오픈 이더넷과 SDN을 통한 Mellanox의 차세대 네트워크 혁신 방안

정연구 Sr.SE

5.Feb 2015

Mellanox TECHNOLOGIES

Connect. Accelerate. Outperform.™

Agenda

- DataCenter Network 트랜드?
- Next Generation Software Defined Networks
- Mellanox Open Ethernet & OCP 소개

데이타센타 Network Trends

The Evolution of SDN

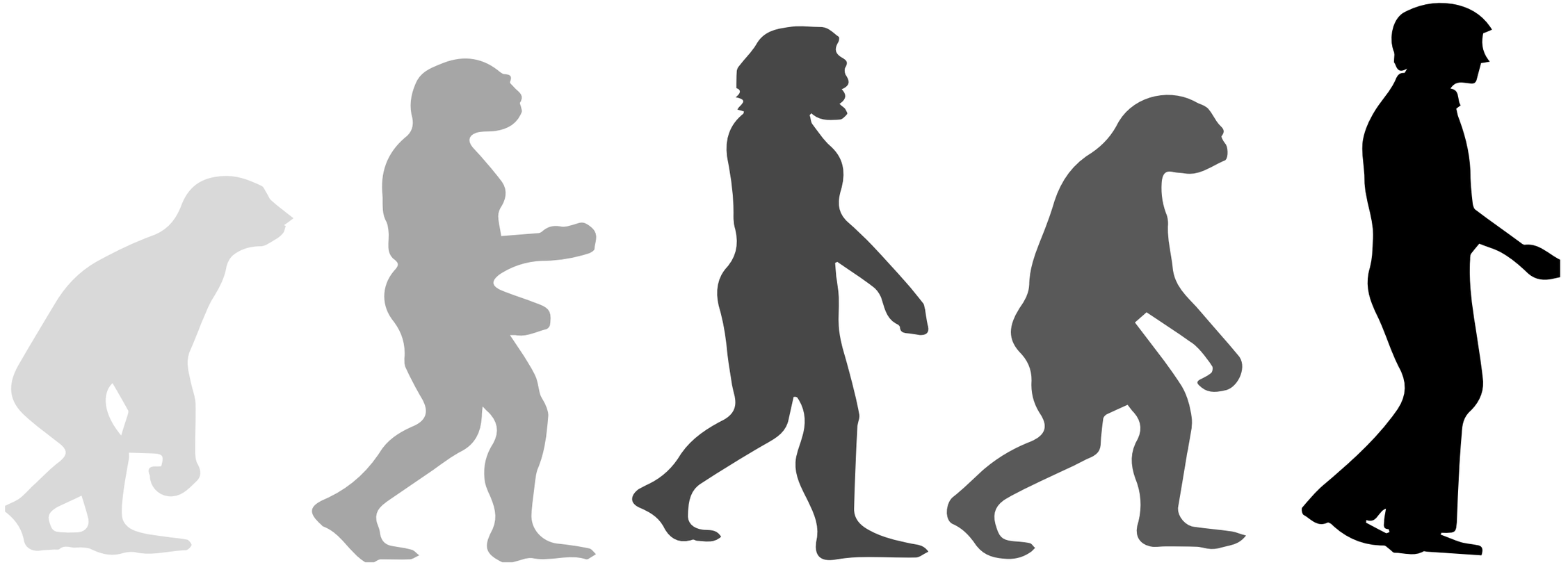
Switches are a build of closed software being sold as a package from switch vendors

Stanford guys wanted to change the networking world using OpenFlow

“SDN” became a popular term – everything is “SDN”

Switch vendors made OpenFlow “an important protocol”

The networking industry finally delivers on the promise of “SDN”



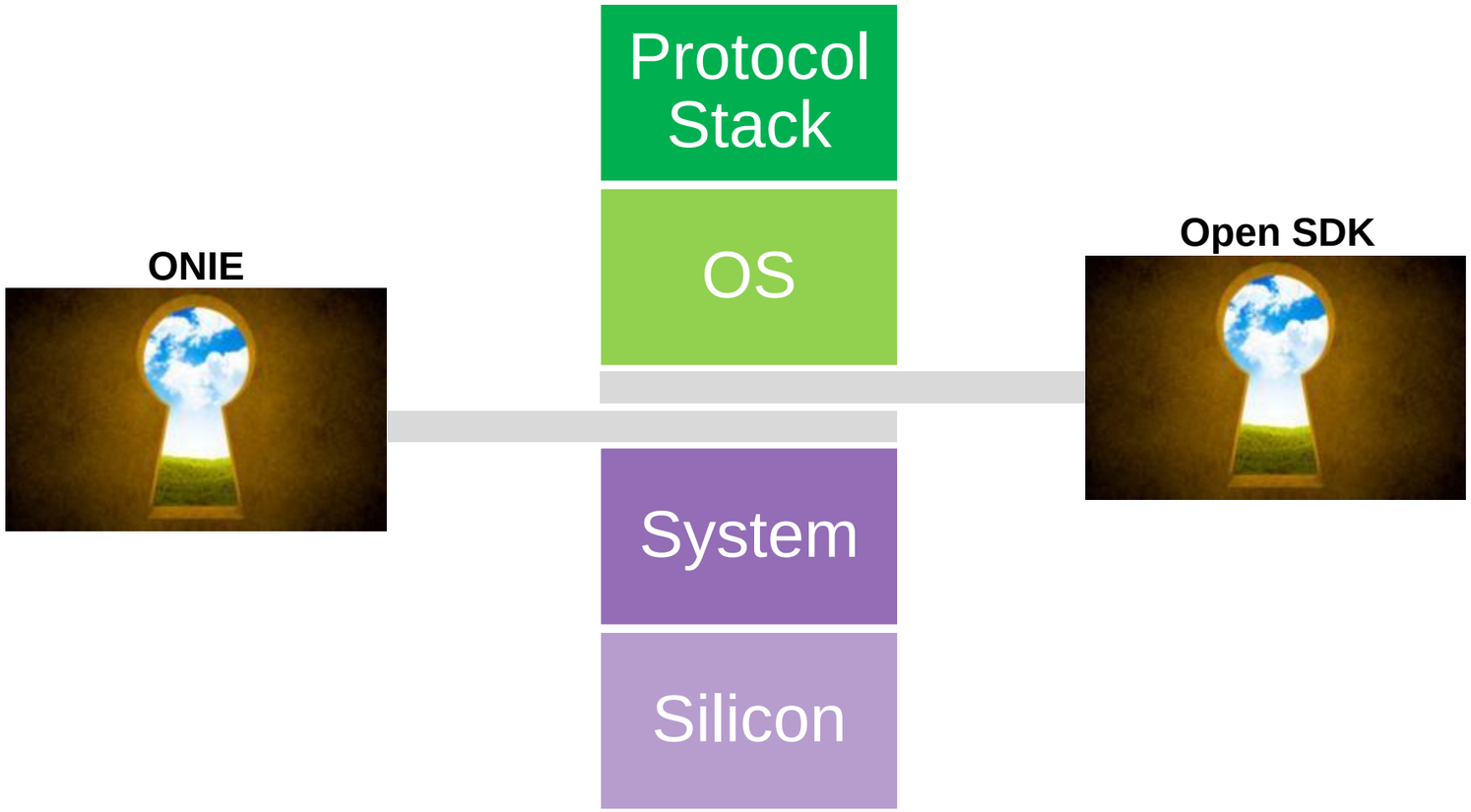
OPEN ETHERNET

OPEN Compute Project

onie

2010 **2011** **2012** **2013** **2014**

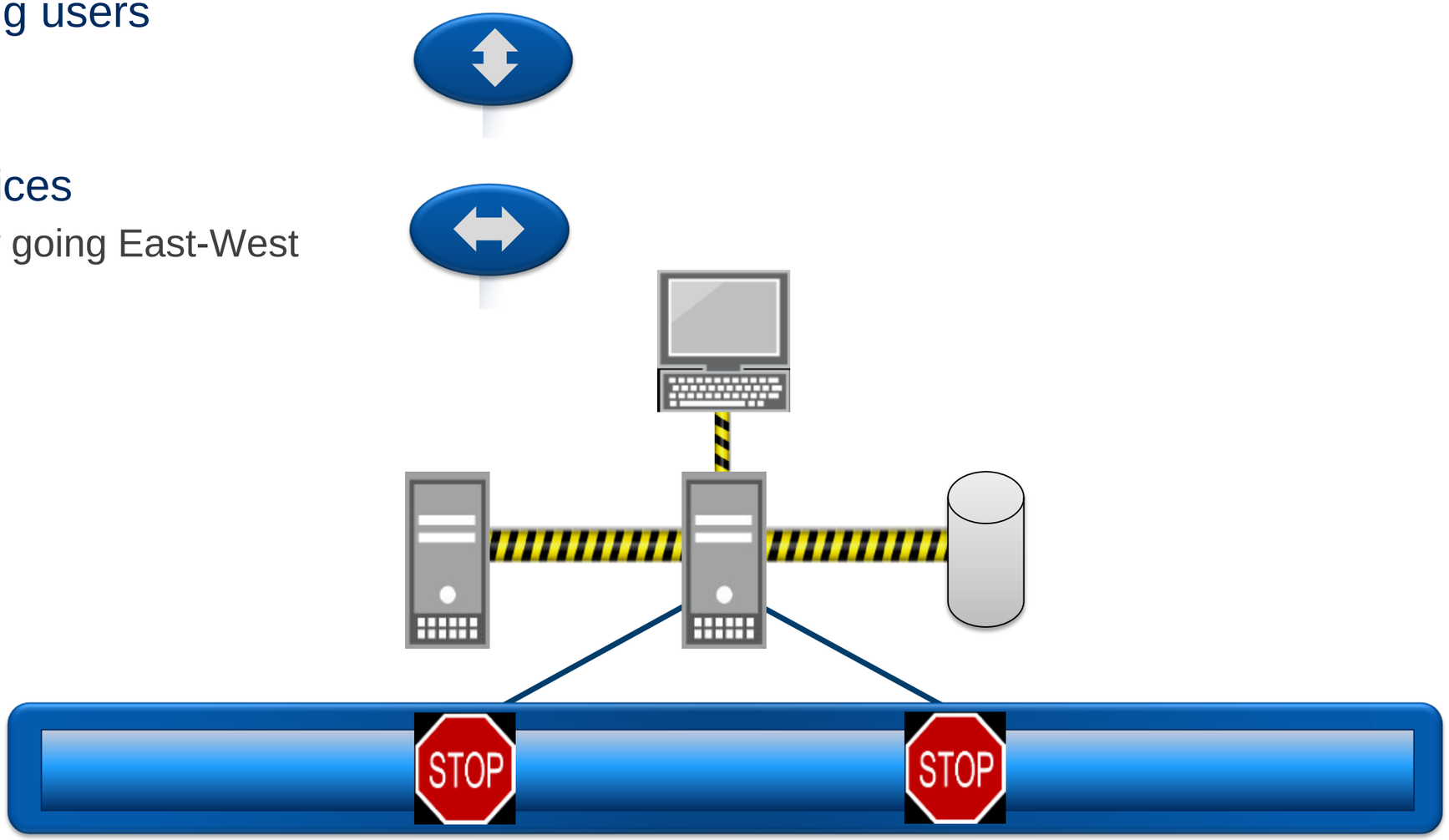
기술 트랜드 : from Closed to Open



OPEN ETHERNET™

기술 트랜드 : Interconnect 네트워크의 중요성 증대

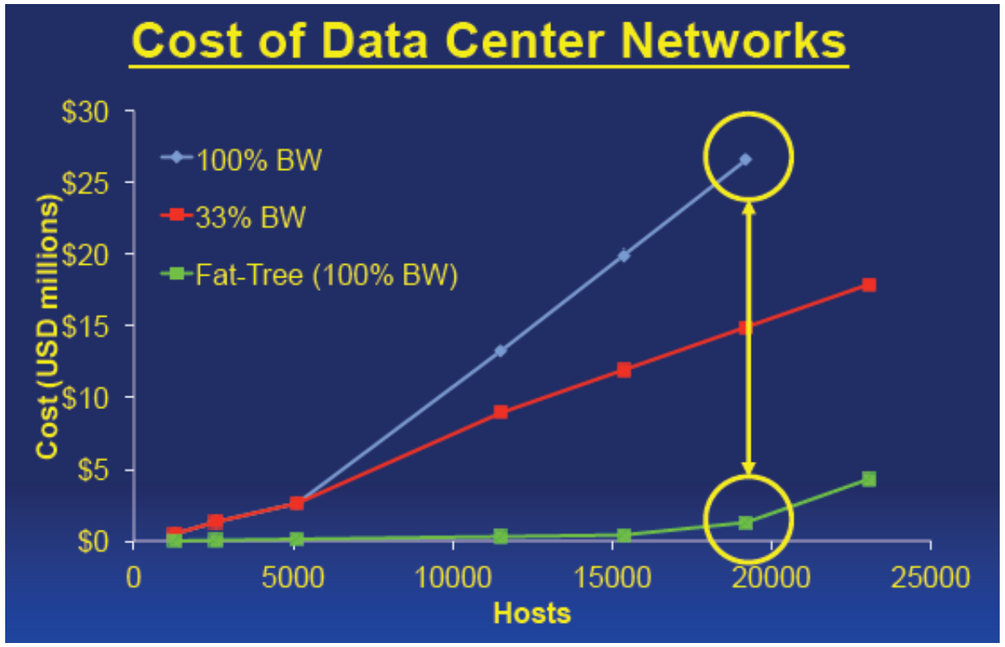
- Traditional role – Connecting users
  - 80% of traffic is North-South
- New role – Connecting devices
  - 75% of network traffic is now going East-West
- New challenges
  - Storage bottlenecks
  - Latency and slow I/O



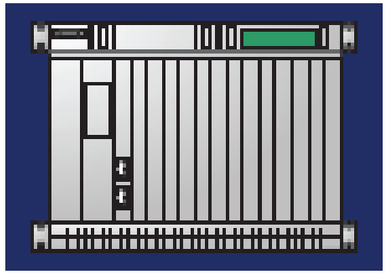
Application Running

기술 트랜드 : from Modular to Scalable

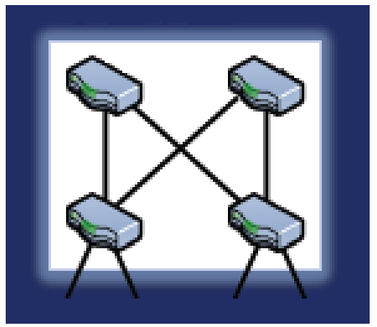
- Data Center Network Requirements
  - Cost and absolute performance
  - Packaging
  - Energy/heat
  - Fault tolerance
- Fat-tree as basis for delivering on above challenges
  - Regularity of the structure allows simplifying assumptions to address challenges above
  - Fat-tree built from 48-port switches support 27k hosts
  - Holds promise for cost, performance, energy



Source: Scale and Efficiency in Data Center Networks, Google / UC San Diego



Modular Switch



Virtual Modular Switch

Expensive, Complex

Cost Effective, Flexible

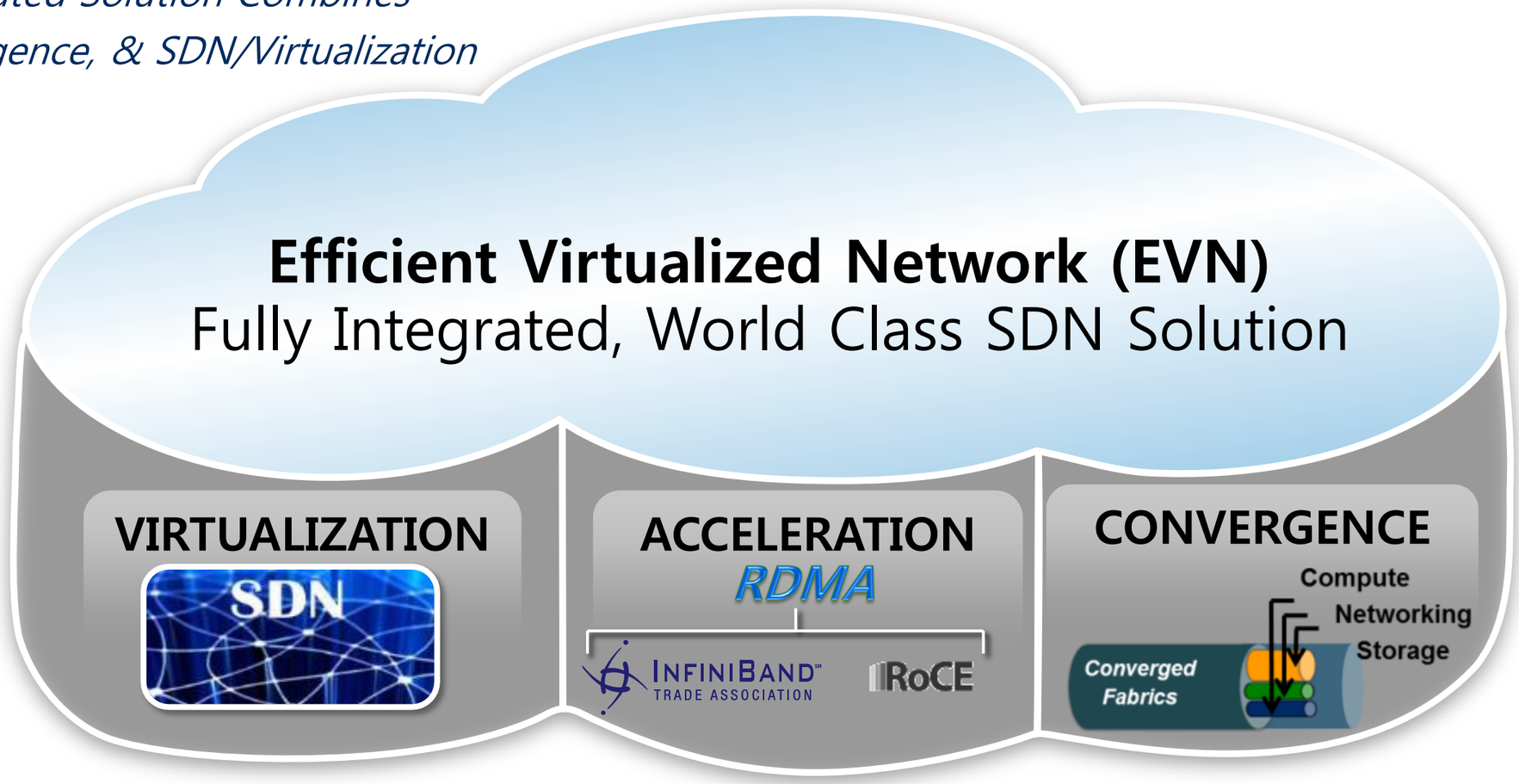
Next Generation Software Defined Networks

차세대 Software Defined Networks

EVN: Efficient Virtualized Network

Fully Integrated Solution Combines

RDMA, Convergence, & SDN/Virtualization



openstack

Windows Azure

vmware vCLOUD POWERED

redhat.

EVN: More than SDN - Efficient Clouds Need an Efficient Virtualized Network

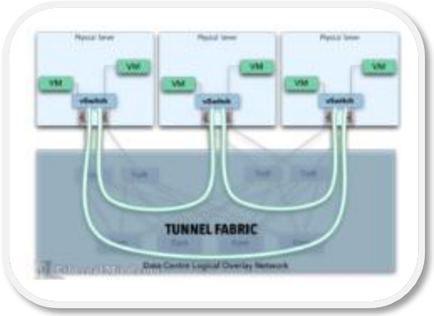
SDN 네트워크를 가속화 하기위한 세가지 요소

Software Defined Networks

=

Virtual Network Management

+



Overlay Network Tunnels

+

OpenFlow

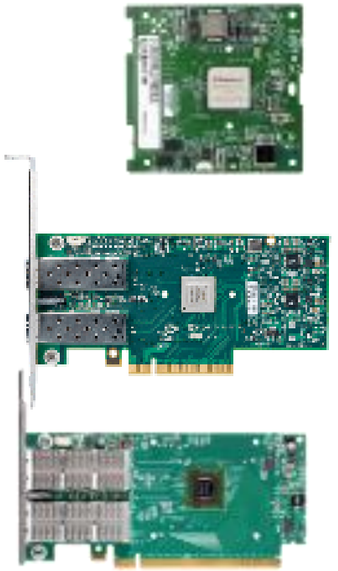
1. Centralized Software Based Control Plane
   - Enables network virtualization
2. Overlay Networks – NVGRE/VXLAN/ Geneve
   - Isolation, Scalability, Simplicity
   - Mellanox accelerates overlay networks to offer bare metal speed
3. Industry Standard API – OpenFlow
   - Enables an industry ecosystem and innovation

최초의 End to End SDN Interconnect solution

– Interconnect Offload Engines

- **Challenge**: Implementing overlay networks in software dramatically increases overheads and costs
- **Solution**: Use interconnect offload engines to handle all networking operations up to the VM
- **Benefit**: Reduce application cost, cloud CAPEX and OPEX

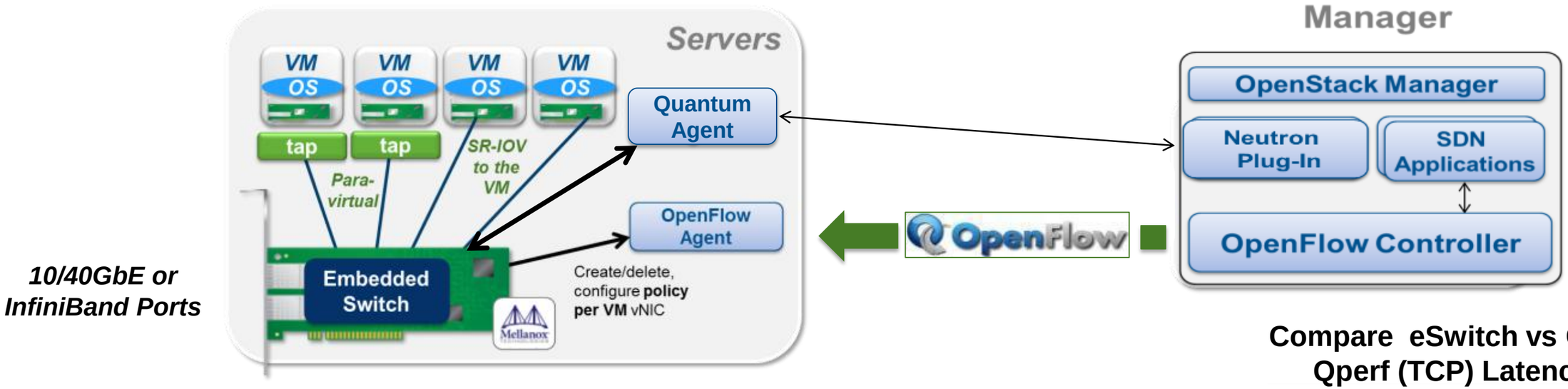
ConnectX®-3 PRO



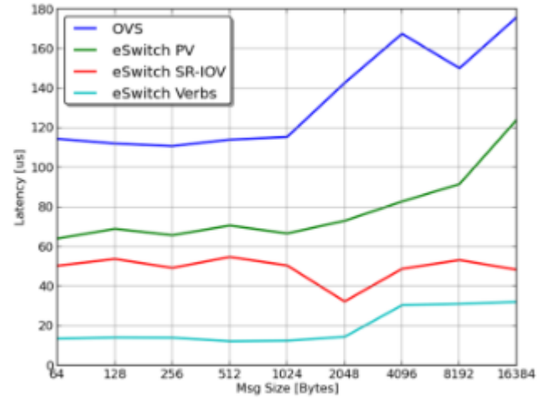
The Foundation of Cloud 2.0

The World's First NVGRE / VXLAN Offloaded NIC

최초의 End to End SDN Interconnect solution
- Mellanox 네트워크 가상화 (Neutron) 플러그 인



Compare eSwitch vs OVS
Qperf (TCP) Latency



- **Deliver the best application performance**
- **Simpler to manage - Real-time NIC provisioning via OpenFlow**
- **Provide tenant & application security/isolation**

The Benefits of VM Provision & Fabric Policy in Hardware
Isolation, Performance & Offload, Simpler SDN

ConnectX-3 Pro 오버레이 네트워크를 기속화

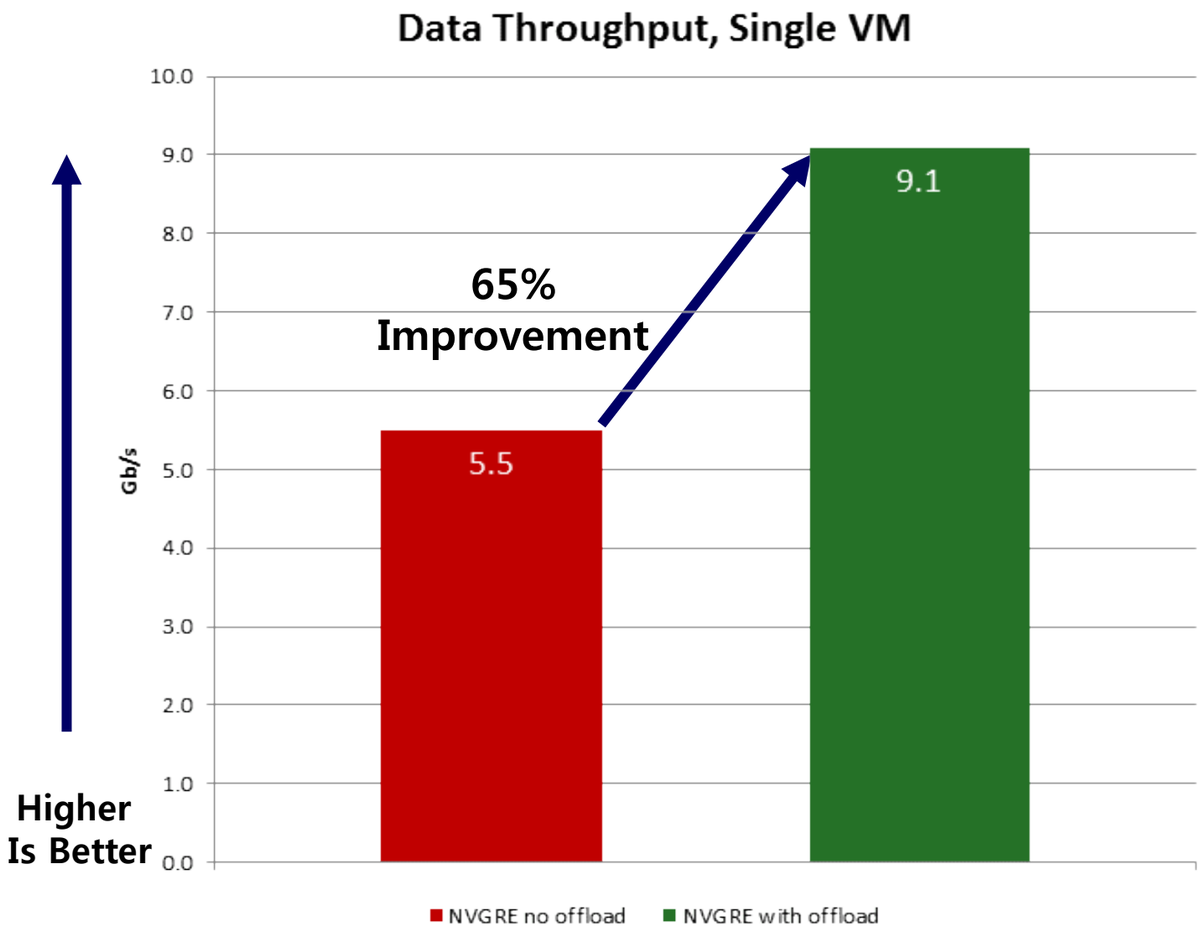
OpenFlow
Virtual Network
Management API

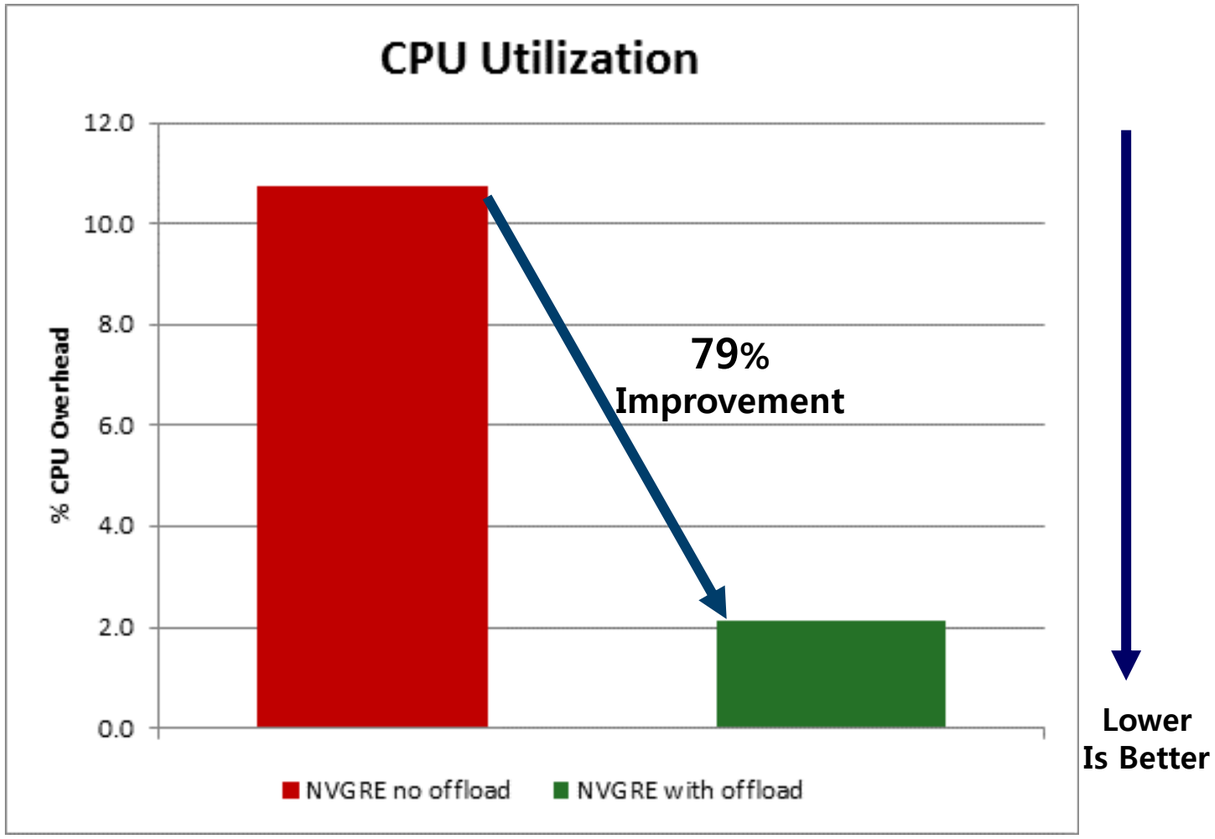
NVGRE/VXLAN Overlay Networks

Virtual Overlay Networks Simplifies Management and VM Migration
ConnectX-3 Pro
Overlay Accelerators Enable Bare Metal Performance

Overlay Network Virtualization: Isolation, Simplicity, Scalability

ConnectX-3 Pro 10GbE 어댑터를 통한 클라우드 오버레이 가속화 결과

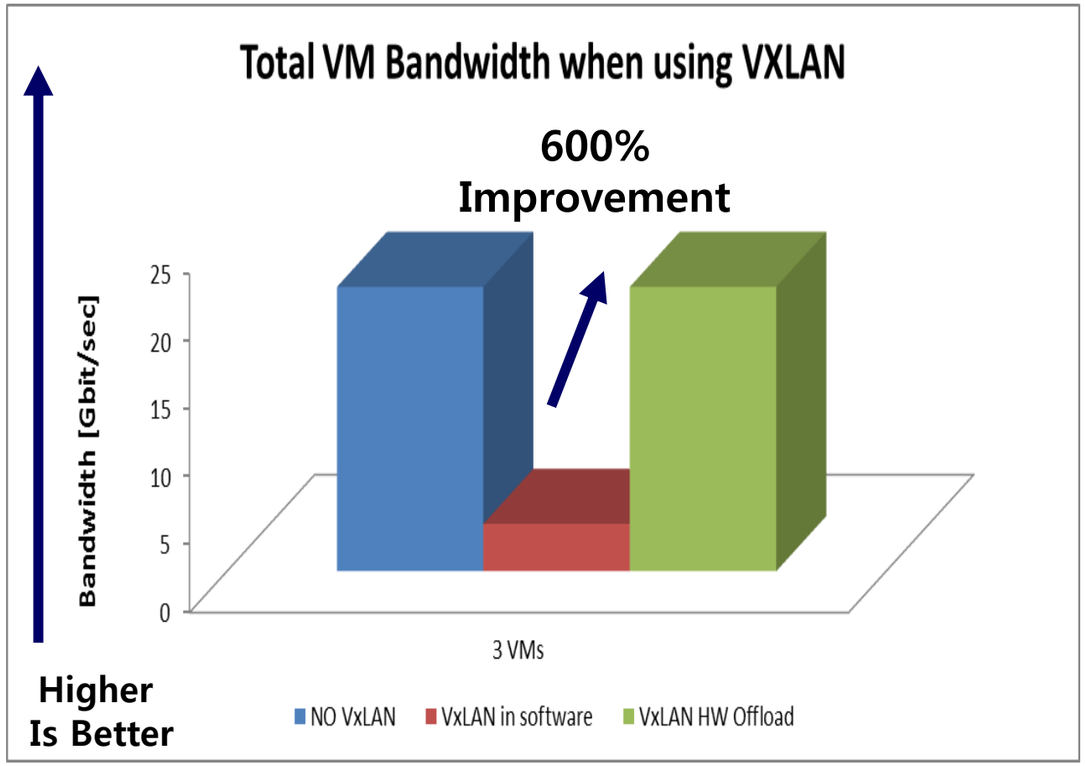


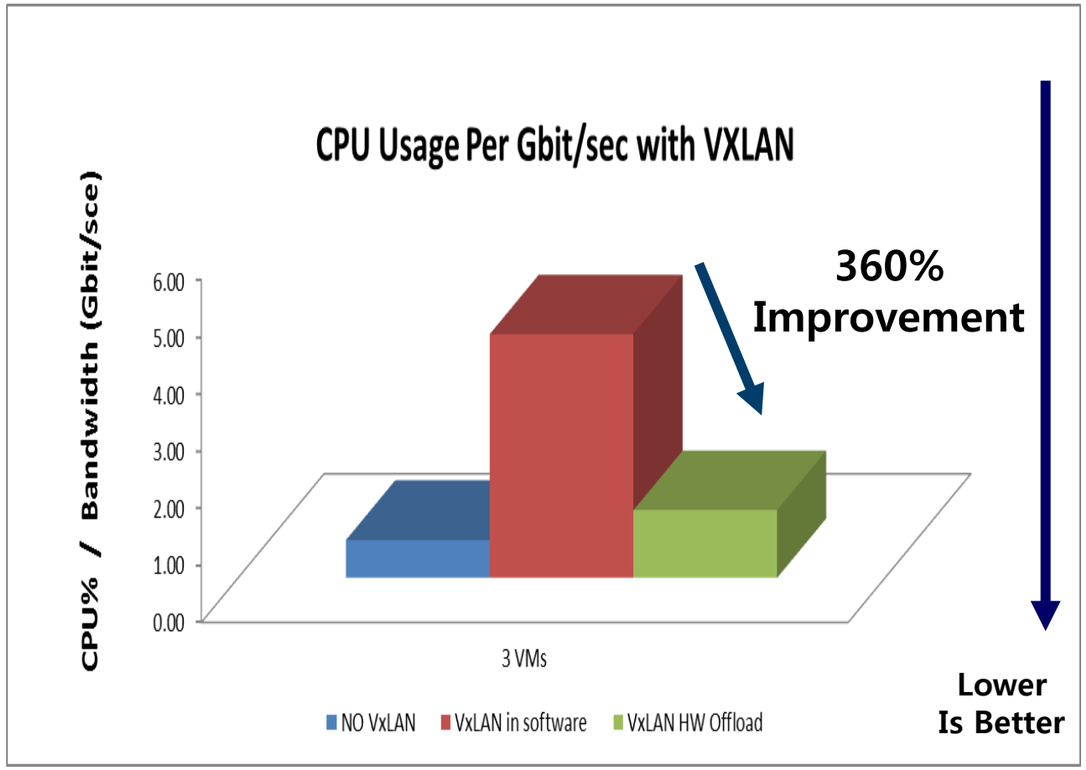


NVGRE Initial Results

Higher Throughput for Less CPU Transport Overhead

ConnectX-3 Pro 10GbE – VXLAN 오버레이 네트워크 가속화 결과





VXLAN Initial Results

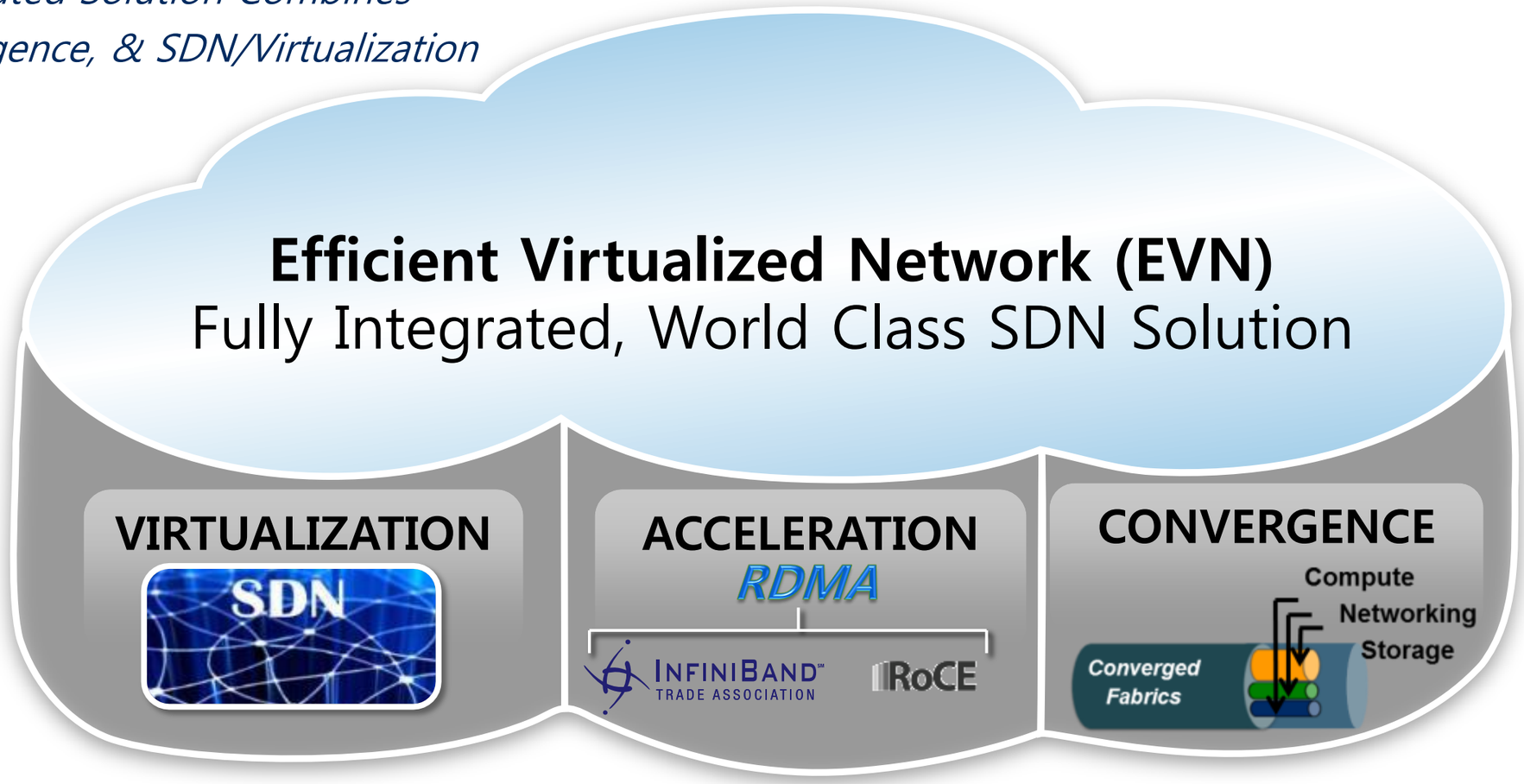
Higher Throughput for Less CPU Transport Overhead

Next Generation Software Defined Networks

EVN: Efficient Virtualized Network

Fully Integrated Solution Combines

RDMA, Convergence, & SDN/Virtualization



openstack

Windows Azure

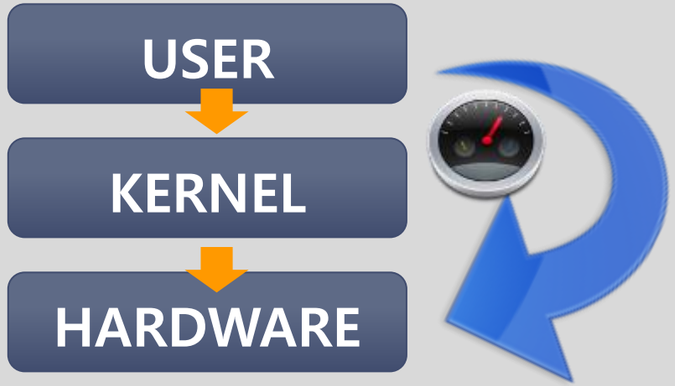
vmware vCLOUD POWERED

redhat

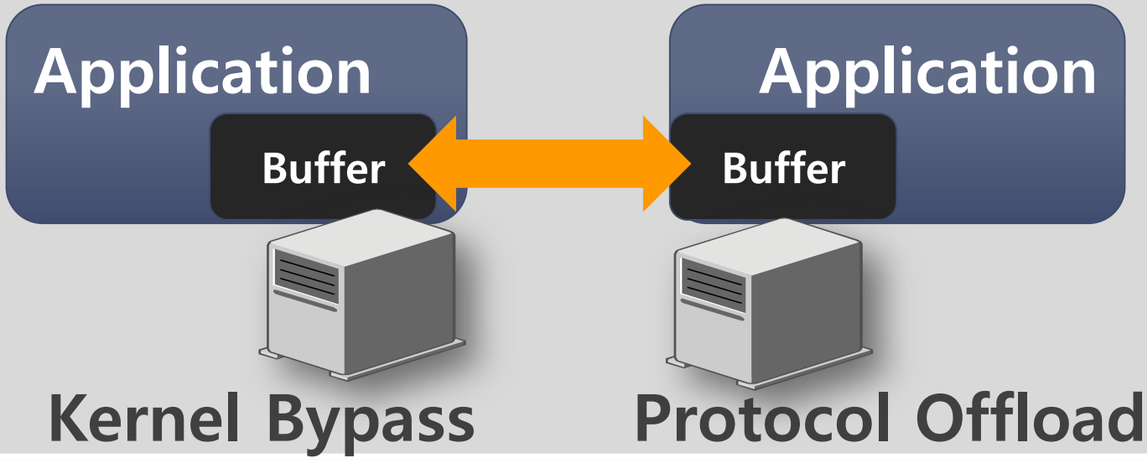
EVN: More than SDN - Efficient Clouds Need an Efficient Virtualized Network

RDMA: 데이타센타 내 자원 활용율을 높이는 가장 중요한 인터컨넥트 기술

ZERO Copy



Remote Data Transfer



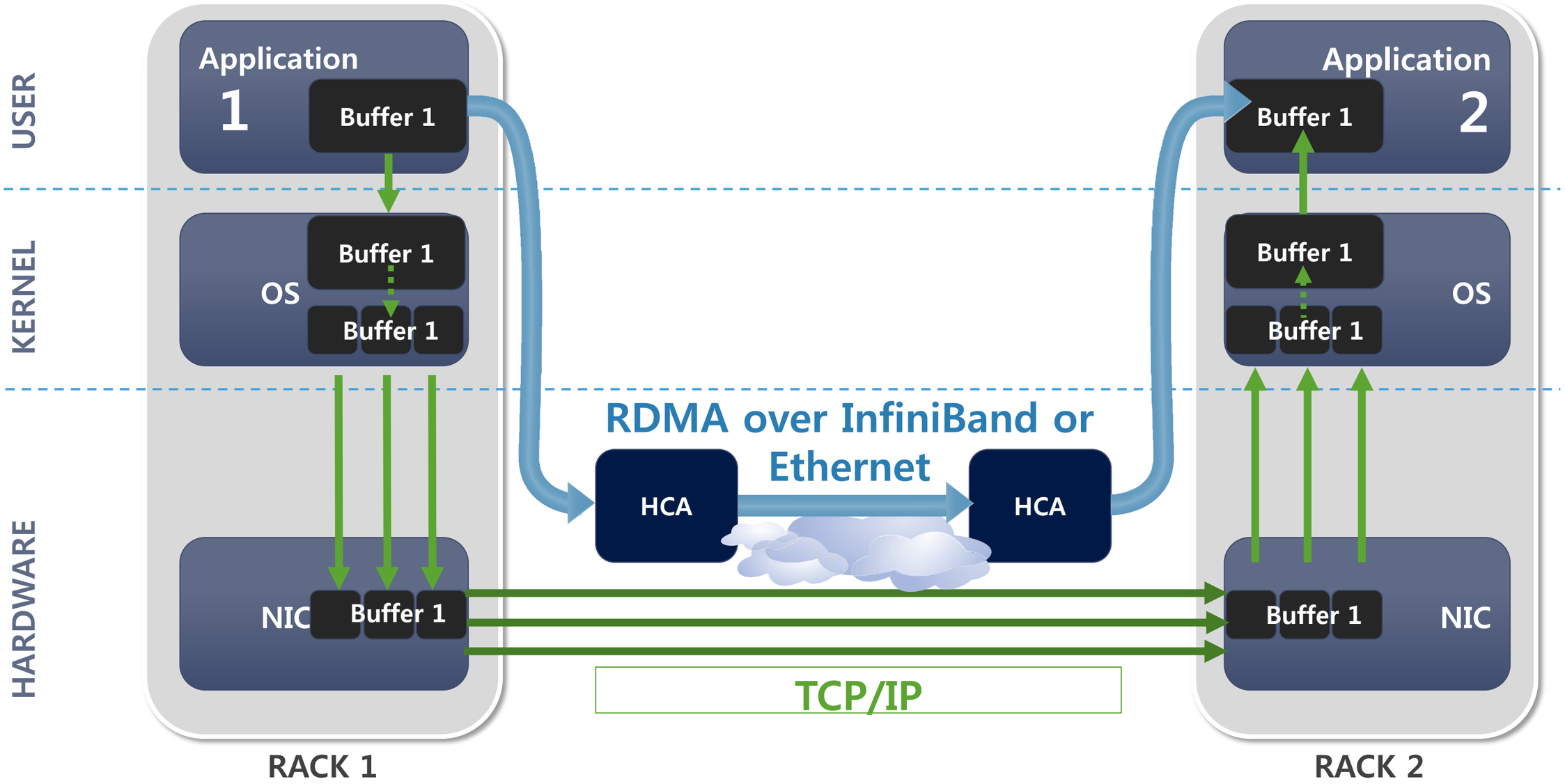
Low Latency, High Performance Data Transfers

InfiniBand - 56Gb/s

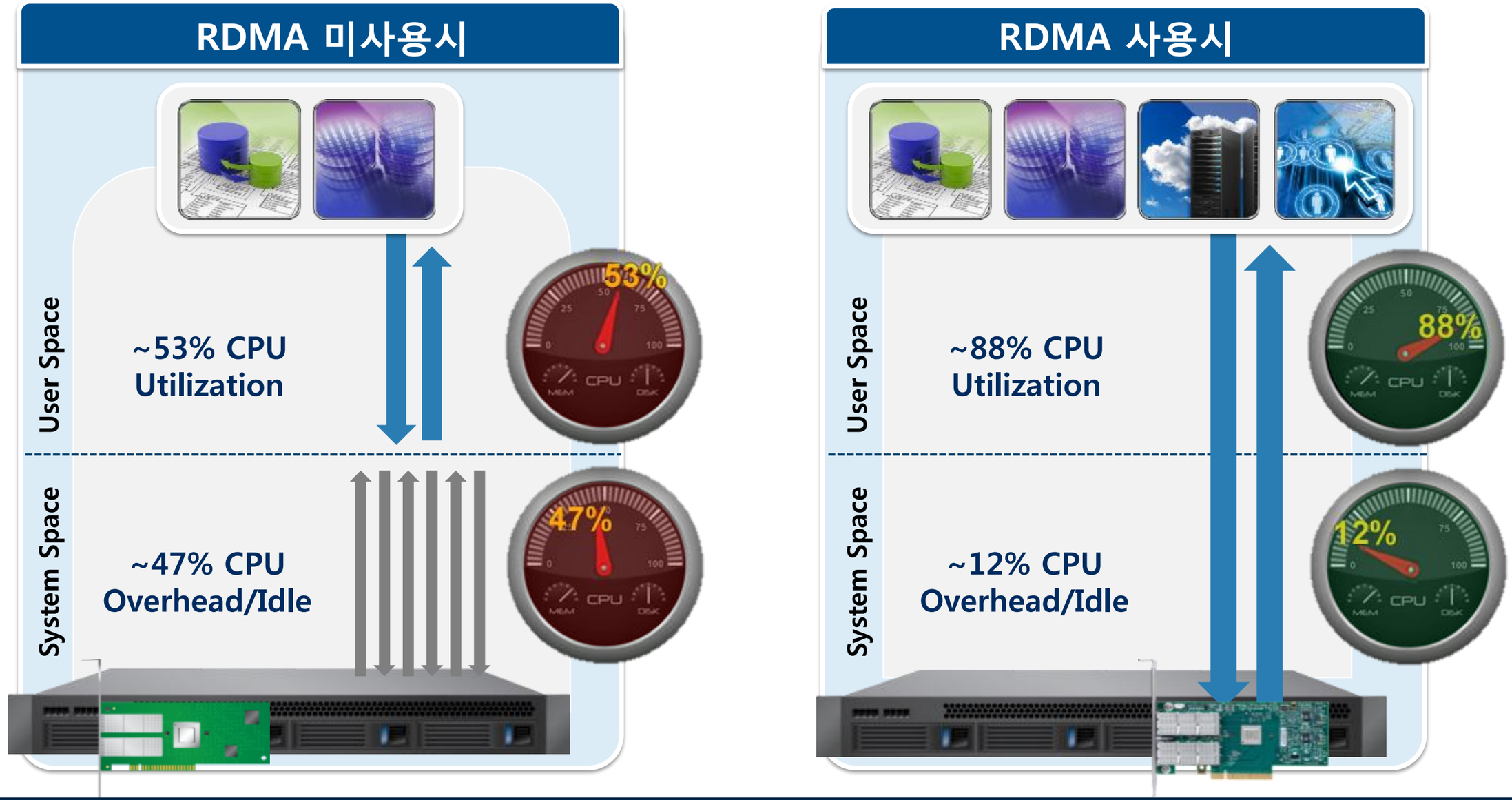
RoCE* – 40Gb/s

* RDMA over Converged Ethernet

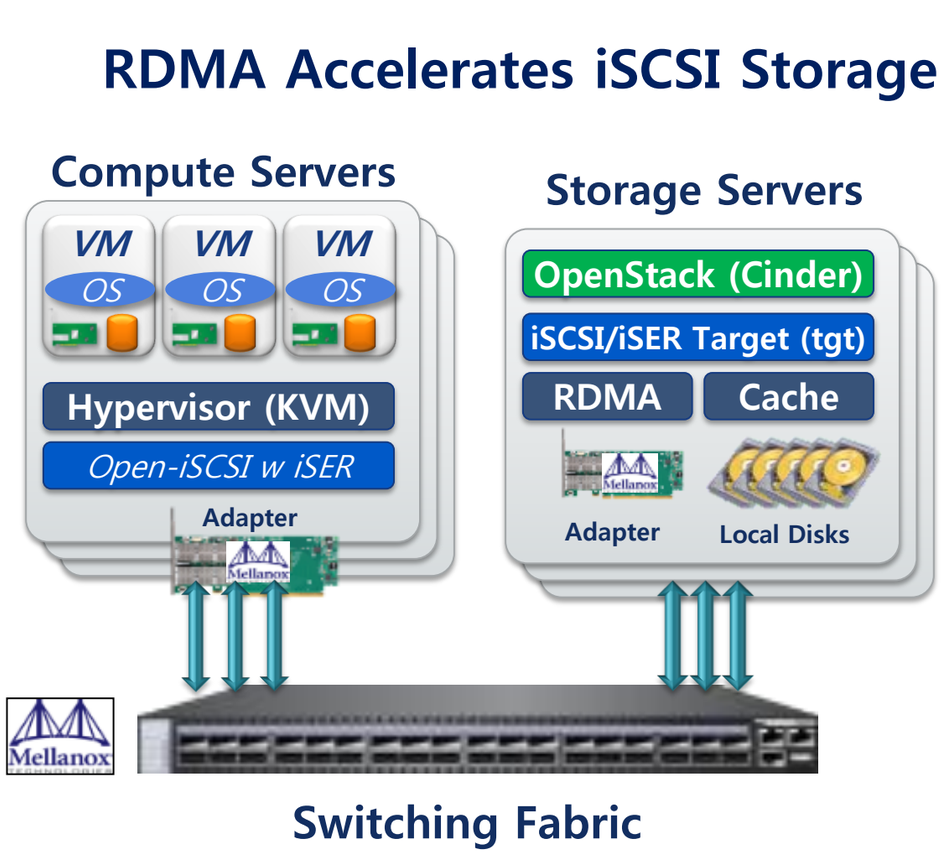
RDMA – 동작 방식

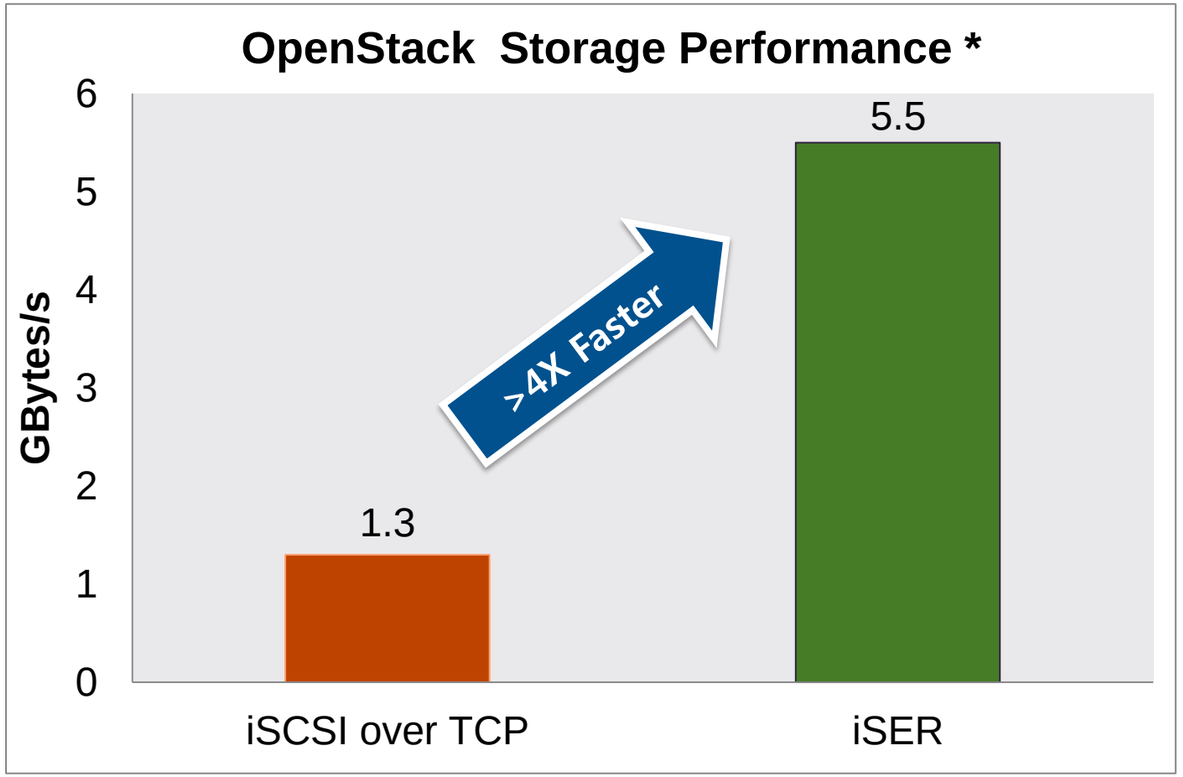

USER
KERNEL
HARDWARE
Application 1
Buffer 1
OS
Buffer 1
Buffer 1
NIC
Buffer 1
RDMA over InfiniBand or Ethernet
HCA
HCA
Application 2
Buffer 1
Buffer 1
OS
Buffer 1
Buffer 1
NIC
TCP/IP
RACK 1
RACK 2

인터커넥트 기술 적용에 따른 CPU 효율 차이



iSCSI over RDMA (iSER): 4배 이상의 Throughput

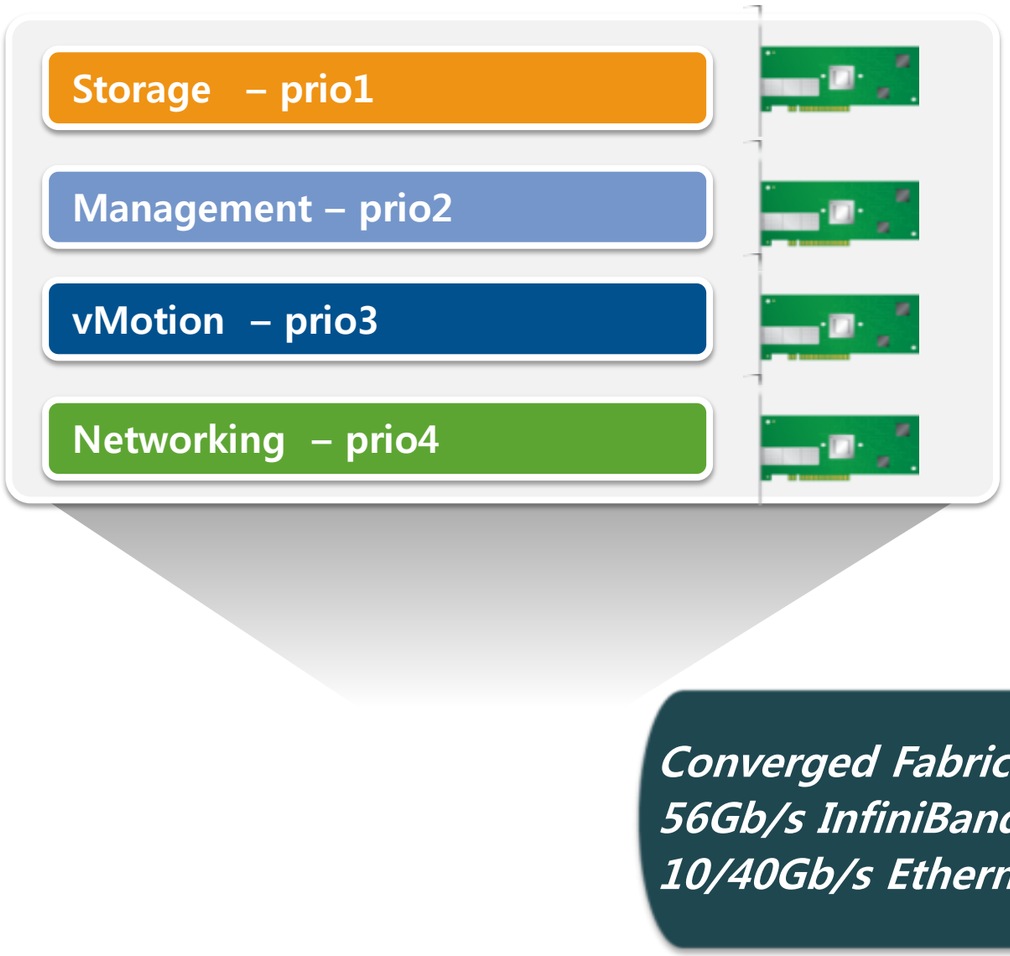




* iSER patches are available on OpenStack branch: https://github.com/mellanox/openstack

Built-in OpenStack Components/Management & Cinder/iSER to Accelerate Storage Access

컨버전스: 스토리지 만을 위한 전용 네트워크의 필요성 제거



Single Interconnect for Compute, Networking, Storage

RDMA: InfiniBand & Ethernet (RoCE*)

There is no Fibre Channel in the Cloud!

Windows Server Hyper-V

vmware

KVM

* RoCE: RDMA over Converged Ethernet

Web 2.0, Public & Private Clouds Converging on Fast RDMA Interconnects

Mellanox 기반 클라우드 랙 솔루션 building Block

CloudX Rack

Mellanox Cables

Industry Standard Servers

- Each equipped with Mellanox ConnectX-3 10GbE/40GbE/InfiniBand Adapter
- Running Cloud Stack / Hypervisor VMware, Hyper-V, KVM, or OpenStack
- Potentially using local disks/SSD for cloud storage or cache

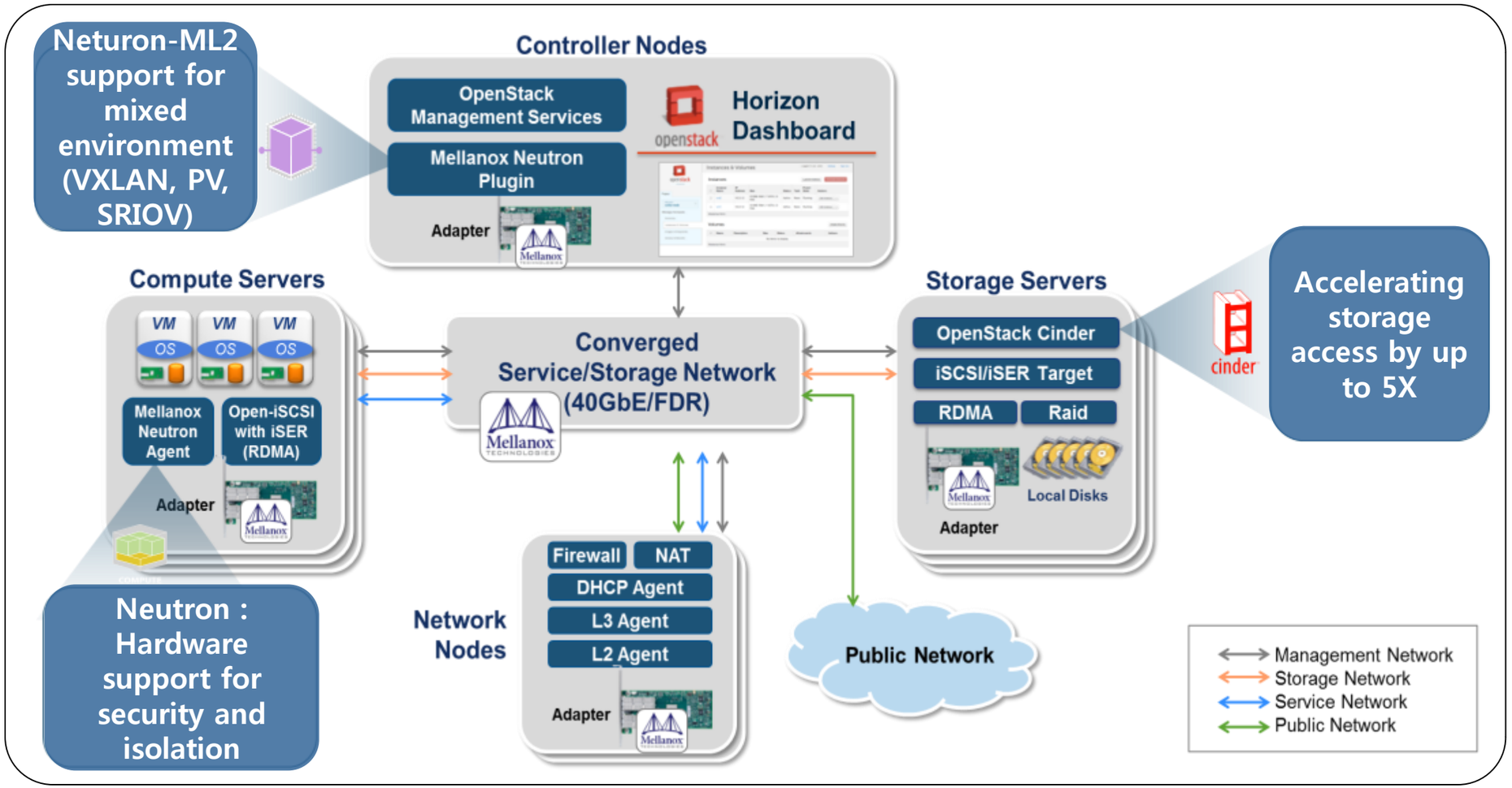
Mellanox High Performance & Density Switches

- Connected via 10GbE/40GbE/InfiniBand
- Delivering worlds' best cost/performance

High Performance RDMA Attached Storage

- From variety of partners
- And/or software based storage appliances (SDS)

스위치와 어댑터에 대한 전체적인 OpenStack 통합



Integrated with Major OpenStack Distributions

Mirantis Pure Play OpenStack — Mirantis Partner

juju — ubuntu Certified

redhat CERTIFIED Partner

In-Box With Havana and Ice House

THE NINTH OPENSTACK RELEASE ICEHOUSE

InfiniBand Trade Association

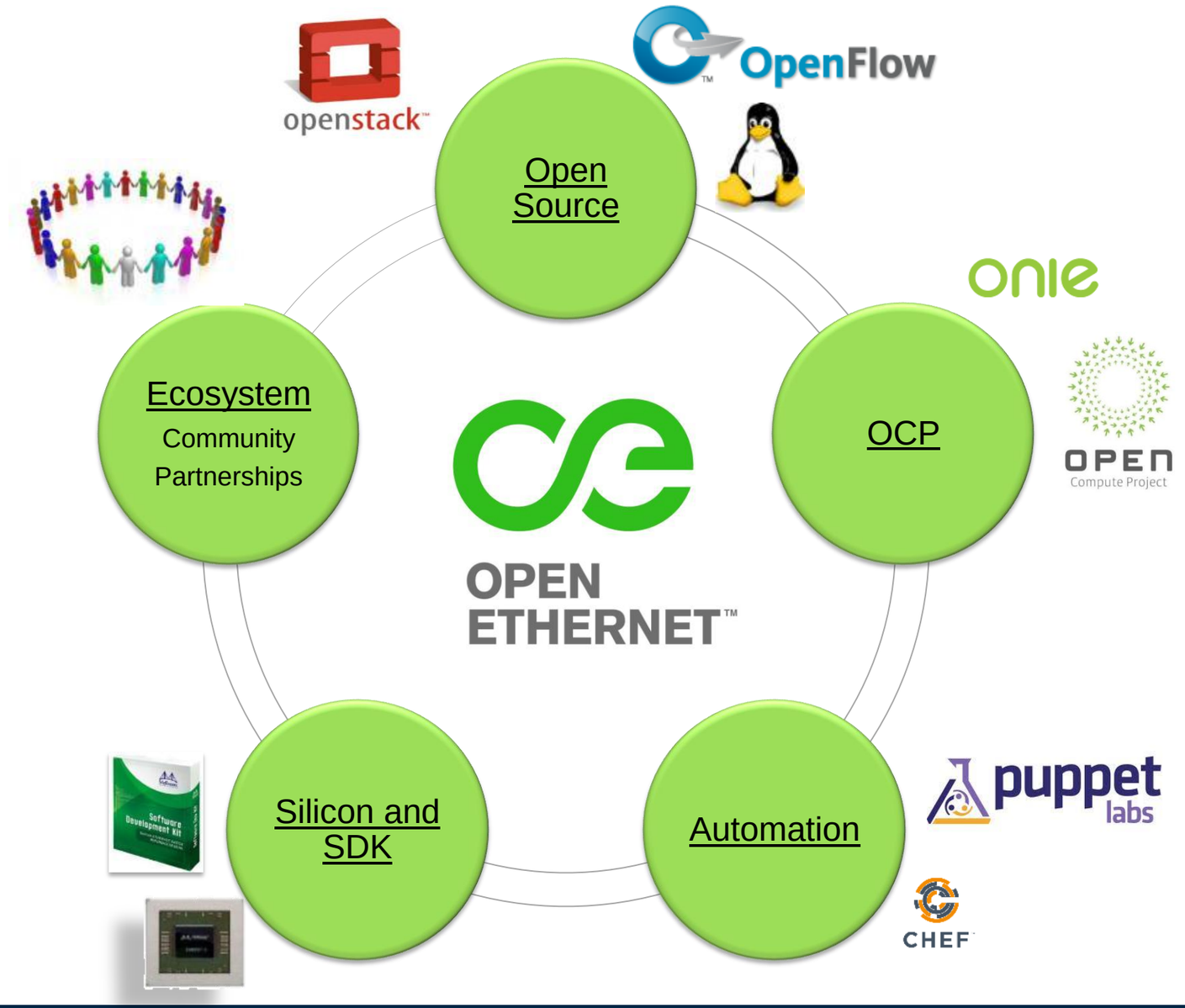
Ethernet

OpenStack Plugins Create Seamless Integration , Control, & Management

Mellanox Open Ethernet & OCP 소개

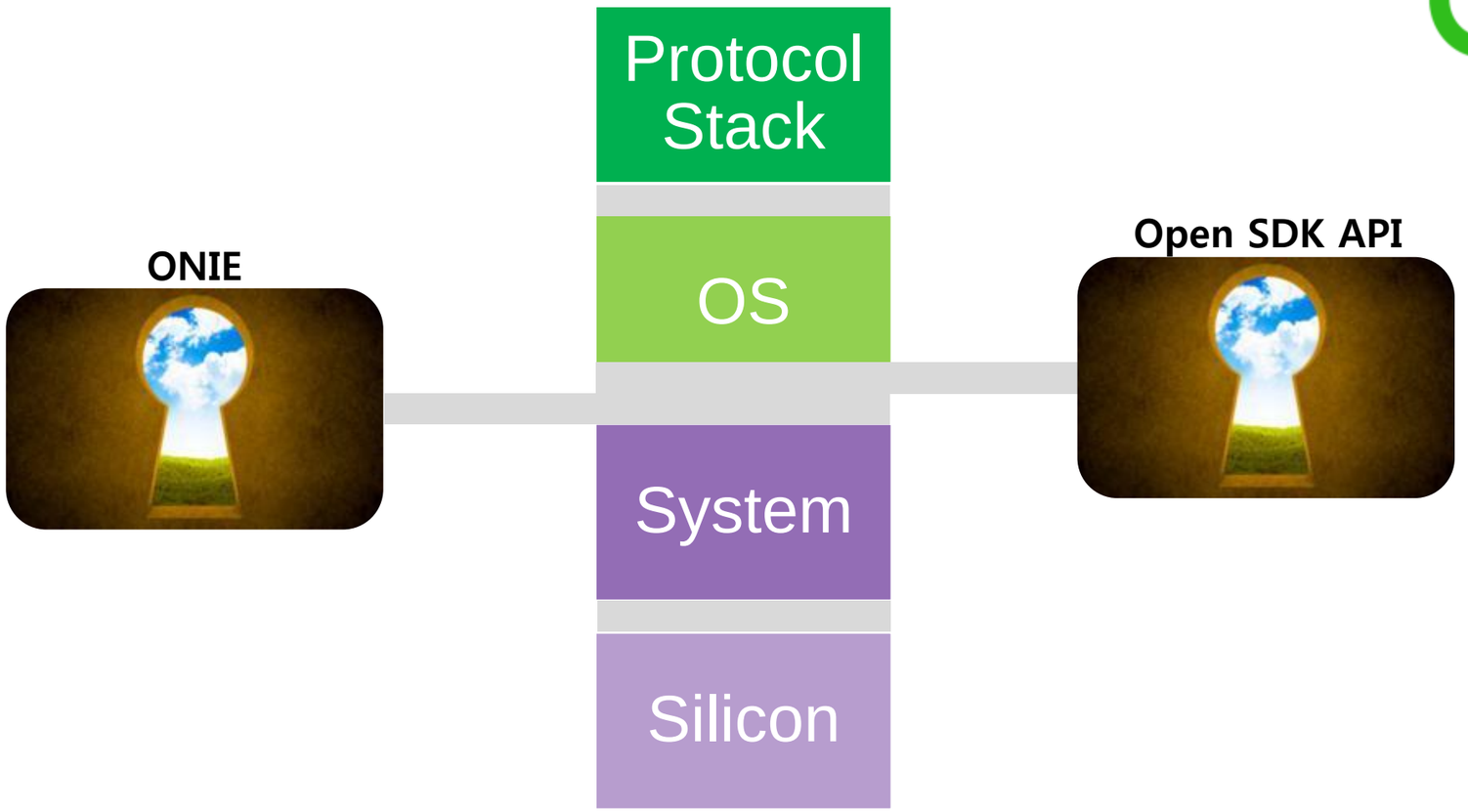
오픈 이더넷 이란?

- **Provide better control over network resources for the various, changing data center needs by**
- **Separation of hardware and software**
- **Free choice of**
  - The most adequate platform
  - With the most suitable software
  - Run it on the most efficient hardware
- **Free choice of**
  - Switch Silicon
  - Hardware
  - Operating System
  - Protocol Stack



하드웨어와 소프트웨어의 분리

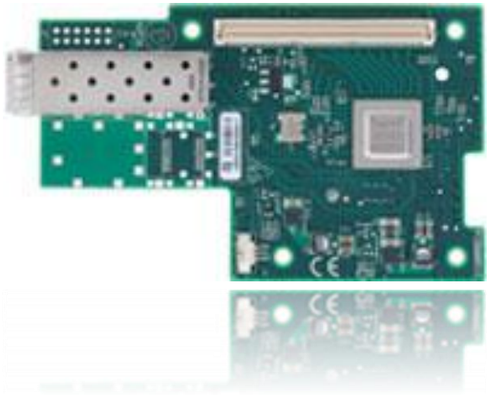
- Open SDK + ONIE (Open Network Install Environment)

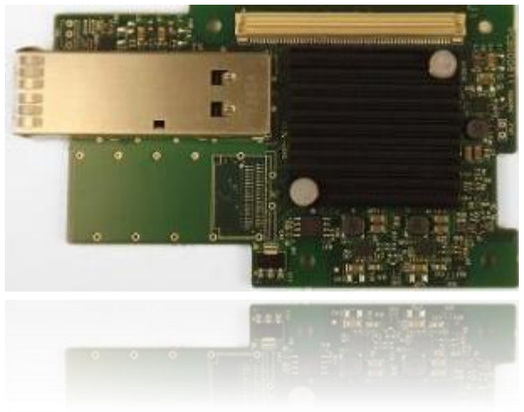


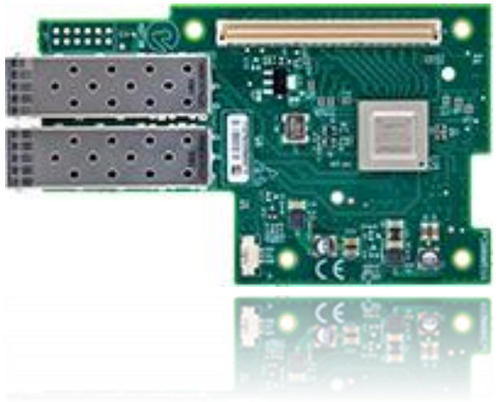
- A switch is a server with many ports: ASIC, Hardware, OS, Applications

OCP 어댑터

- Mellanox is first to market
  - First OCP-compliant 10GbE adapter
  - First OCP-compliant 40GbE adapter
- Shipping in high volume
  - Single and dual port options
  - Servers sold by multiple ODMs
- Support Microsoft OCP Cloud Server Specification







Post This Tweet This

Mon, Jan 27, 2014

Mellanox Releases World's First 40 Gigabit Ethernet NIC Based on Open Compute Project (OCP) Designs

ConnectX®-3 Pro 40GbE OCP-based NICs are built to OCP specifications and optimize the performance of scalable and virtualized environments by providing virtualization and overlay network offloads

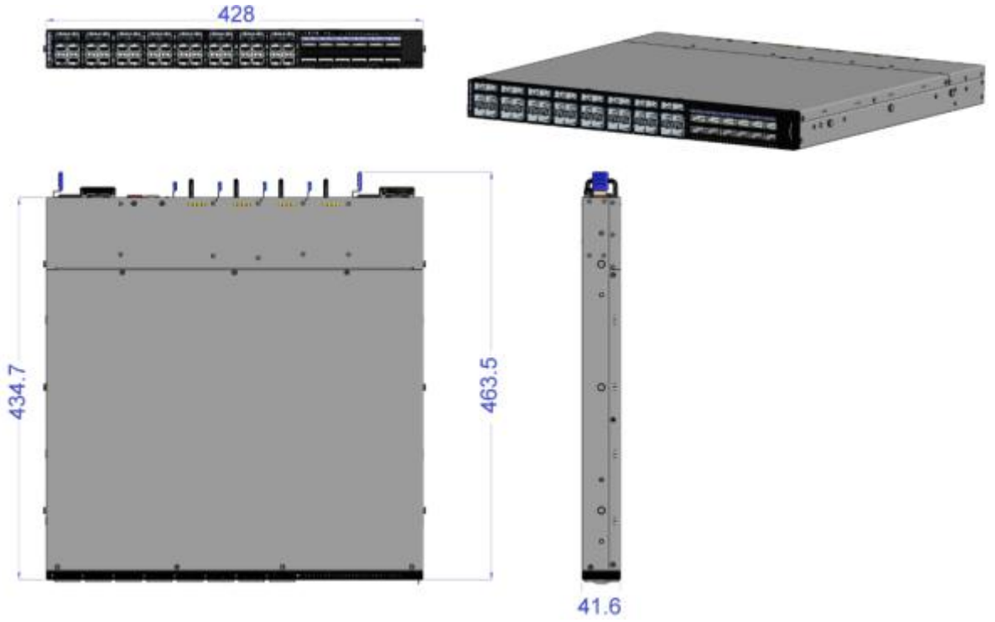
OCP Summit, San Jose, CA – January 27, 2014 – Mellanox® Technologies, Ltd. (NASDAQ: MLNX), a leading supplier of high-performance, end-to-end interconnect solutions for data center servers and storage systems,

OPEN Compute Project

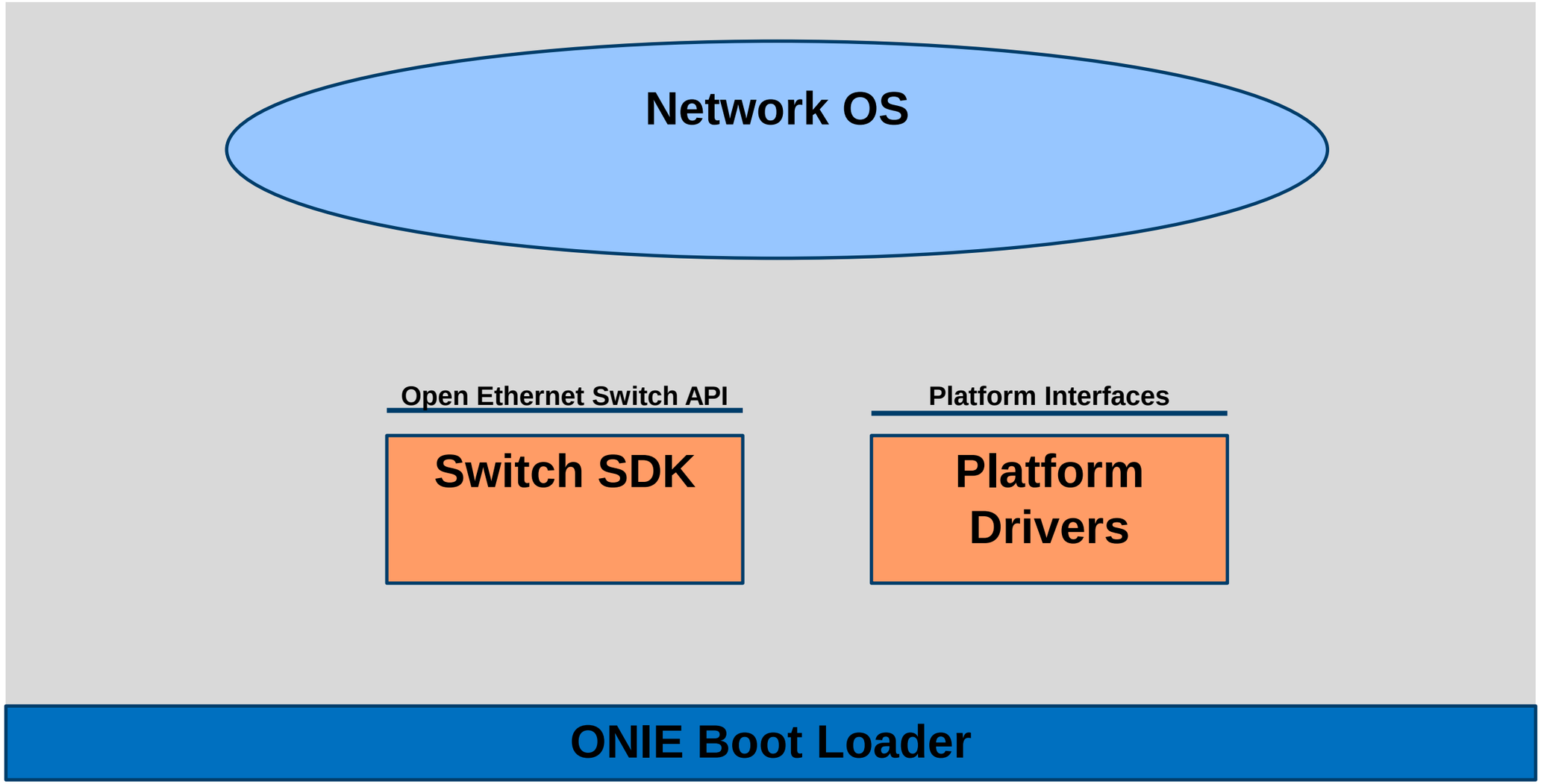
SX1410 ToR 스위치

- 1.92Tb/s throughput
- 200/300ns L2/L3 latency
- x86 dual core
- Low power
- Unified management interfaces and FRUs
- Non blocking 10GbE ToR
- ONIE memory compliant
- SwitchX-2 based
- Port Configurations
  - 48 SFP+ 10GbE – SR/LR/DAC
  - 12 QSFP+ 40GbE – SR4/LR4/DAC
  - Break-out capabilities:

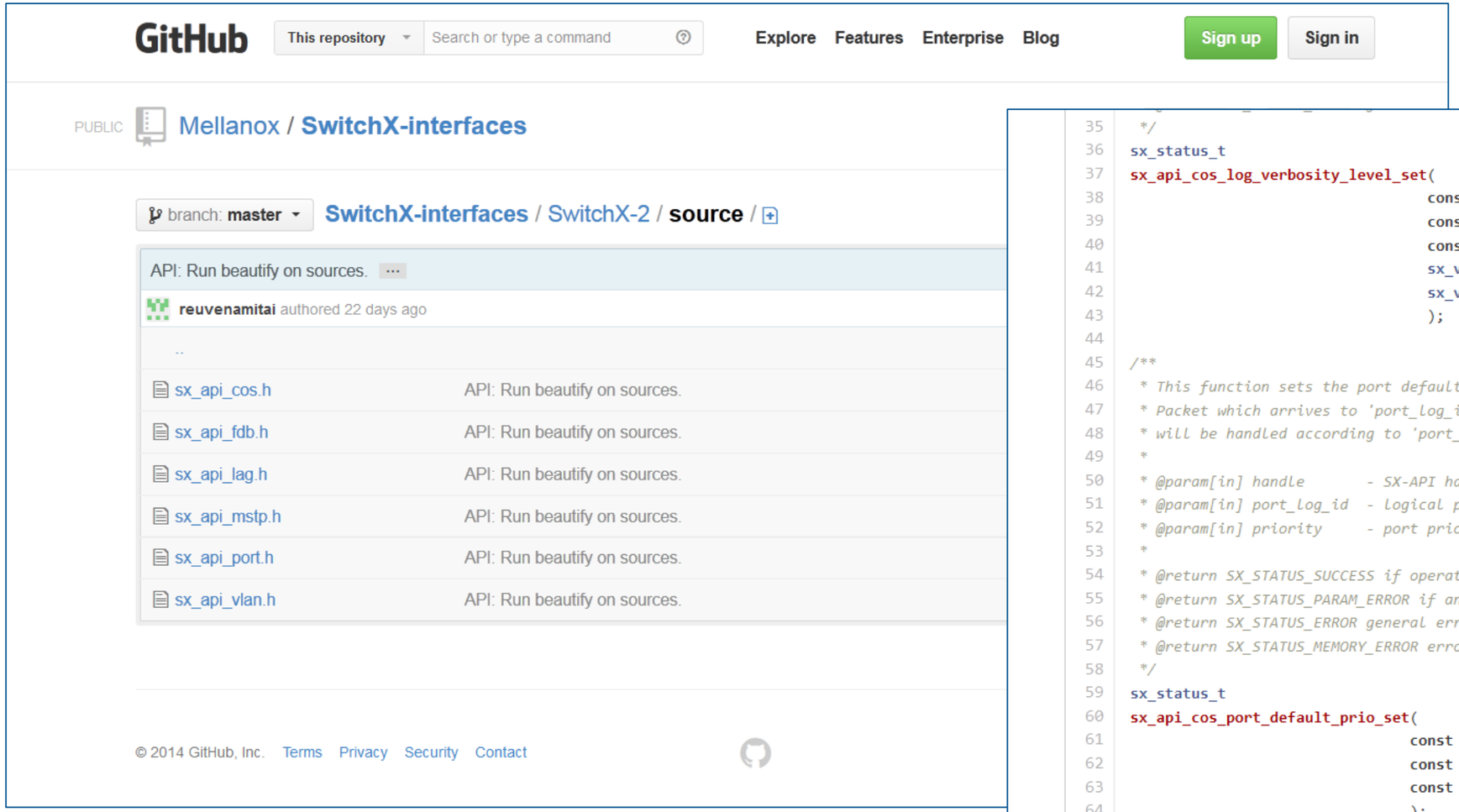
40GbE ports	12	10	8	6	4	2	0
10GbE ports	48	54	56	58	60	62	64

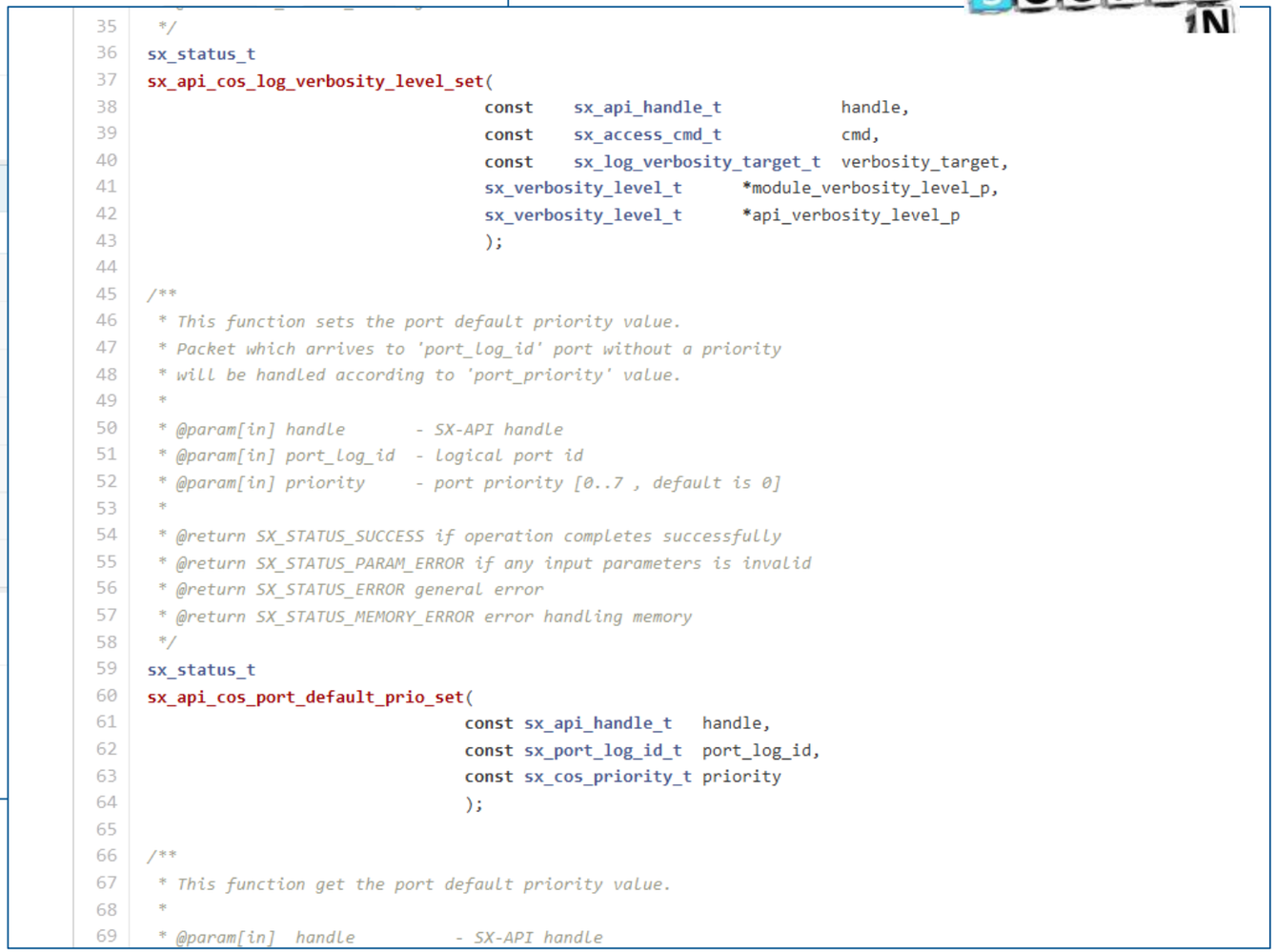


Open Platform System Software Blocks



오픈 스위치 SDK API 제공





```
 */
sx_status_t
sx_api_cos_log_verbosity_level_set(
                                   const    sx_api_handle_t           handle,
                                   const    sx_access_cmd_t           cmd,
                                   const    sx_log_verbosity_target_t  verbosity_target,
                                   sx_verbosity_level_t      *module_verbosity_level_p,
                                   sx_verbosity_level_t      *api_verbosity_level_p
                                   );

/**
 * This function sets the port default priority value.
 * Packet which arrives to 'port_log_id' port without a priority
 * will be handled according to 'port_priority' value.
 *
 * @param[in] handle       - SX-API handle
 * @param[in] port_log_id  - logical port id
 * @param[in] priority     - port priority [0..7 , default is 0]
 *
 * @return SX_STATUS_SUCCESS if operation completes successfully
 * @return SX_STATUS_PARAM_ERROR if any input parameters is invalid
 * @return SX_STATUS_ERROR general error
 * @return SX_STATUS_MEMORY_ERROR error handling memory
 */
sx_status_t
sx_api_cos_port_default_prio_set(
                                 const sx_api_handle_t   handle,
                                 const sx_port_log_id_t  port_log_id,
                                 const sx_cos_priority_t priority
                                 );

/**
 * This function get the port default priority value.
 *
 * @param[in]  handle          - SX-API handle
```

https://github.com/Mellanox/SwitchX-interfaces/tree/master/SwitchX-2

오픈 이더넷 하드웨어 제품 군

1U, 36 x QSFP Ethernet Switch
The Ideal 40GbE ToR/Aggregation

1U, 48 x SFP+ and 12 x QSFP Ethernet Switch
Non-blocking 10GbE➔ 40GbE ToR

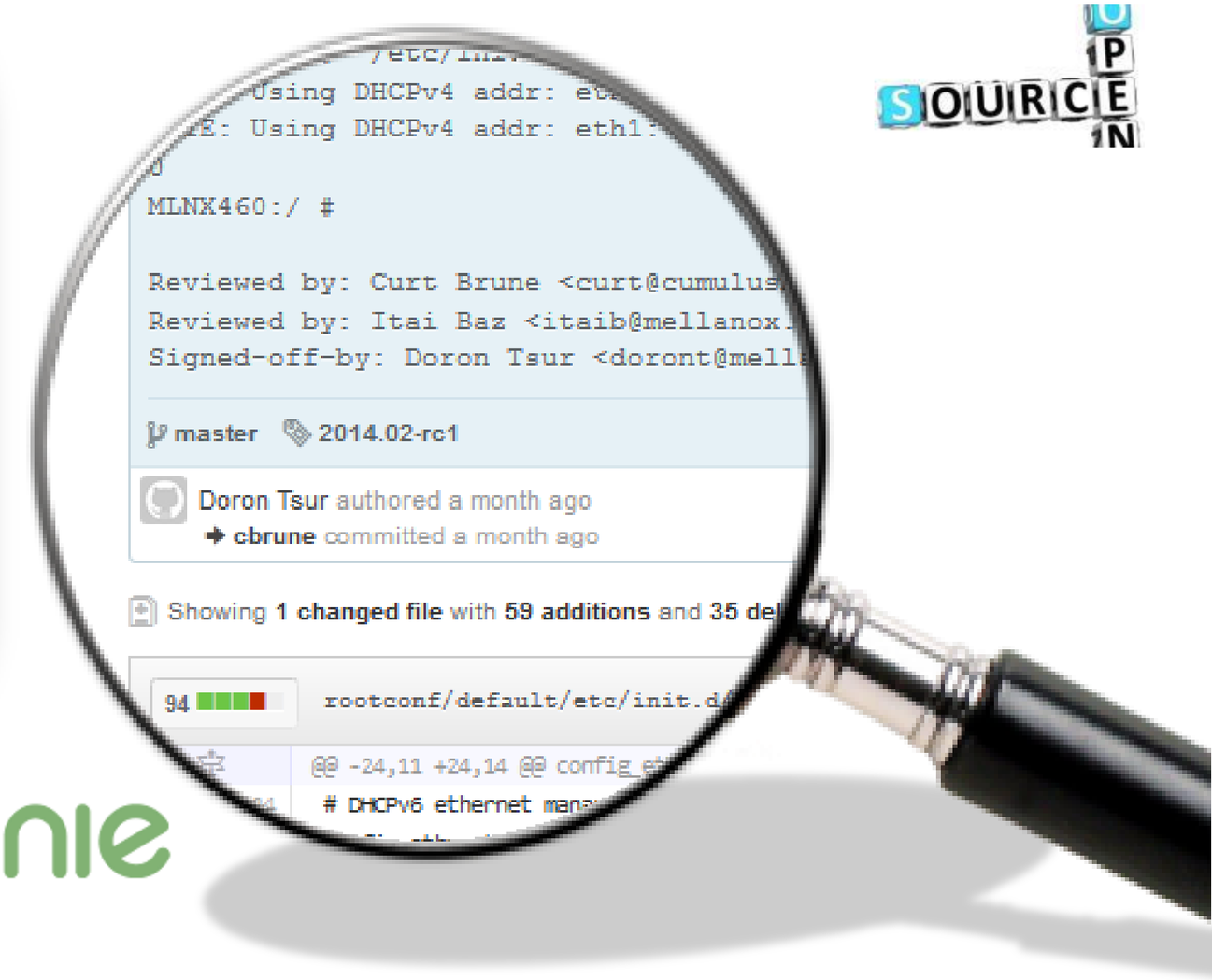
1U, 64 x SFP+ Ethernet Switch
Highest density 10GbE ToR

1U, Half Width, 12 x QSFP Ethernet Switch
Ideal storage/Database 10/40GbE Switch

- Highest Capacity in 1RU
  - From 12 QSFP to 36 QSFP
- Unique Value Proposition
  - 56GbE
  - VPI
  - End to end solution
  - Lowest power
- Low latency
  - 220ns L2 latency
  - 330ns L3 latency
- Low power
  - Under 1W per 10GbE interface
  - 2.3W per 40GbE interface
  - 0.6W per 10GbE of throughput

ONIE를 위한 통합과 기여

- Open Network Install Environment
  - Boot loader + Linux kernel + BusyBox
- Freedom of choice
  - Load any Net-OS on any hardware
- ONIE for SwitchX-2 available on
  - PPC
  - x86



Source: http://onie.github.io/onie

최적의 데이타 센타를 위한 Mellanox 인터컨넥트

RDMA

Lowest Latency

Highest Throughput

Highest Utilization

Unlimited Scalability

Best Offload Engines

Compute and Storage

Congestion Control

Overlay Networks

InfiniBand and Ethernet

Virtualization

Best Return on Investment

Most Cost-Effective Compute and Storage Interconnect Solution

Thank You

Mellanox
TECHNOLOGIES

Connect. Accelerate. Outperform.™