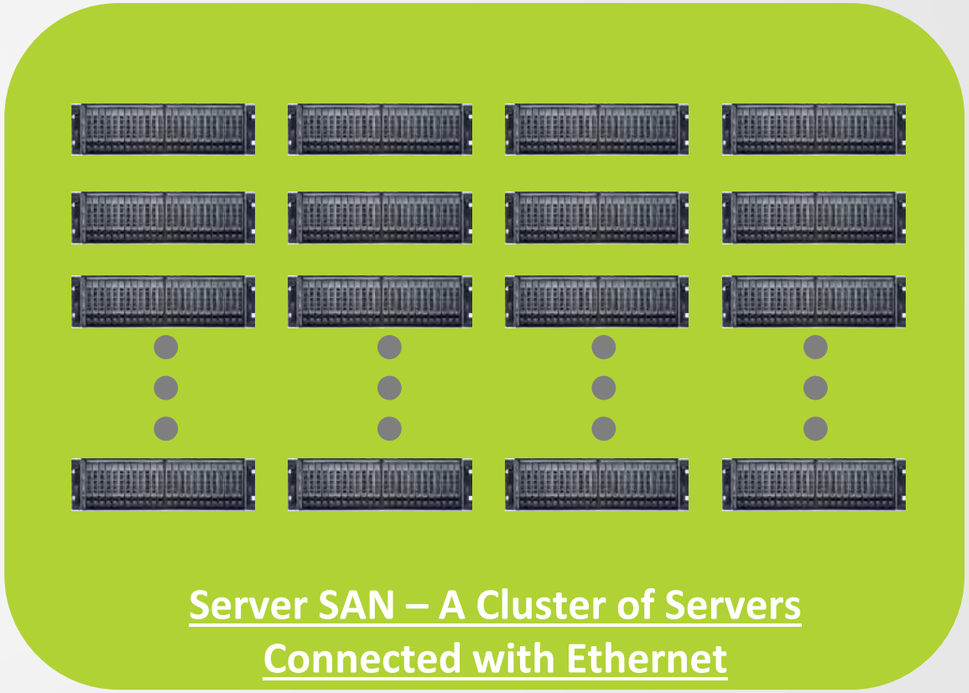
Смена парадигмы в СХД для облачных ЦОД

Legacy ЦОД – FC SAN



- Проприетарный «черный ящик»
- Сложно обслуживать
- Дорого масштабировать

Современные облачные ЦОД



- На базе стандартных серверов
- Простое управление (software-defined)
- Легко и дешего масштабировать (scale as you grow)

© 2018 Mellanox Technologies | Confidential

Король умер, да здравствует король!

Ethernet вырывается вперед

1998

Feature	Fibre Channel	Ethernet
Bandwidth	1 G	100 M
Supports	Block	Block, file
Lossless	Yes	No
Cost	High $$$$	Medium $$
Cloud / HCI	No / No	No / No
Vendors	Several	Many
SDS / Scale-out	No / No	No / No

2018

Feature	Fibre Channel	Ethernet
Bandwidth	8/16/32 G	10/25/40/100 G
Supports	Block	Block, file, object
Lossless	Yes	Yes
Cost	High $$	Low $
Cloud / HCI	No / No	Yes / Yes
Vendors	2 / 2	Many / Many
SDS / Scale-out	Rare / No	Yes / Yes

Вчера: Storage сеть = FC

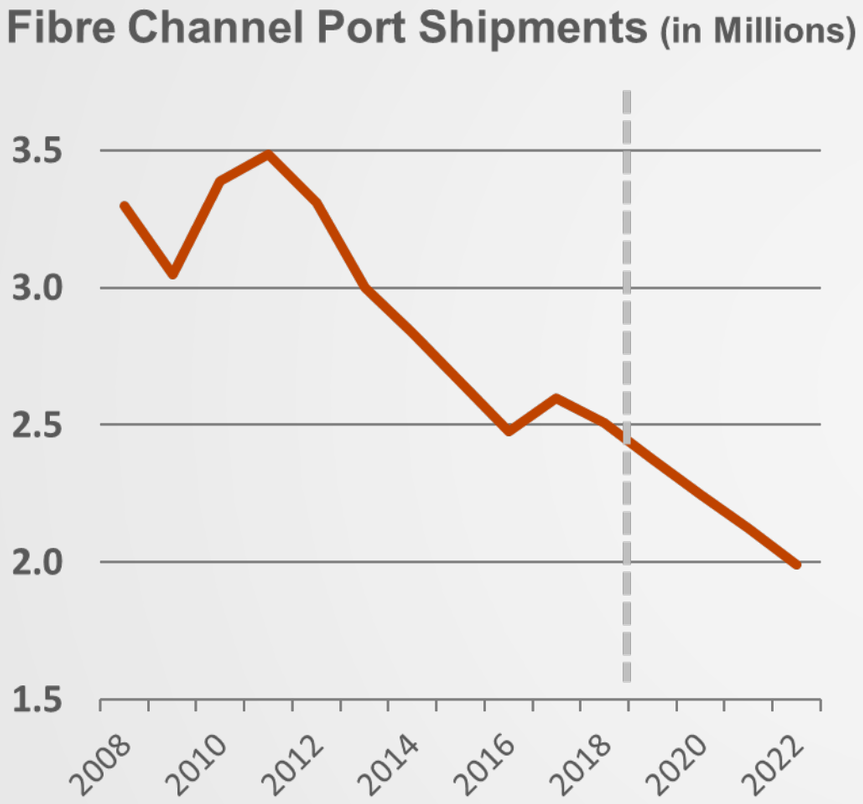
- Fibre Channel давал лучшую производительность
- Все СХД относились к tier-1 storage
- Не было облаков и виртуализации

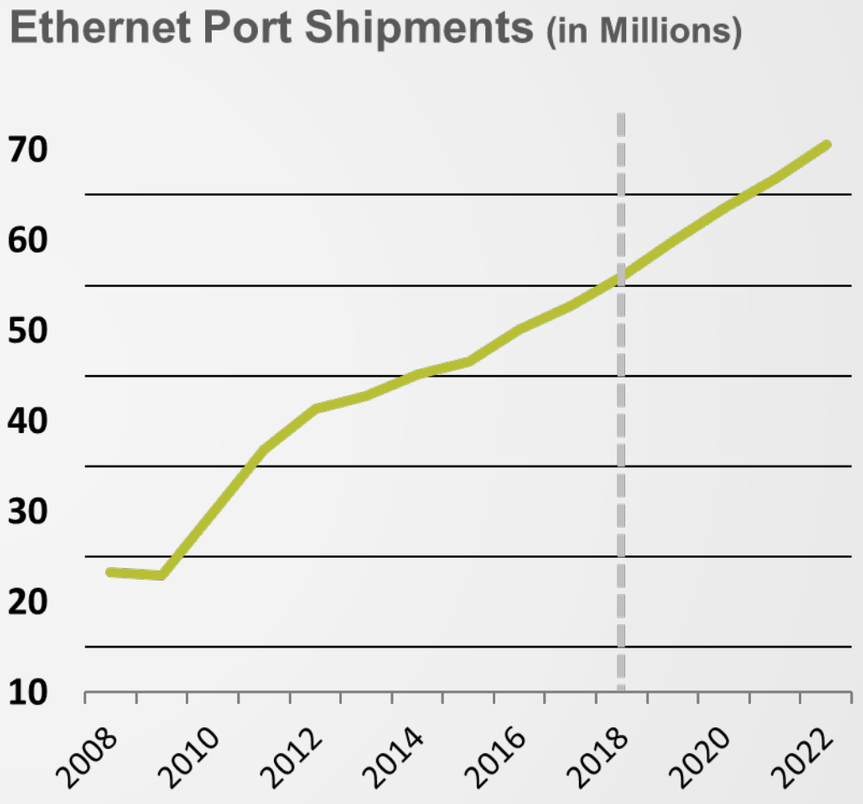
Сегодня: Ethernet как универсальная сеть также и для СХД

- Ethernet - выбор для Primary + Secondary storage!
- Block, Object, NAS, Cloud, Hyperconverged, Big Data
- Ethernet выгоднее и производительнее чем Fibre Channel

© 2018 Mellanox Technologies | Confidential

Поставки Ethernet vs. FC (в кол-ве портов)





Source: Crehan Research, January 2018

© 2018 Mellanox Technologies | Confidential

Ethernet-фабрика для СХД

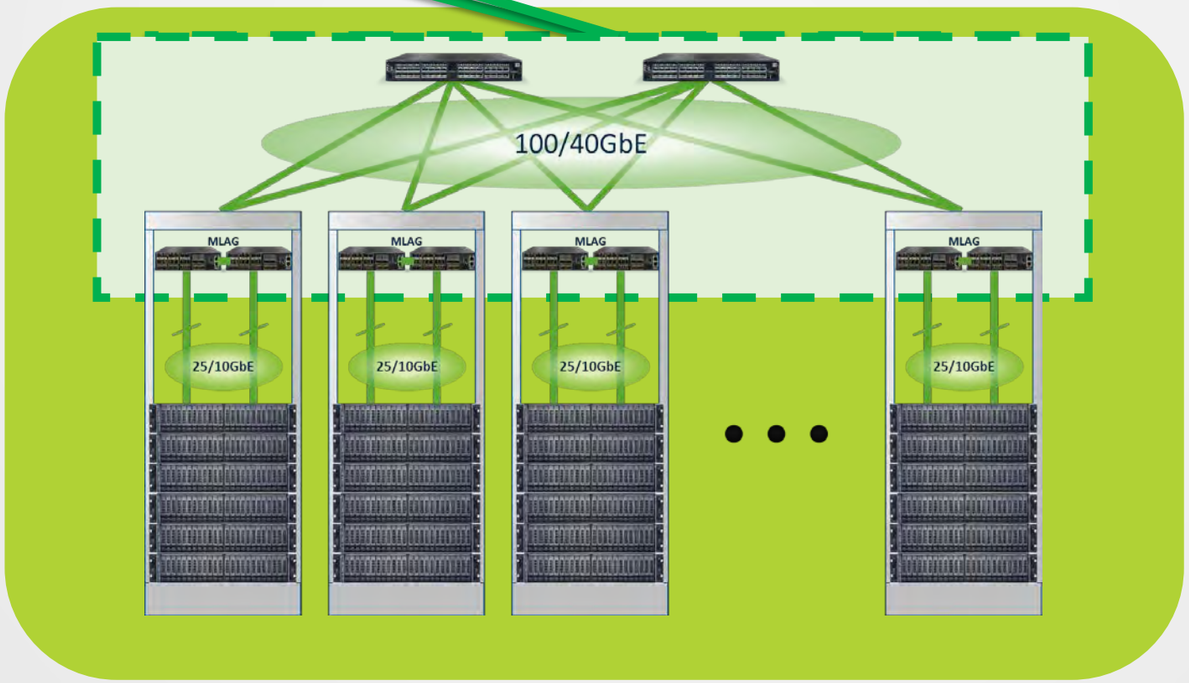
Ethernet Storage Fabric (ESF)

Ethernet Storage Fabric

- Производительность
- Масштабируемость
- Автоматизация
- Эффективность

Storage протоколы и приложения для ESF

- iSCSI/iSER
- NFS/NFSoRDMA
- CIFS/SMB Direct
- NVMeOF
- Ceph
- HDFS
- Lustre/GPFS/GlusterFS



Дизайн сети для ESF

- 100GbE **Layer 3** Clos фабрика
- **Маршрутизация OSPF/BGP**
- **LAG/MLAG** для СХД и серверов
- Масштабирование
  - **от 1 до 1000+** стоек

© 2018 Mellanox Technologies | Confidential

Решения Mellanox для ESF
Коммутаторы 10/25/40/50/100GbE

SN2010

Optimized 10/25G ToR for HCI and storage

- ½ width ToR
- 18x10/25GbE + 4x40/100GbE
- Supports 1GbE ports

SN2100

Ideal high-speed ToR for HCI and storage

- ½ width ToR
- 16x 40/100GbE
- 32x 50GbE or 64x 10/25GbE
- Supports 1GbE ports

SN2410

10/25GbE ToR for servers and storage

- 48x 10/25GbE + 8x 40/100GbE
- Supports 1GbE ports

SN2700

40/100GbE aggregation for servers and storage

- 32x 40/100GbE
- 64x 10/25/50GbE
- Supports 1GbE ports

cumulus®

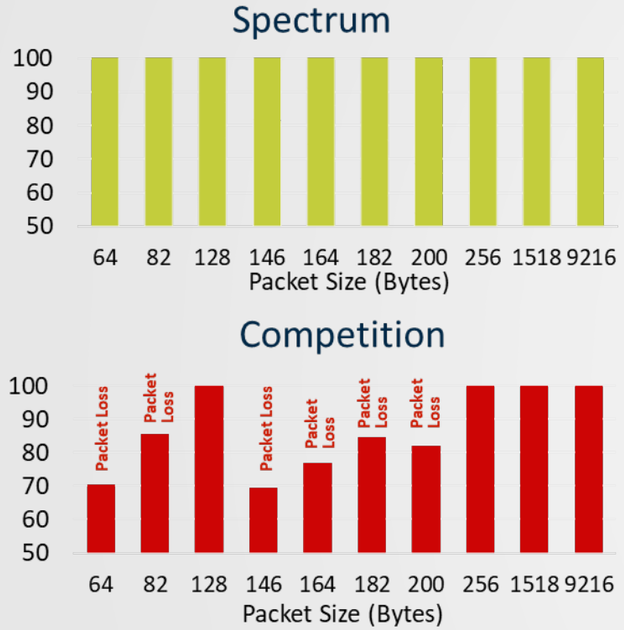
MELLANOX ONYX

Spectrum™

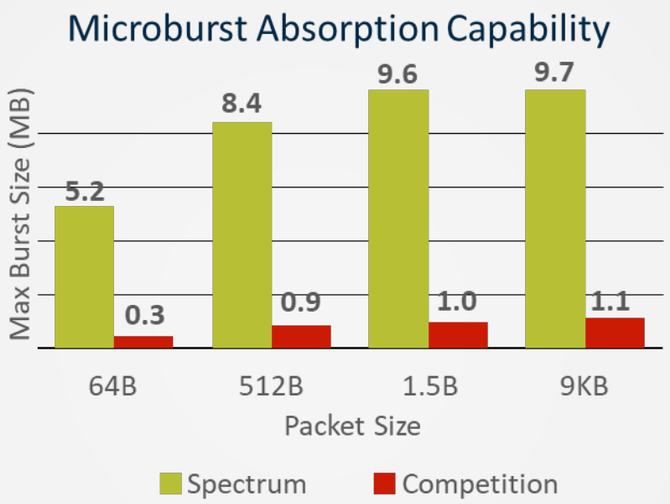
© 2018 Mellanox Technologies | Confidential

Spectrum - лучшее решение для ESF

Отсутствие потерь

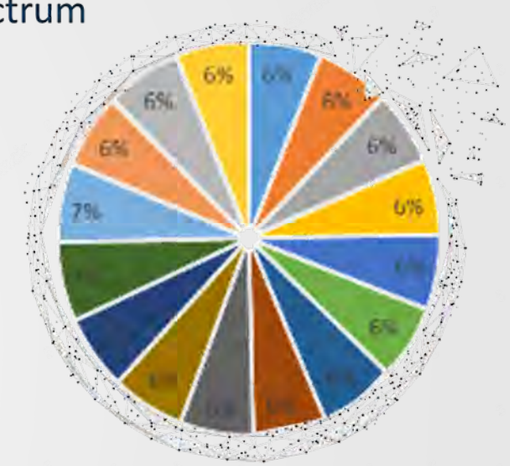


Буферизация

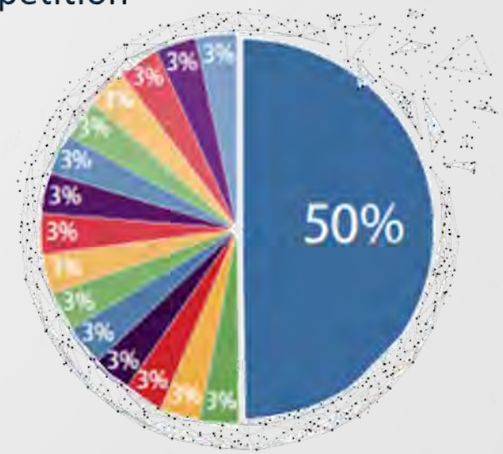


Качество обслуживания

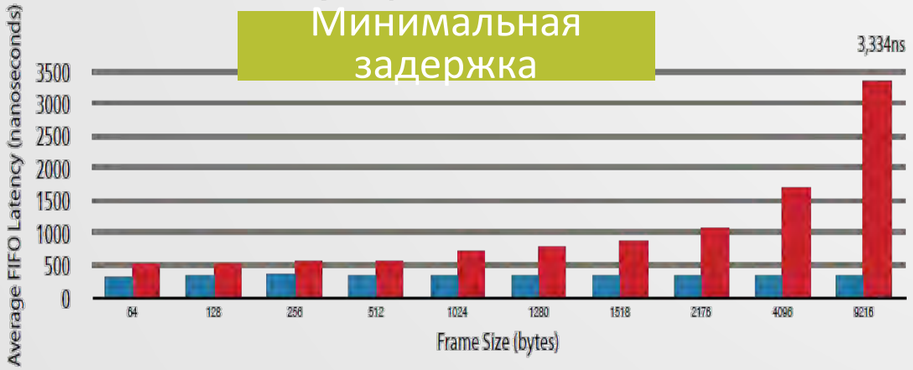
Spectrum



Competition



Минимальная задержка



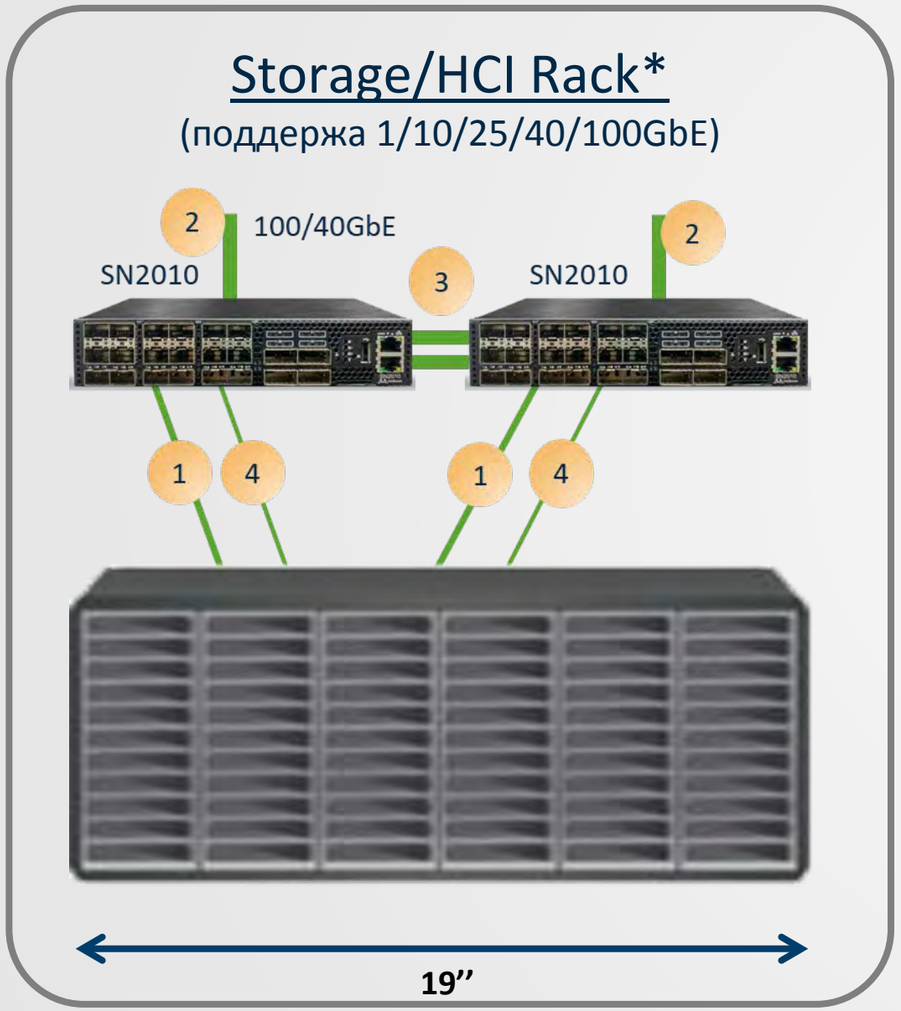
TEST REPORT
Tolly.

www.zeropacketloss.com

www.Mellanox.com/tolly

© 2018 Mellanox Technologies | Confidential

Оптимальное подключение стойки СХД с коммутатором SN2010



1. 10/25GbE link: SFP+/28 to SFP+/28

2. 40/100GbE link: QSFP/28 to QSFP/28

3. 100GbE MLAG Links: QSFP28 to QSFP28

4. 1GbE link: 1GbE Transceiver

- ½ 19” width, 1RU height
- 57W typical (ATIS)
- 18x10/25GbE + 4x40/100GbE
- Supports 1GbE ports

Performance

10G Optimized

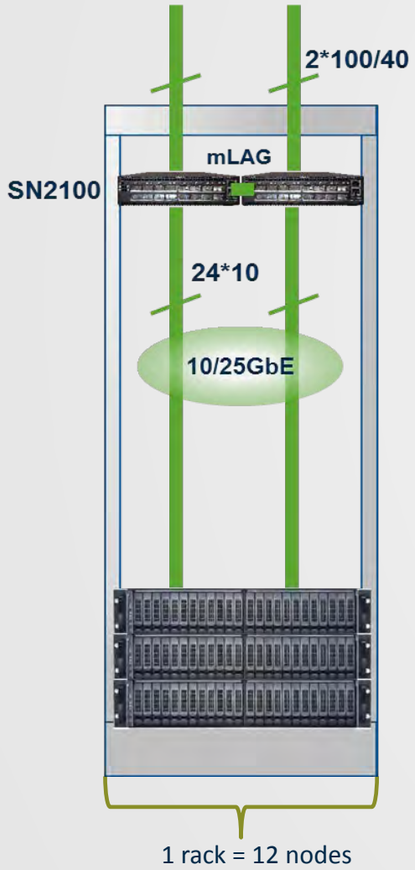
25G Ready

Best $/Gb/s

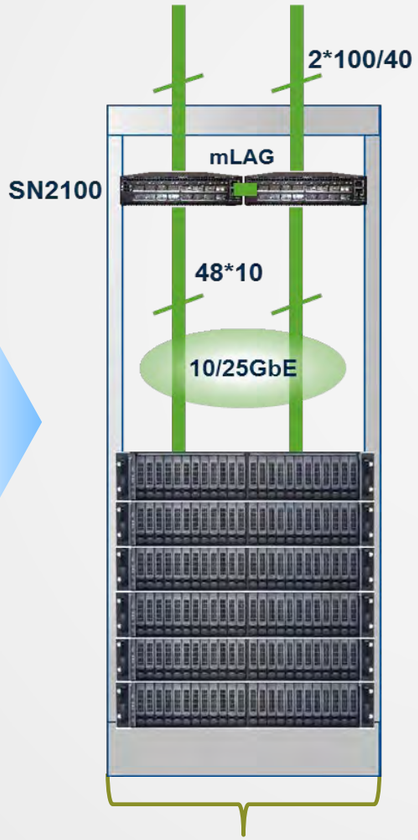
Blog [Jeff Shao: Mellanox SN2010 the Best HCI Switch]

Масштабирование Storage-фабрики

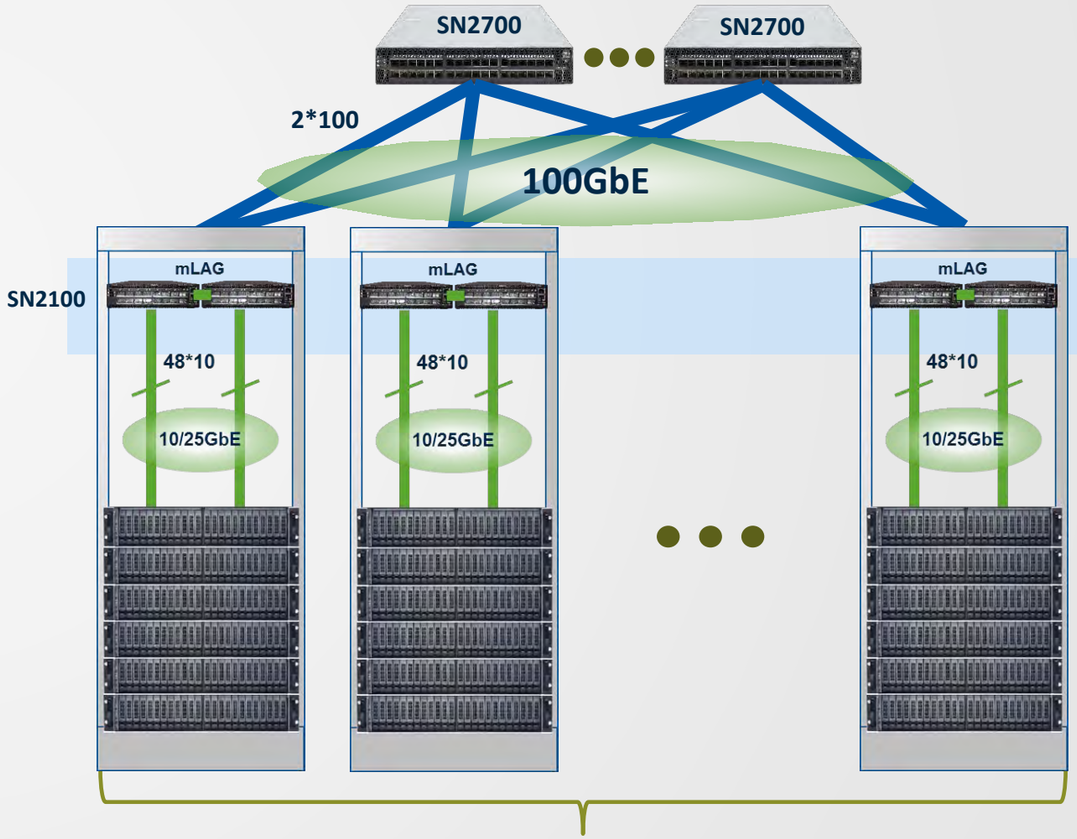
Часть стойки



Одна стойка



Несколько стоек



© 2018 Mellanox Technologies | Confidential

Акселерация СХД и приложений с RoCE

- Highest performance and lowest latency
- Hardware RDMA offload
- Hardware offload of RoCE congestion control
- Hardware offload of data path and NVMe command offload

- Zero packet loss, line-rate performance at all packet sizes and port combination
  - 30% loss in competitive solution
- Predictable buffer allocation to any port & packet sizes
  - Competitive variance spreading by ~600%
- Low latency, up to 90% latency in a typical TOR deployment scenario

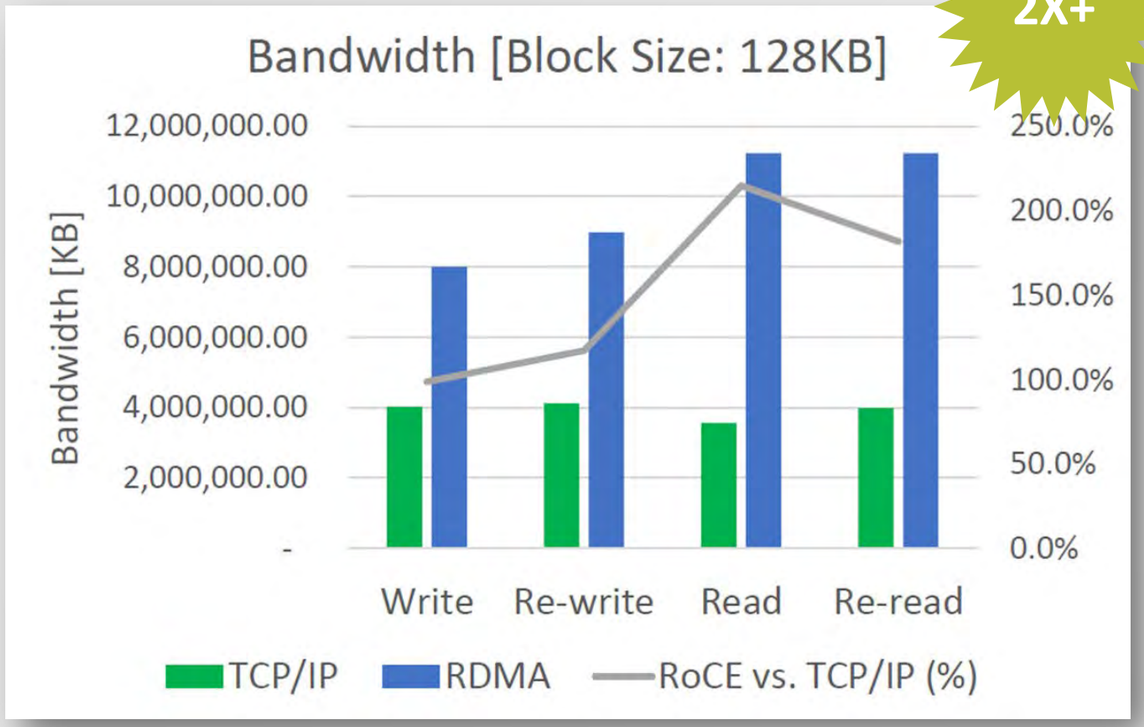
RoCE Enabled Storage

- **iSER**
- **NVMe-oF**
- **Microsoft SMB 3.0**
- **vSphere 6.5**
- **Ceph**
- **Spark**

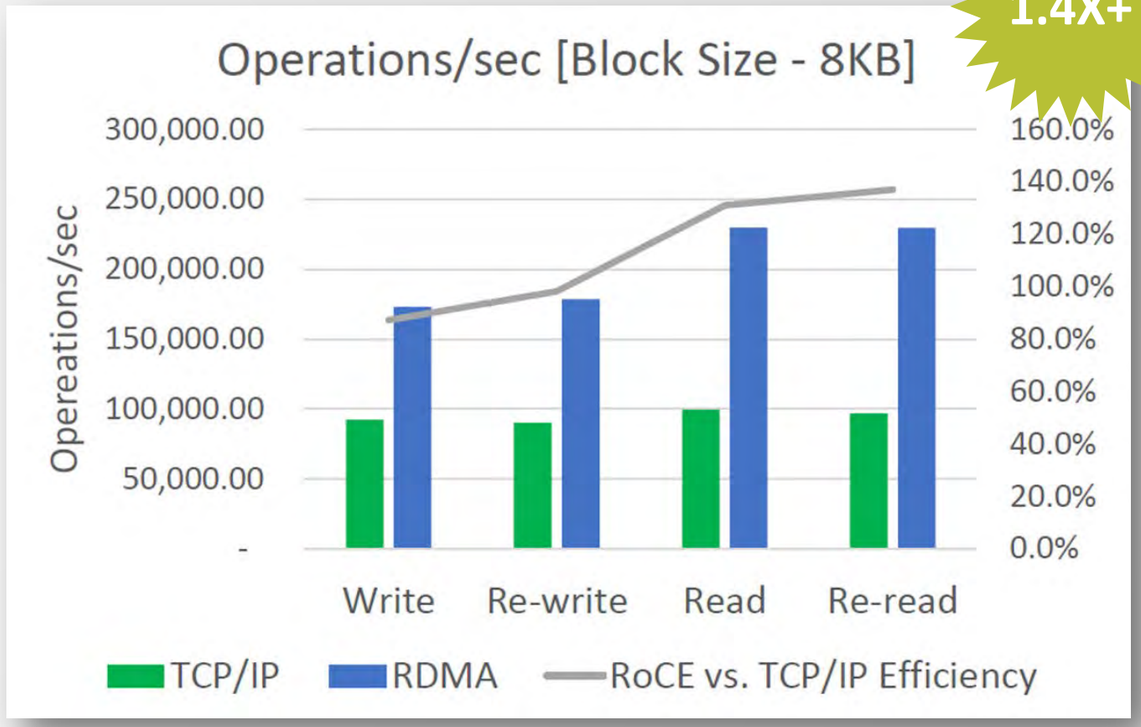
© 2018 Mellanox Technologies | Confidential

Пример: RoCE ускоряет NFS

2X+



1.4X+



Test Configurations:

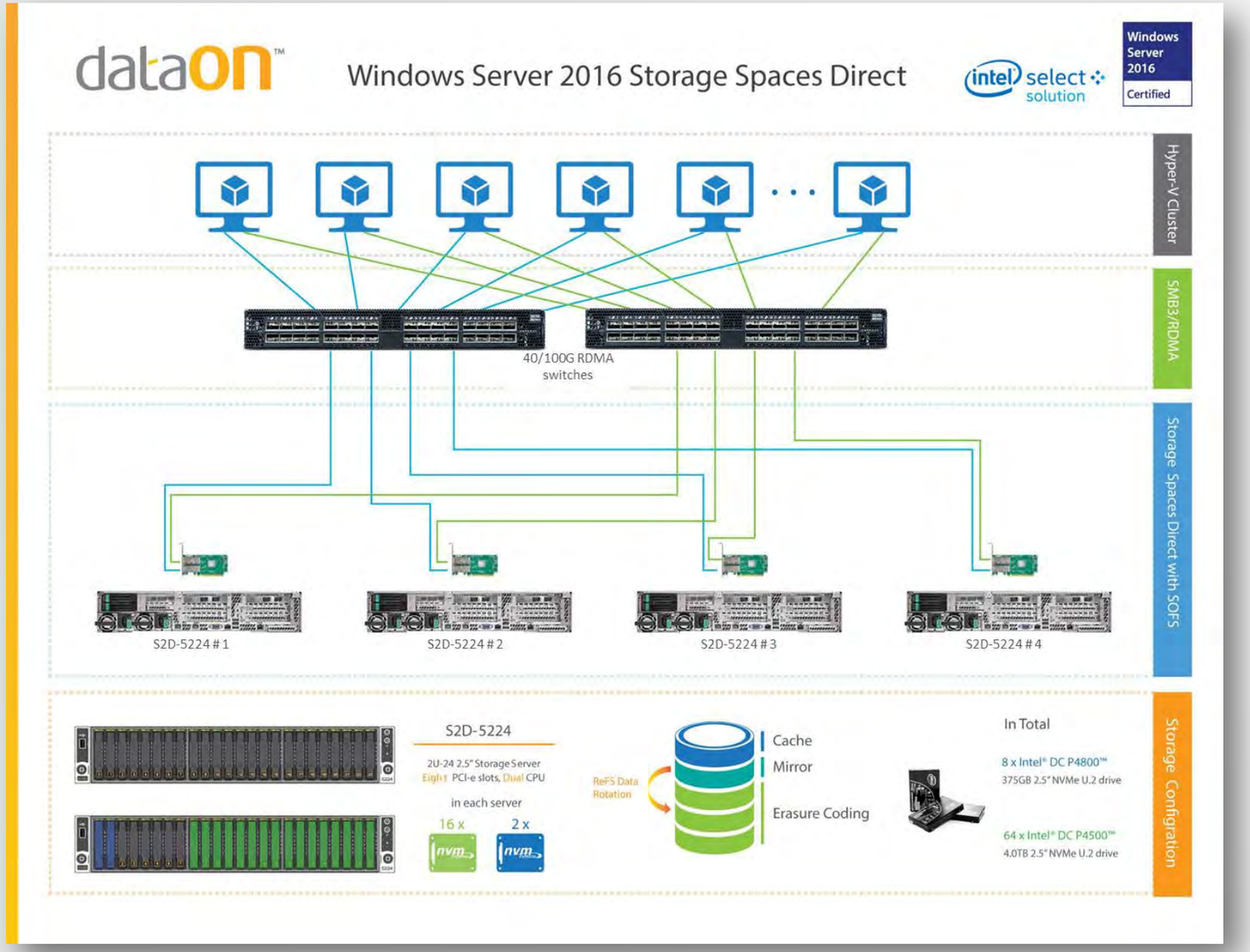
- A set of ‘iozone’ tests were performed on a single client against a Linux NFS server using a tmpfs, so storage latencies are removed from the picture and transport behavior is clearly exposed.
- The client included Intel(R) Core(TM) i5-3450S CPU @ 2.80GHz one socket, four cores, HT disabled 16GB RAM, 1333MHz DDR3, non-ECC HCA together with Mellanox’s ConnectX-5 100GbE NIC.
- The NFS Server included Intel(R) Xeon(R) CPU E5-1620 v4 @ 3.50GHz one socket, four cores, HT disabled 64GB RAM, 2400MHz DDR4 HCA, together with Mellanox’s ConnectX-5 100GbE NIC.
- The Client and the NFS Server were connected to Mellanox’s Spectrum™ SN2700 switch, over a single 100GbE Mellanox LinkX® copper cable.

Blog [Motti Beck: Double Your Network File System Performance with RDMA-Enabled Networking]

© 2018 Mellanox Technologies | Confidential

Сценарии использования ESF

Решение для Microsoft Storage Spaces Direct (S2D)



Source: DataON S2D5224 Deployment Guide

- DataON Hyperconverged Platform S2D 3110/5224
  - Сертификация для Windows Server 2016
  - 40+ Hyper-V VMs на ноду и до 16 нод в кластере
- Windows Server 2016 Datacenter Storage Spaces Direct
- Коммутаторы Mellanox Spectrum switches и адаптеры ConnectX-4/5
  - поддержка RoCE на 10/25/40/50/100GbE
- 3M IOPS с одного кластера на 4 ноды с NVMe SSD дисками; 40+ VM на одну ноду

© 2018 Mellanox Technologies | Confidential

Гибридное облако с Nutanix

Основной ЦОД

Резервный ЦОД

Data Center Interconnect (DCI)

- Объединение нескольких ЦОД с помощью VXLAN/EVPN
- Репликация данных на расстоянии
- Отказоустойчивость и DR

Любой транспорт для DCI:

- темная оптика
- DWDM
- L3 VPN
- L2 VPN

Mellanox TECHNOLOGIES

NUTANIX READY • AHV • • INTEGRATED •

NEO

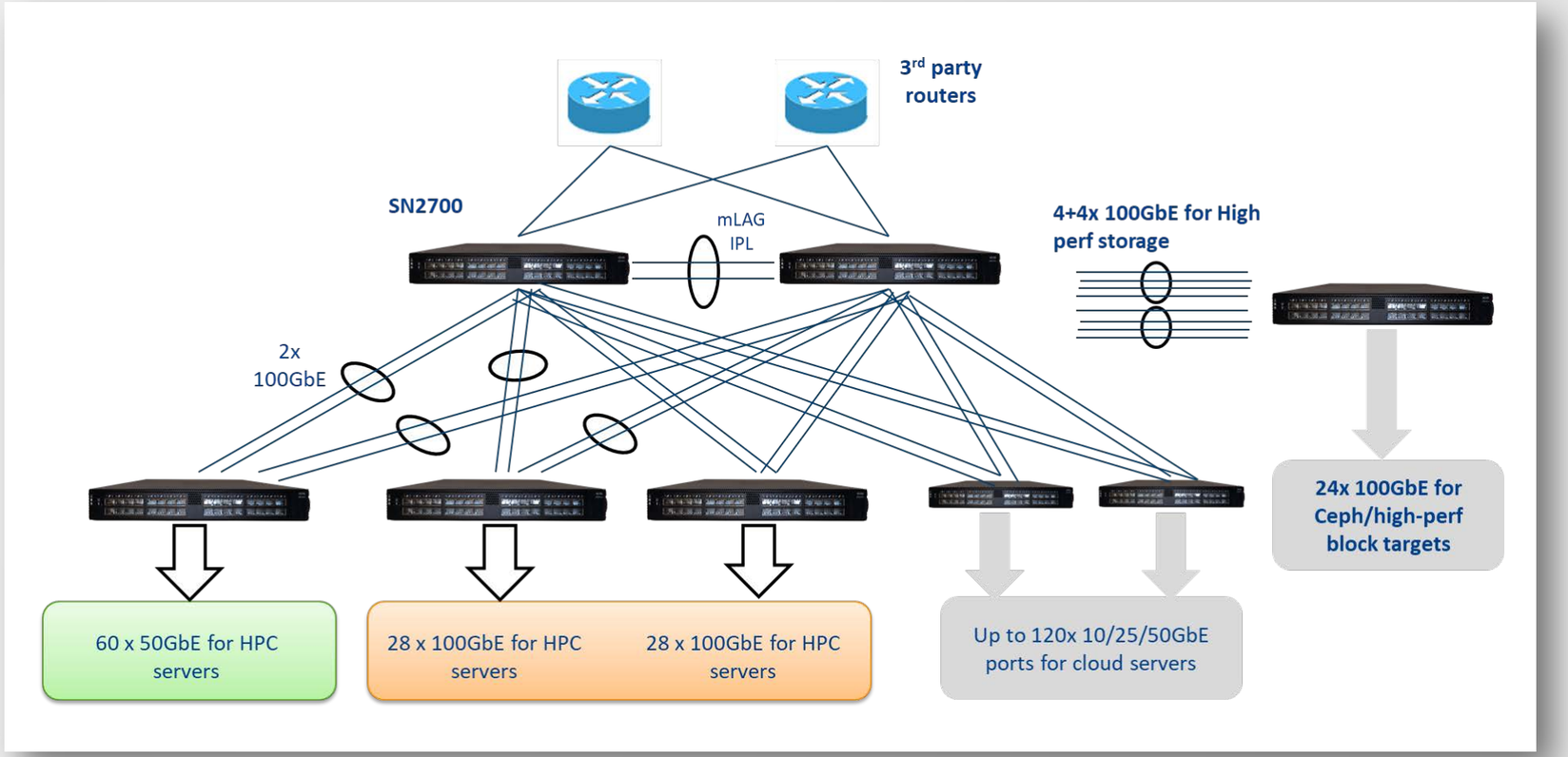
Acropolis

Prism Central

Standard-based VXLAN/EVPN with Automated Network Provisioning

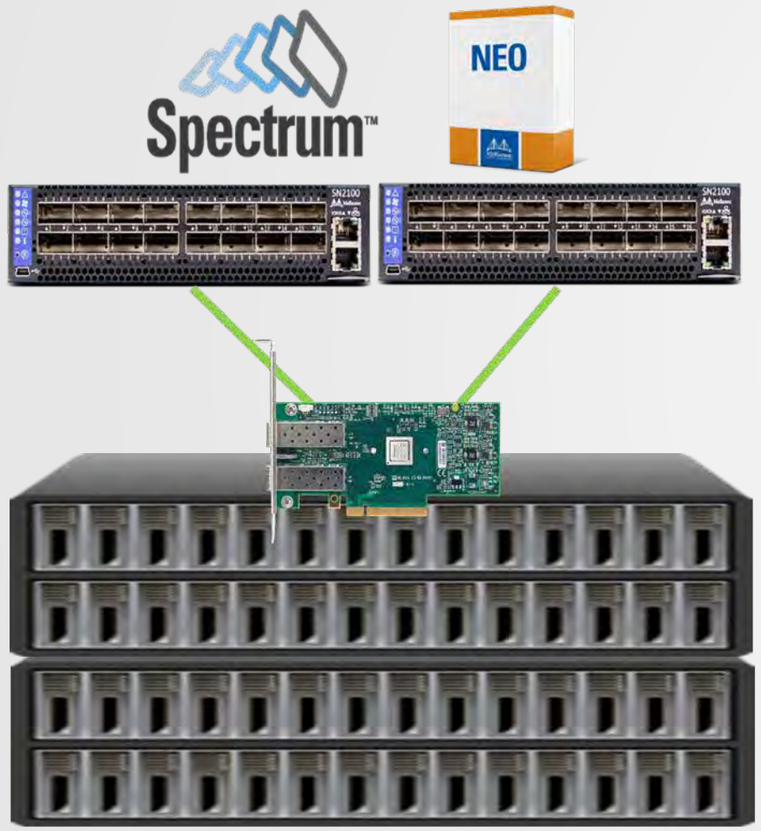
Решение для объектного хранения с Ceph

- Mellanox ESF
  - Leaf/spine
  - 100G RoCE
- Ceph Storage Clusters
- OpenStack Storage
  - Cinder для блочного СХД
  - Swift/S3 для объектного СХД



Solution Brief: Accelerate Redhat Ceph Storage

Готовые решения для ESF



Microsoft Azure

ORACLE RAC – Real Application Clusters

vmware

ceph

hadoop

Apache Spark

dataON DataONSTORAGE

E8 storage Rack-Scale Flash. No Compromise.

Excelero

Hewlett Packard Enterprise

IBM

Micron SOLIDSCALE

NetApp

NUTANIX The Enterprise Cloud Company

PURESTORAGE

- Low Latency
- Fast Deployment
- Guaranteed QoS
- Automated Mgmt

Полный спектр решений для ESF в 2018

Scale-out/Software-defined Storage

ceph | Compuverde | DATERA | Datrium | E8 storage | elastifile | EMC² ScaleIO | Excelero

iguazio | MINIO | nexenta | Open vStorage | PLEXISTOR | PURESTORAGE | QNAP | Quobyte

rozo systems | SCYLLA | StorPool DISTRIBUTED STORAGE | VAST | WEKA.IO | ZADARA STORAGE

Converged Systems & Hyper Converged Infrastructure

dataON | HYPERGRID | iguazio | maxta | NUTANIX The Enterprise Cloud Company | Pivot | SCALE COMPUTING | smartx MAKE IT SIMPLE

STRATOSCALE | ZERO STACK

Media & Entertainment

ELEMENTS ULTIMATE PERFORMANCE | LUMA FORGE | opendrives | pixitmedia | Qumulo

Source: [Mellanox Solution Partner Catalog]

Заключение

Ethernet и ESF - лучшая сеть для СХД в 2018+

- Производительность
  - Скорости 25/50/100GbE vs. 16/32G FC
  - Сверхнизкая задержка в 300нс или 1 мкс end-to-end
  - Акселерация приложений и протоколов с RoCE
- Экономия
  - Низкая стоимость за порт
  - Единая инфраструктура интерконнекта (кабели/трансиверы)
- Универсальность
  - Ethernet - единый транспорт для сети ЦОД и СХД
  - Конвергенция и гиперконвергенция
  - Любые storage-протоколы и приложения: iSCSI/iSER, NVMe, NFS, SMB, S2D, Ceph
- Доступность
  - Готовые tier-1 решения: HPE, NetApp, IBM, Dell/EMC, Nutanix
  - Новые scale-out решения на RoCE/NVMeOF: E8, Excellero, Ceph, S2D, Zadara, etc.
  - Open source решения: Ceph, OpenStack, Linux NVMf, Open vStorage, etc.

© 2018 Mellanox Technologies | Confidential

Спасибо!

© 2018 Mellanox Technologies