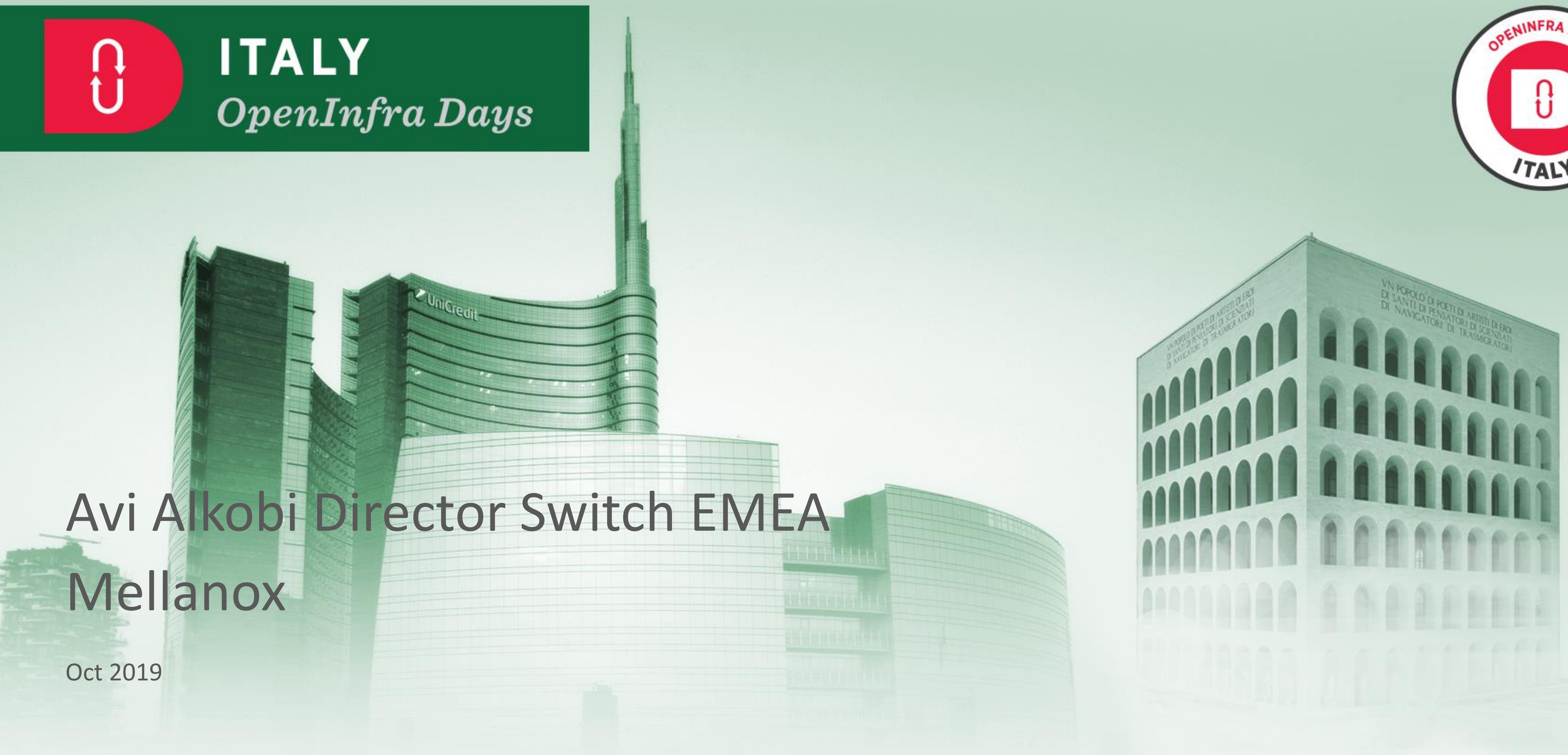
ITALY
OpenInfra Days



Avi Alkobi Director Switch EMEA Mellanox

Oct 2019

Organized by

IRIDEOS

Binario Etico

Under the patronage of

AGID | Agenzia per l'Italia Digitale

Sponsored by

Mellanox TECHNOLOGIES

MESOSPHERE

gci SERVICE FACTORY

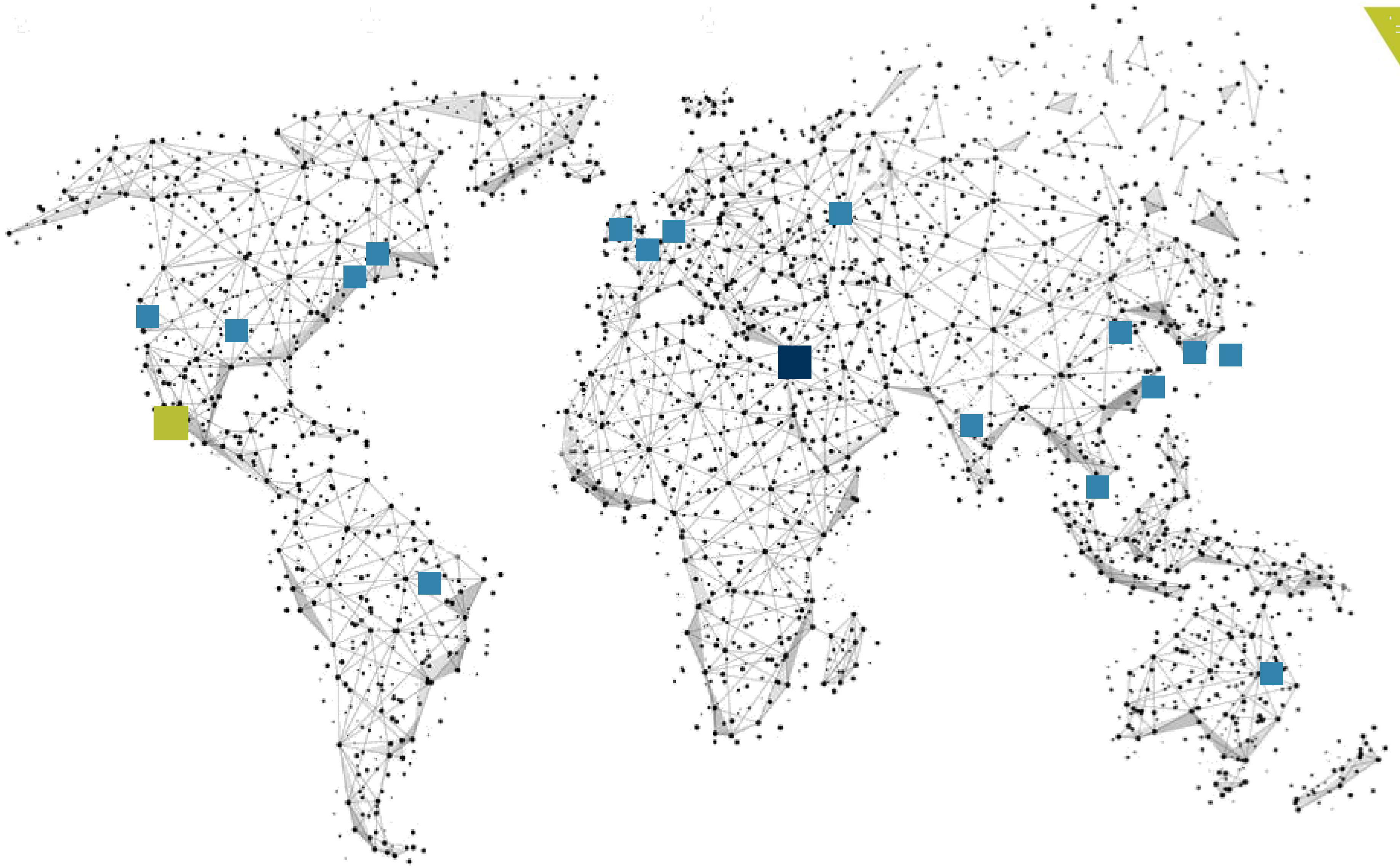
Mellanox Overview

1999
Mellanox Founded

$1.09B
2018 Revenue

~2,500
Employees worldwide

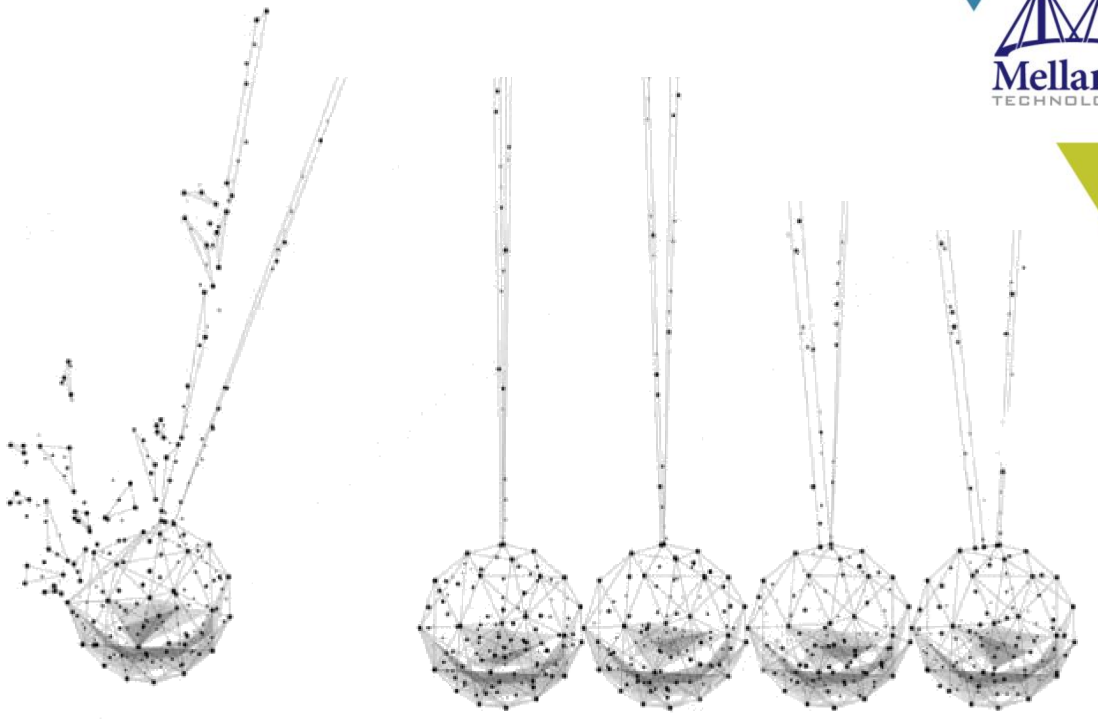
Ticker: MLNX

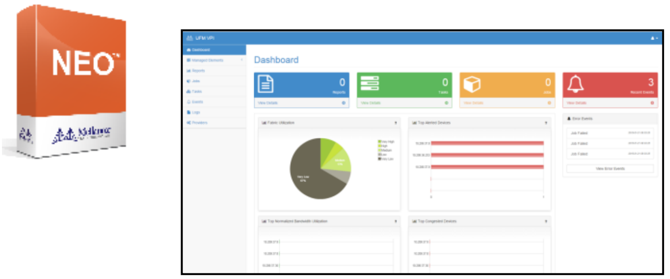


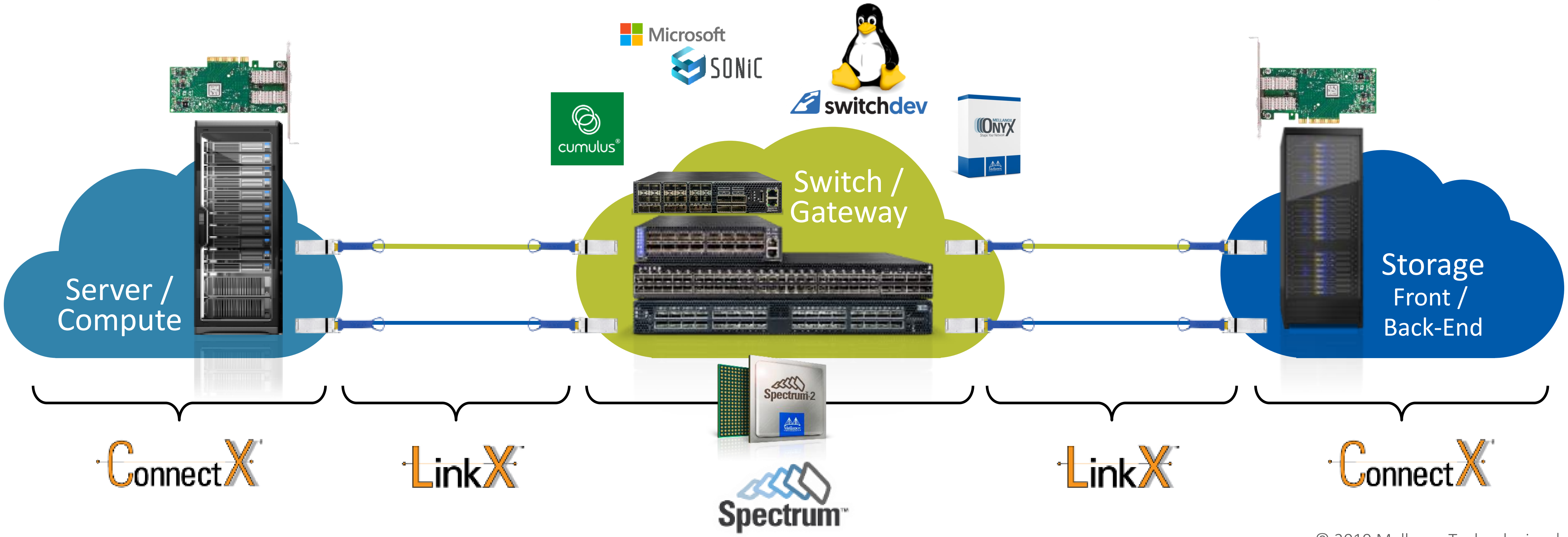
Company Headquarters:

- ■ Yokneam, Israel
- 🟩 Sunnyvale, California
- 🟦 Worldwide Offices

End-to-End Interconnect Solutions

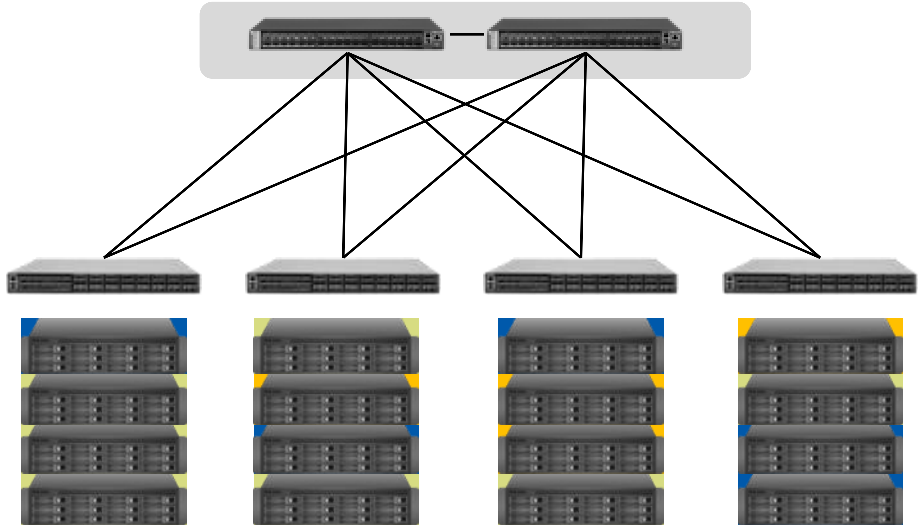






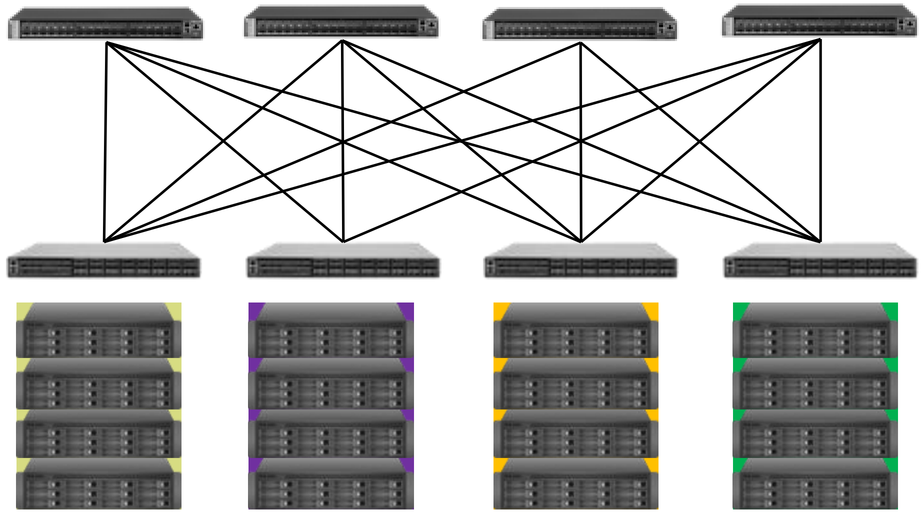
Leaf/Spine Deployments

Layer 2 / MLAG



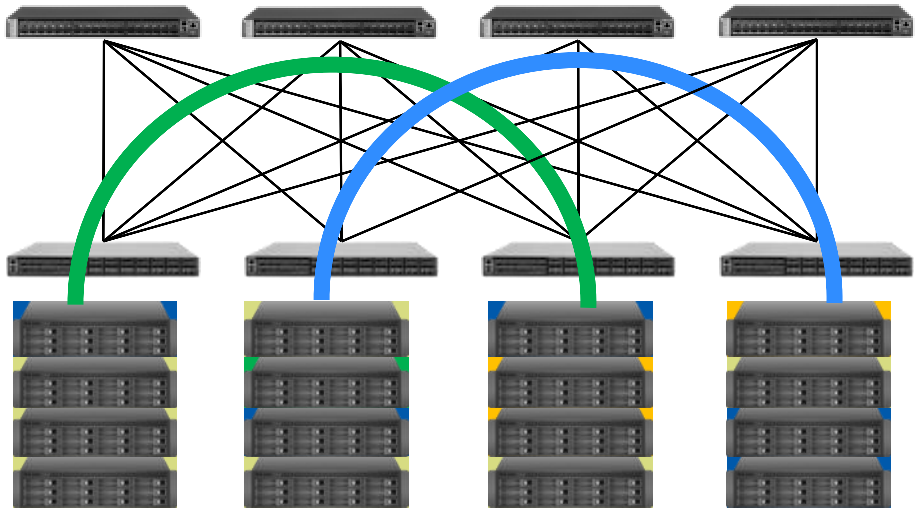
Connect out via spines, L3 GW on spines or above

Layer 3 / ECMP



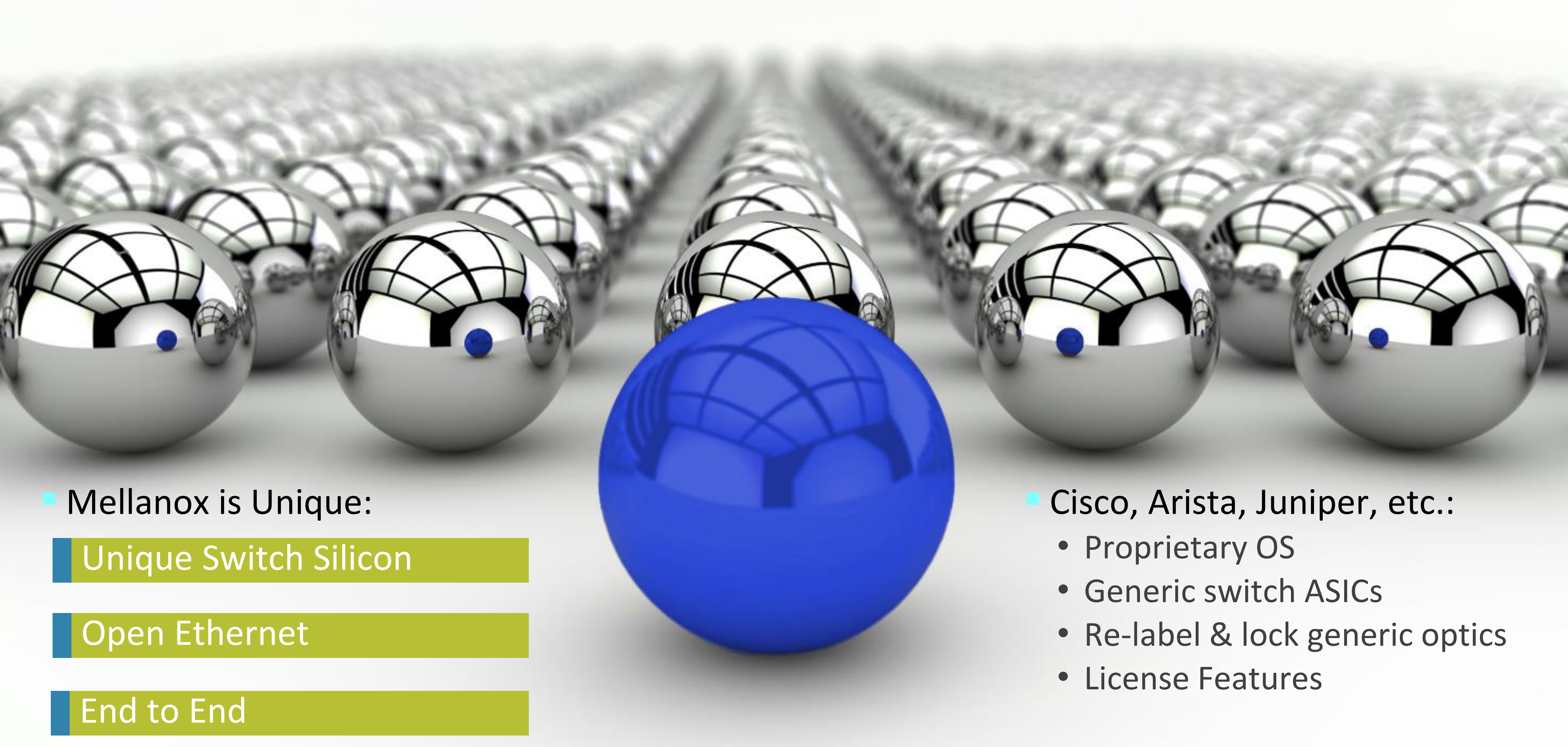
BGP from the Host – Kuberentes
VXLAN from the host- VMware , OpenStack, Kubernetes

L2 over Layer 3 VXLAN



Anycast L3 GW on TORs or FW as GW located on the border leaf.
EVPEN Type 5 for Routes out of the fabric.

Mellanox – Not Like Other Network Vendors



- Mellanox is Unique:

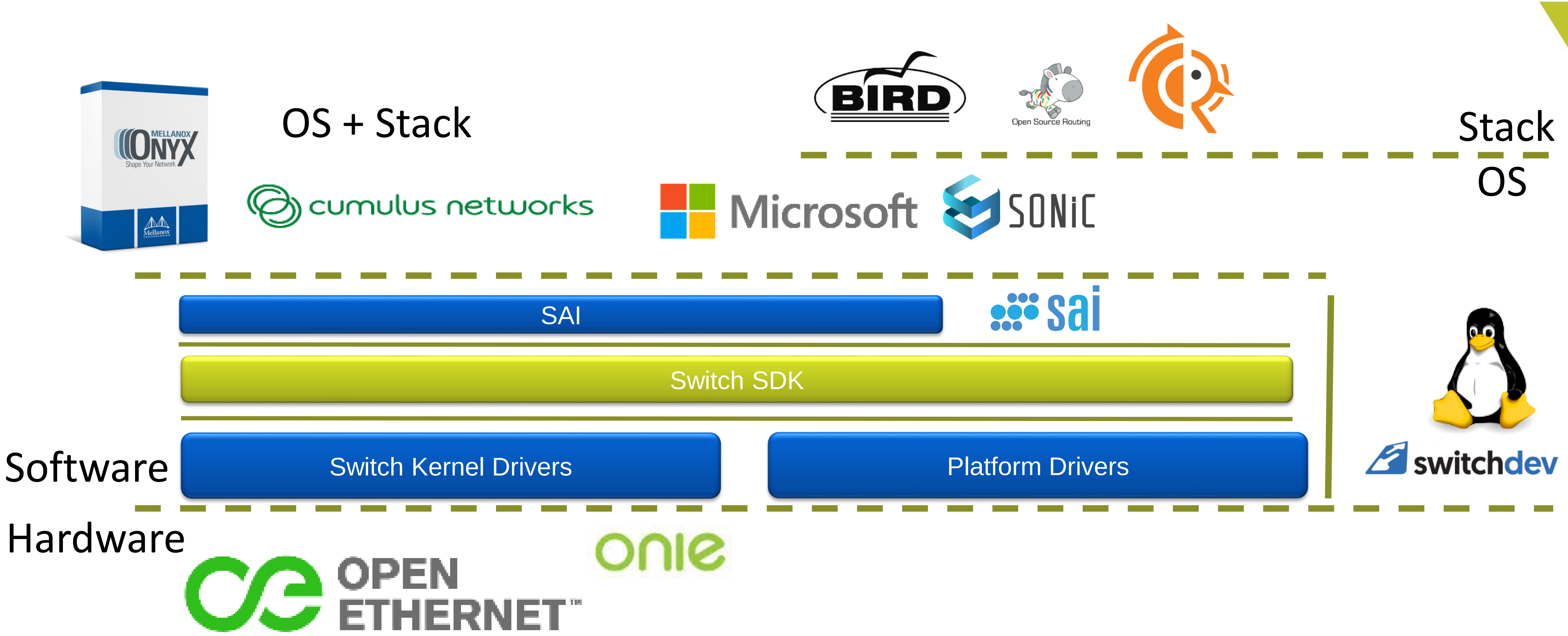
Unique Switch Silicon

Open Ethernet

End to End

- Cisco, Arista, Juniper, etc.:
  - Proprietary OS
  - Generic switch ASICs
  - Re-label & lock generic optics
  - License Features

We Are Building an Open Ecosystem



OS + Stack

Stack

OS

SAI

Switch SDK

Software

Switch Kernel Drivers

Platform Drivers

Hardware

Open Ethernet SN2000 Series

SN2700 – 32x100GbE (up to 64 x 50/25/10GbE)
The Ideal 100GbE ToR / Aggregation

SN2410 – 8x100GbE + 48x25GbE
25GbE → 100GbE ToR

SN2100 – 16x100GbE ports (64x25GbE)
Ideal storage/Database Switch
Highest 25GbE Density per rack unit

SN2010 – 18x10/25GbE + 4x40/100GbE
Ideal HCI ToR Switch

- Predictable Performance
- Fair Traffic Distribution for Cloud
- Best-in-Class Throughput, Latency, Power Consumption
- Zero Packet Loss

300ns

SN2700 – 169W
SN2410 – 165W
SN2100 – 94W

Energy efficiency

Spectrum™

Spectrum 2 - Open Ethernet SN3000 Series

SN3700C – 32x100GbE (128x 1-25GbE)
100GbE Spine/ToR

SN3700 – 32x200GbE (128x 1-50GbE)
200GbE Spine

SN3800 – 64x100GbE
Spine/Super Spine

WHAT JUST HAPPENED

SN3510 – 48x25/50GbE + 6x400GbE
25/50GbE → 400GbE ToR
(Q2 2020)

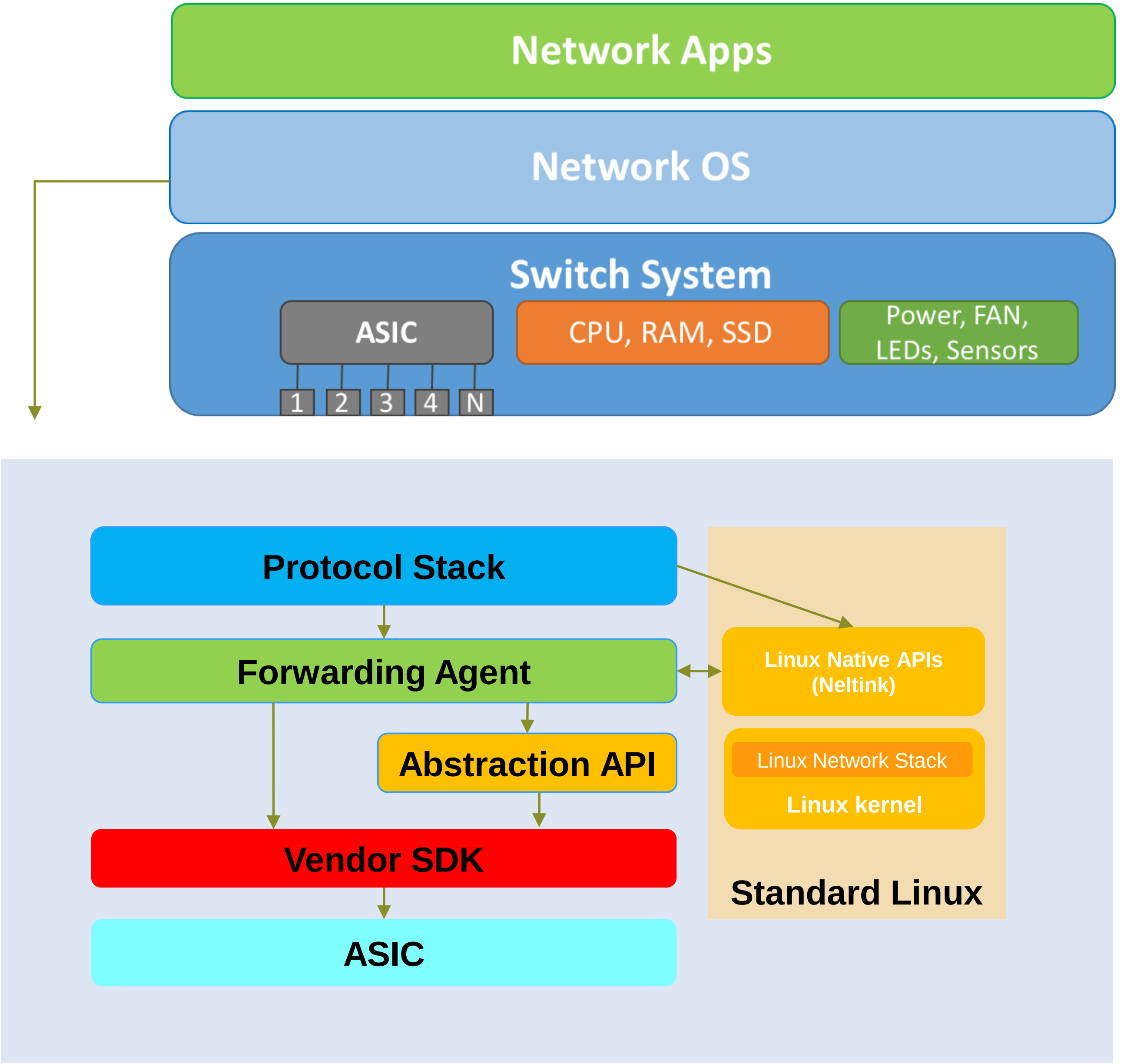
- Best-in-Class Buffers
- Best-In-Class Virtualization
- Best-In-Class Telemetry

cumulus®

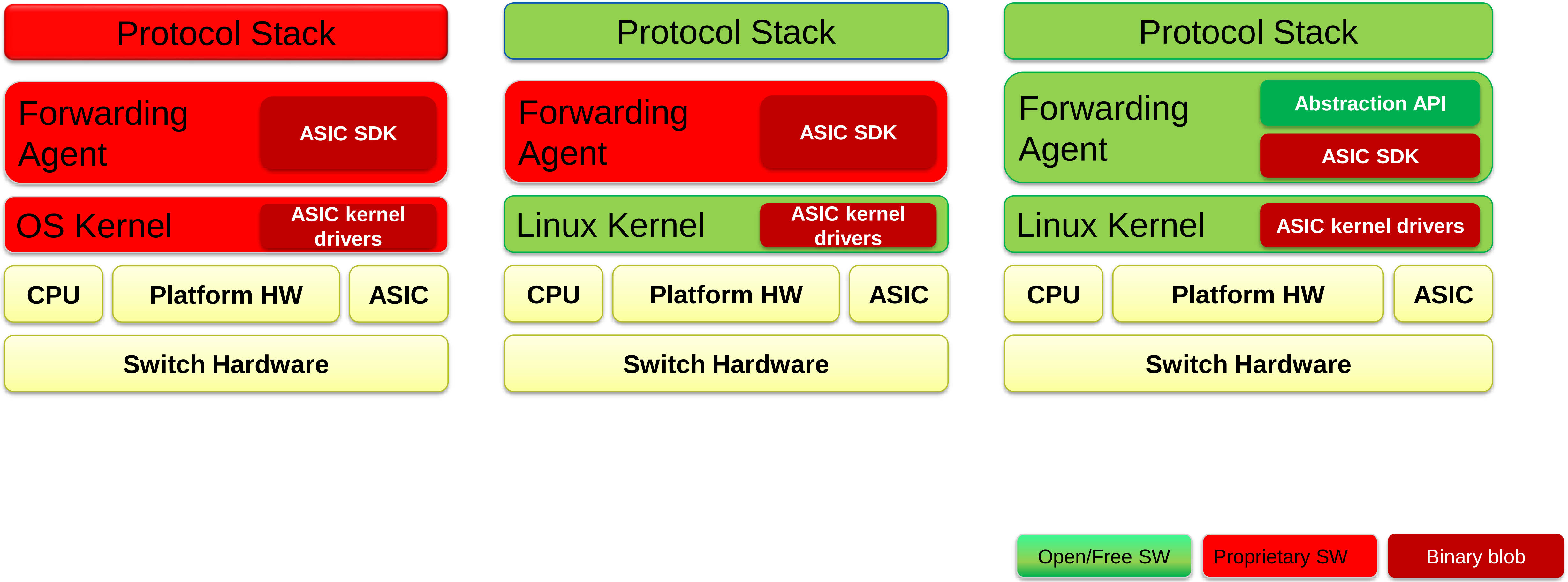
MELLANOX ONYX

Switch NOS Reference Architecture

- Protocol Stack
  - Network protocols (RIB)
    - Bridge, STP, OSPF, BGP.
  - Forming FIB our of RIB
- Forwarding Agent
  - Middleware between Protocol Stack & ASIC
  - Programming FIB into the ASIC (HW offload)
  - Uses special API to communicate with ASIC
- API
  - Proprietary ASIC vendor SDK
  - Broadcom, Mellanox, Cavium, Marvell, etc.
  - Trend to standardize and open SDK
  - SAI, OpenNSL, OF-DPA, P4, etc.

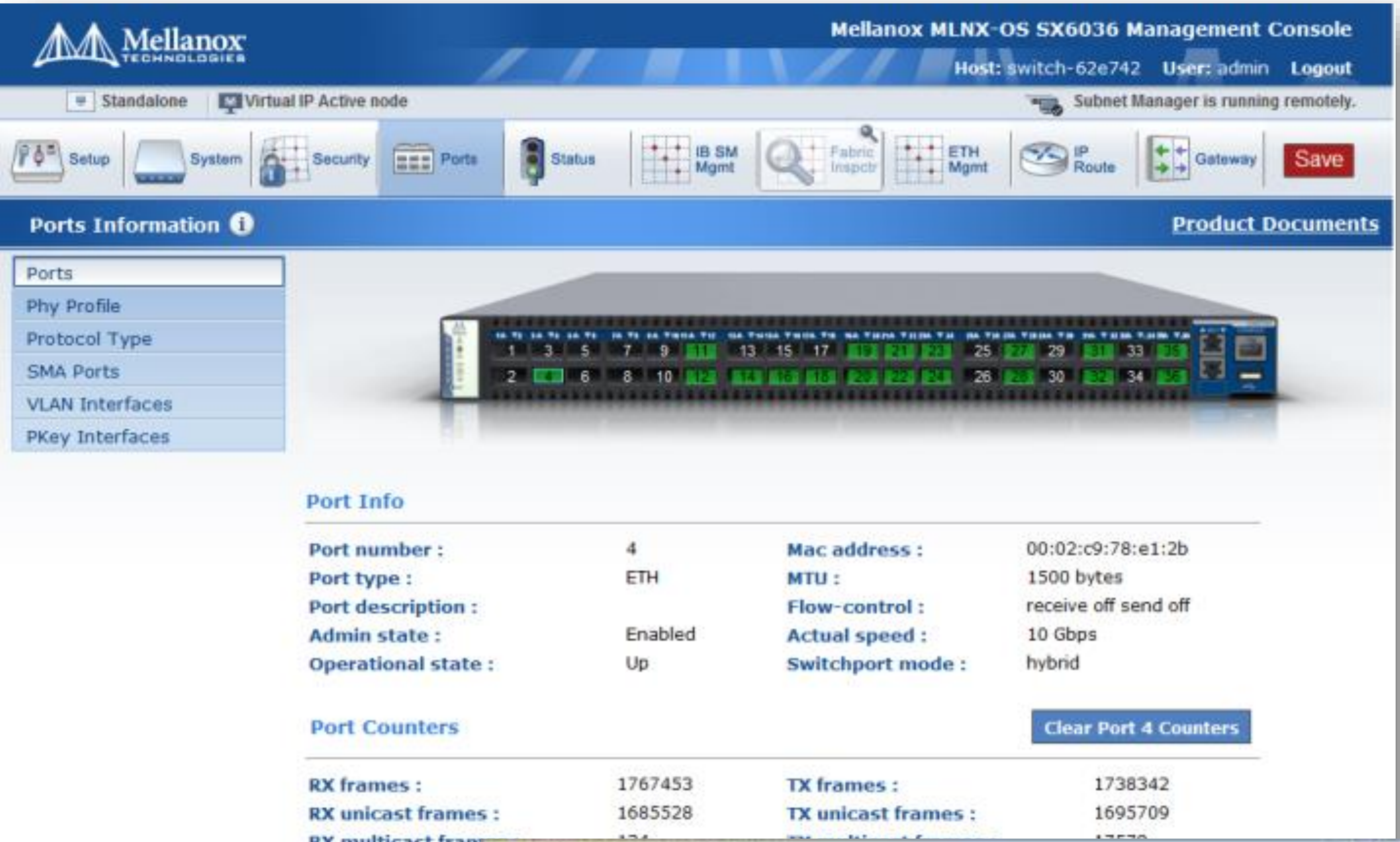


Switch NOS architecture examples



Traditional NOS	Commercial NOS based on Linux	Open NOS based on Linux
Protocol Stack	Protocol Stack	Protocol Stack
Forwarding Agent ASIC SDK	Forwarding Agent ASIC SDK	Forwarding Agent Abstraction API ASIC SDK
OS Kernel ASIC kernel drivers	Linux Kernel ASIC kernel drivers	Linux Kernel ASIC kernel drivers
CPU | Platform HW | ASIC	CPU | Platform HW | ASIC	CPU | Platform HW | ASIC
Switch Hardware	Switch Hardware	Switch Hardware

MLNX-OS traditional Industry like OS



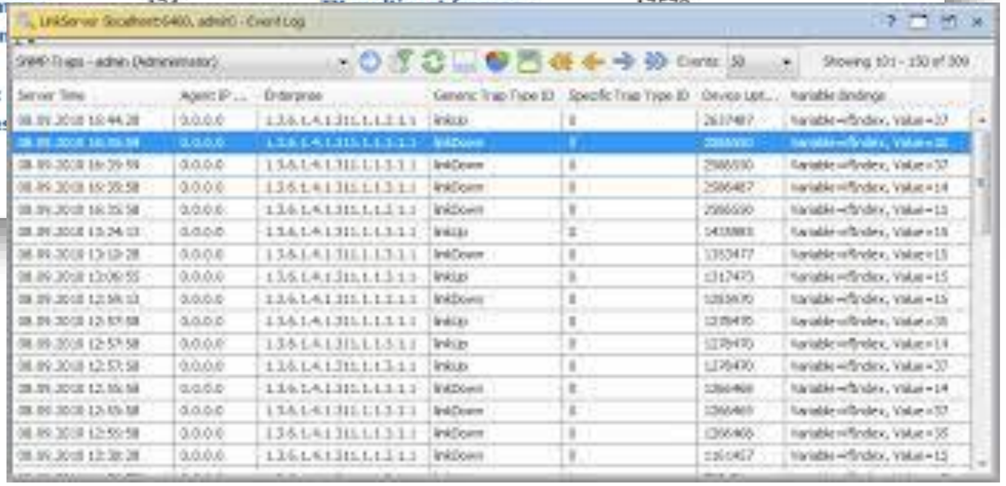
Port Info

Port number :	4	Mac address :	00:02:c9:78:e1:2b
Port type :	ETH	MTU :	1500 bytes
Port description :		Flow-control :	receive off send off
Admin state :	Enabled	Actual speed :	10 Gbps
Operational state :	Up	Switchport mode :	hybrid

Port Counters

RX frames :	1767453	TX frames :	1738342
RX unicast frames :	1685528	TX unicast frames :	1695709
RX multicast fram			
RX broadcast fram			
RX octets :			
RX error frames :			
RX discard frame			

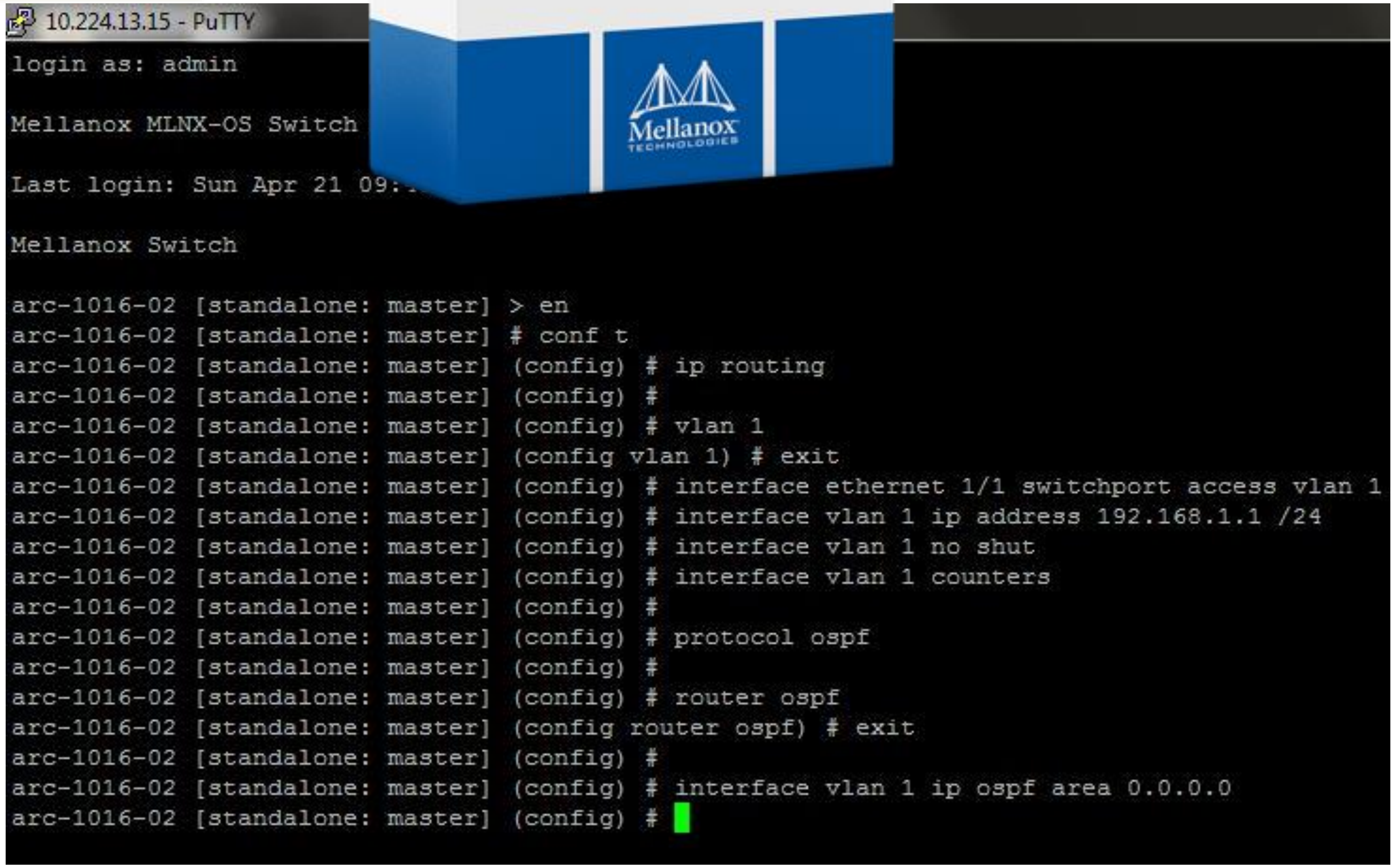
- GUI


LinkServer (localhost:6460, admin0) - Event Log
SNMP Traps - admin (Administrator)
Events: 50
Showing 101 - 150 of 309

Server Time	Agent IP ...	Enterprise	Generic Trap Type ID	Specific Trap Type ID	Device Upt...	Variable Bindings
08.09.2018 16:44:28	0.0.0.0	1.3.6.1.4.1.311.1.1.3.1.1	linkUp	0	2617487	Variable=ifIndex, Value=37
08.09.2018 16:[illegible]	0.0.0.0	1.3.6.1.4.1.311.1.1.3.1.1	linkDown	0	[illegible]	Variable=ifIndex, Value=[illegible]
08.09.2018 16:39:59	0.0.0.0	1.3.6.1.4.1.311.1.1.3.1.1	linkDown	0	2586590	Variable=ifIndex, Value=37
08.09.2018 16:38:58	0.0.0.0	1.3.6.1.4.1.311.1.1.3.1.1	linkDown	0	2586487	Variable=ifIndex, Value=14
08.09.2018 16:35:58	0.0.0.0	1.3.6.1.4.1.311.1.1.3.1.1	linkDown	0	2586590	Variable=ifIndex, Value=13
08.09.2018 13:24:13	0.0.0.0	1.3.6.1.4.1.311.1.1.3.1.1	linkUp	0	1418883	Variable=ifIndex, Value=18
08.09.2018 13:12:28	0.0.0.0	1.3.6.1.4.1.311.1.1.3.1.1	linkDown	0	1353477	Variable=ifIndex, Value=13
08.09.2018 13:06:55	0.0.0.0	1.3.6.1.4.1.311.1.1.3.1.1	linkUp	0	1317475	Variable=ifIndex, Value=15
08.09.2018 12:58:13	0.0.0.0	1.3.6.1.4.1.311.1.1.3.1.1	linkDown	0	1283670	Variable=ifIndex, Value=15
08.09.2018 12:57:58	0.0.0.0	1.3.6.1.4.1.311.1.1.3.1.1	linkUp	0	1278470	Variable=ifIndex, Value=38
08.09.2018 12:57:58	0.0.0.0	1.3.6.1.4.1.311.1.1.3.1.1	linkUp	0	1278470	Variable=ifIndex, Value=14
08.09.2018 12:57:58	0.0.0.0	1.3.6.1.4.1.311.1.1.3.1.1	linkUp	0	1278470	Variable=ifIndex, Value=37
08.09.2018 12:56:58	0.0.0.0	1.3.6.1.4.1.311.1.1.3.1.1	linkDown	0	1266468	Variable=ifIndex, Value=14
08.09.2018 12:55:58	0.0.0.0	1.3.6.1.4.1.311.1.1.3.1.1	linkDown	0	1266469	Variable=ifIndex, Value=37
08.09.2018 12:55:58	0.0.0.0	1.3.6.1.4.1.311.1.1.3.1.1	linkDown	0	1266468	Variable=ifIndex, Value=35
08.09.2018 12:38:28	0.0.0.0	1.3.6.1.4.1.311.1.1.3.1.1	linkDown	0	1261457	Variable=ifIndex, Value=13

- SNMP

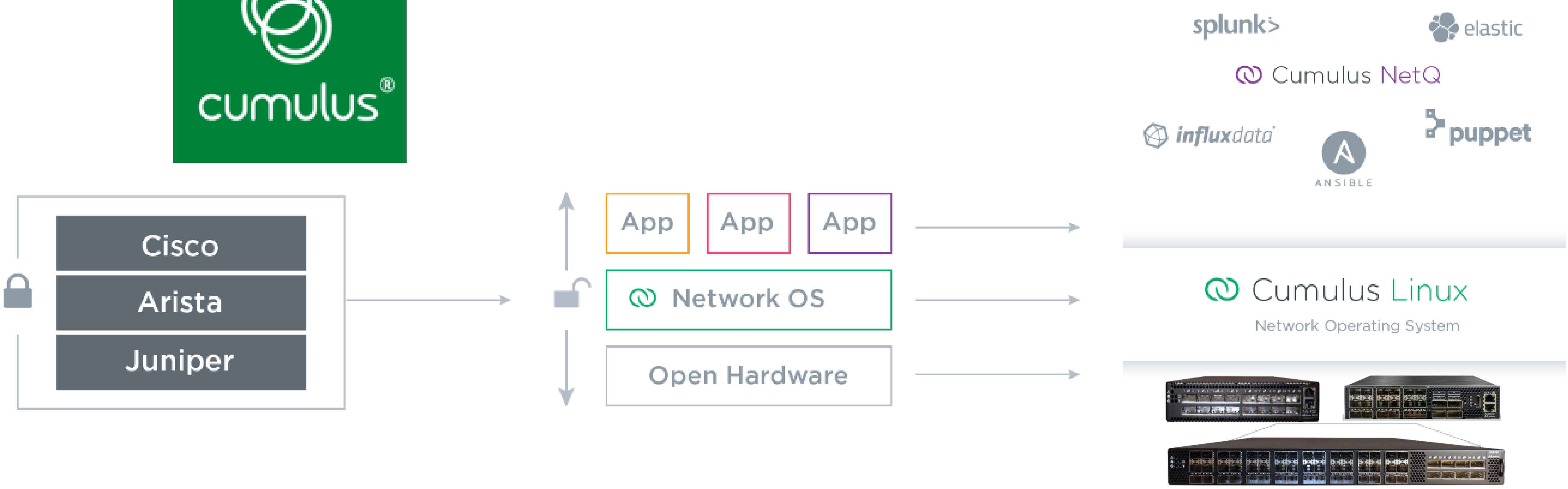
- CLI



```
login as: admin

Mellanox MLNX-OS Switch

Last login: Sun Apr 21 09:

Mellanox Switch

arc-1016-02 [standalone: master] > en
arc-1016-02 [standalone: master] # conf t
arc-1016-02 [standalone: master] (config) # ip routing
arc-1016-02 [standalone: master] (config) #
arc-1016-02 [standalone: master] (config) # vlan 1
arc-1016-02 [standalone: master] (config vlan 1) # exit
arc-1016-02 [standalone: master] (config) # interface ethernet 1/1 switchport access vlan 1
arc-1016-02 [standalone: master] (config) # interface vlan 1 ip address 192.168.1.1 /24
arc-1016-02 [standalone: master] (config) # interface vlan 1 no shut
arc-1016-02 [standalone: master] (config) # interface vlan 1 counters
arc-1016-02 [standalone: master] (config) #
arc-1016-02 [standalone: master] (config) # protocol ospf
arc-1016-02 [standalone: master] (config) #
arc-1016-02 [standalone: master] (config) # router ospf
arc-1016-02 [standalone: master] (config router ospf) # exit
arc-1016-02 [standalone: master] (config) #
arc-1016-02 [standalone: master] (config) # interface vlan 1 ip ospf area 0.0.0.0
arc-1016-02 [standalone: master] (config) #
```

Cumulus – Mellanox Partnership

cumulus®



splunk>

elastic

Cumulus NetQ

influxdata

ANSIBLE

puppet

Cisco

Arista

Juniper

App

App

App

Network OS

Open Hardware

Cumulus Linux

Network Operating System

Best Hardware

1 Economical scalability

With commodity hardware and a standardized Linux stack, achieving a lower TCO by up to 60%

2 Built for the automation age

Making networking repeatable and consistent

3 Standardized toolsets

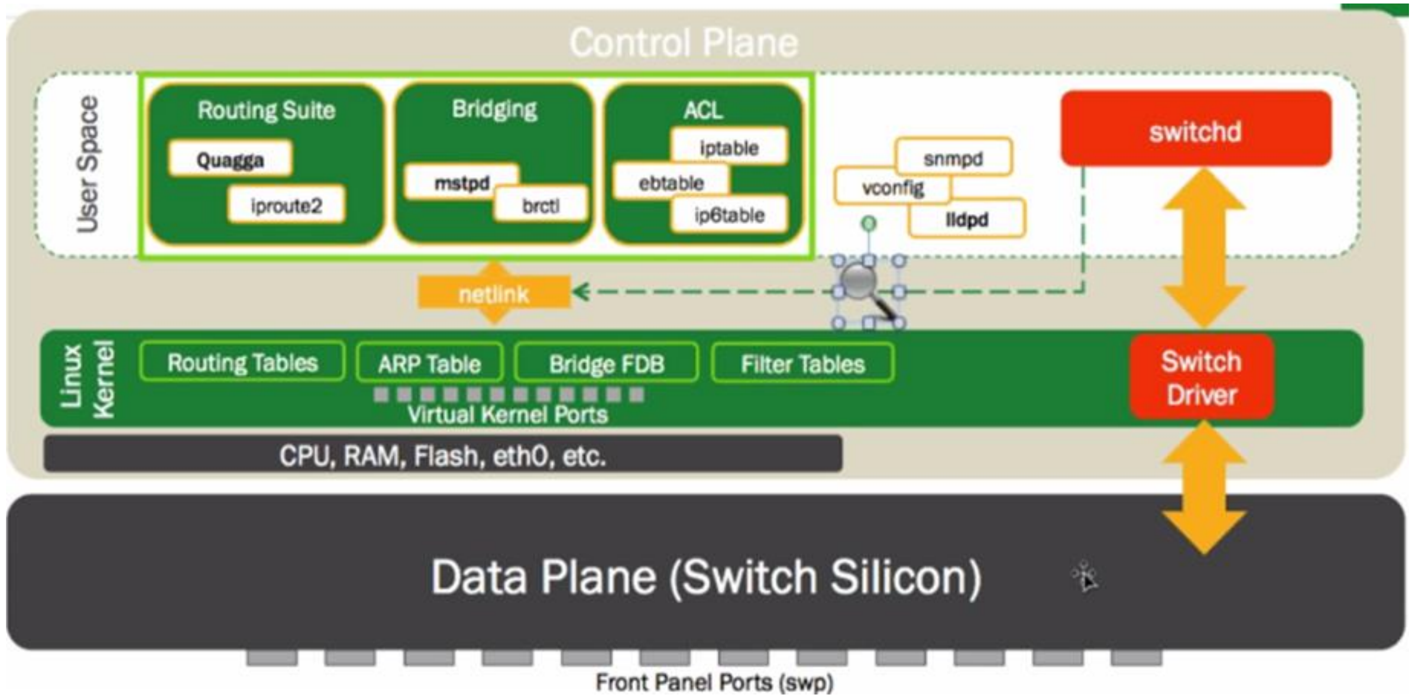
Easily enable Linux tools: automation, monitoring, analytics…

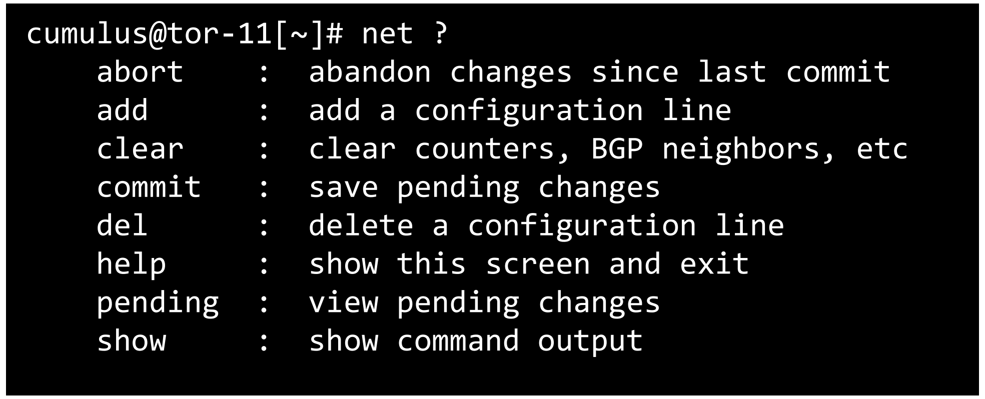
4 Choice and flexibility

50+ hardware platforms, from 11 vendors, and 2 silicon

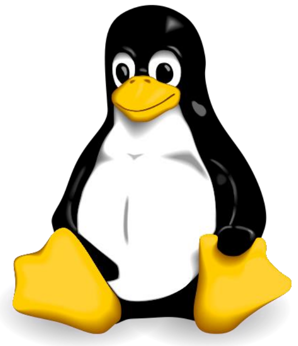
Cumulus

- Cumulus Linux is a Debian Jessie based , much lighter in the size
- Linux is a Linux, you can google it, you can use man command to learn about the different commands
- Except of the code that controls the silicon , all are open source and additions that are pushed to the upstream (in process)
- Cumulus are providing systems wrappers commands for a better user interface
- NCLU





```
cumulus@tor-11[~]# net ?
    abort     :  abandon changes since last commit
    add       :  add a configuration line
    clear     :  clear counters, BGP neighbors, etc
    commit    :  save pending changes
    del       :  delete a configuration line
    help      :  show this screen and exit
    pending   :  view pending changes
    show      :  show command output
```



SONiC - Software for Open Networking in the Cloud

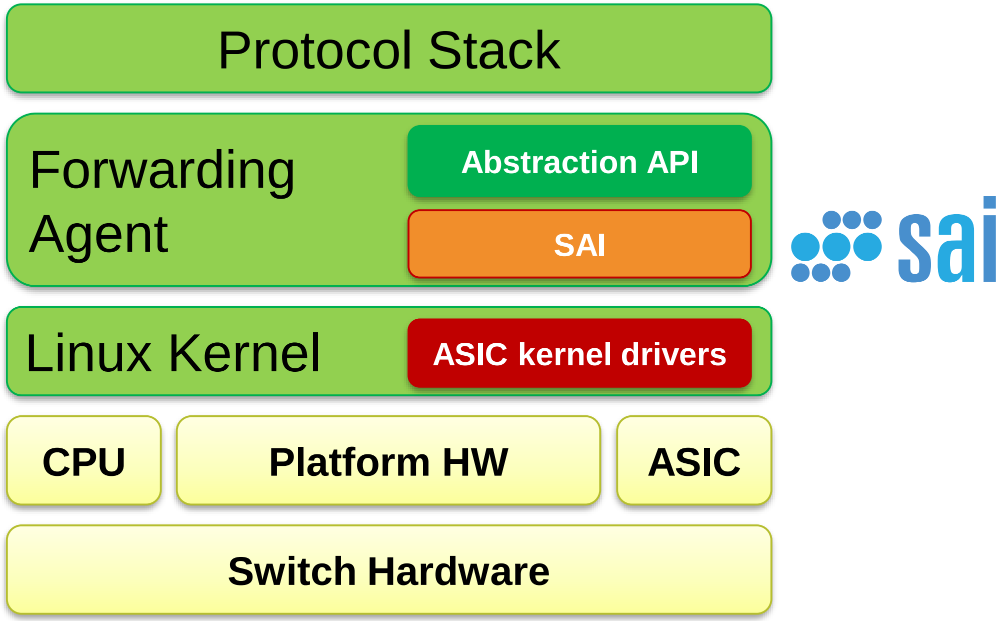
- SONiC is a collection of software packages installed on Linux running on a network hardware switch which make it a complete, functional router targeted at data center networks. Runs on Debian 8 ‘Jessie‘ distribution.

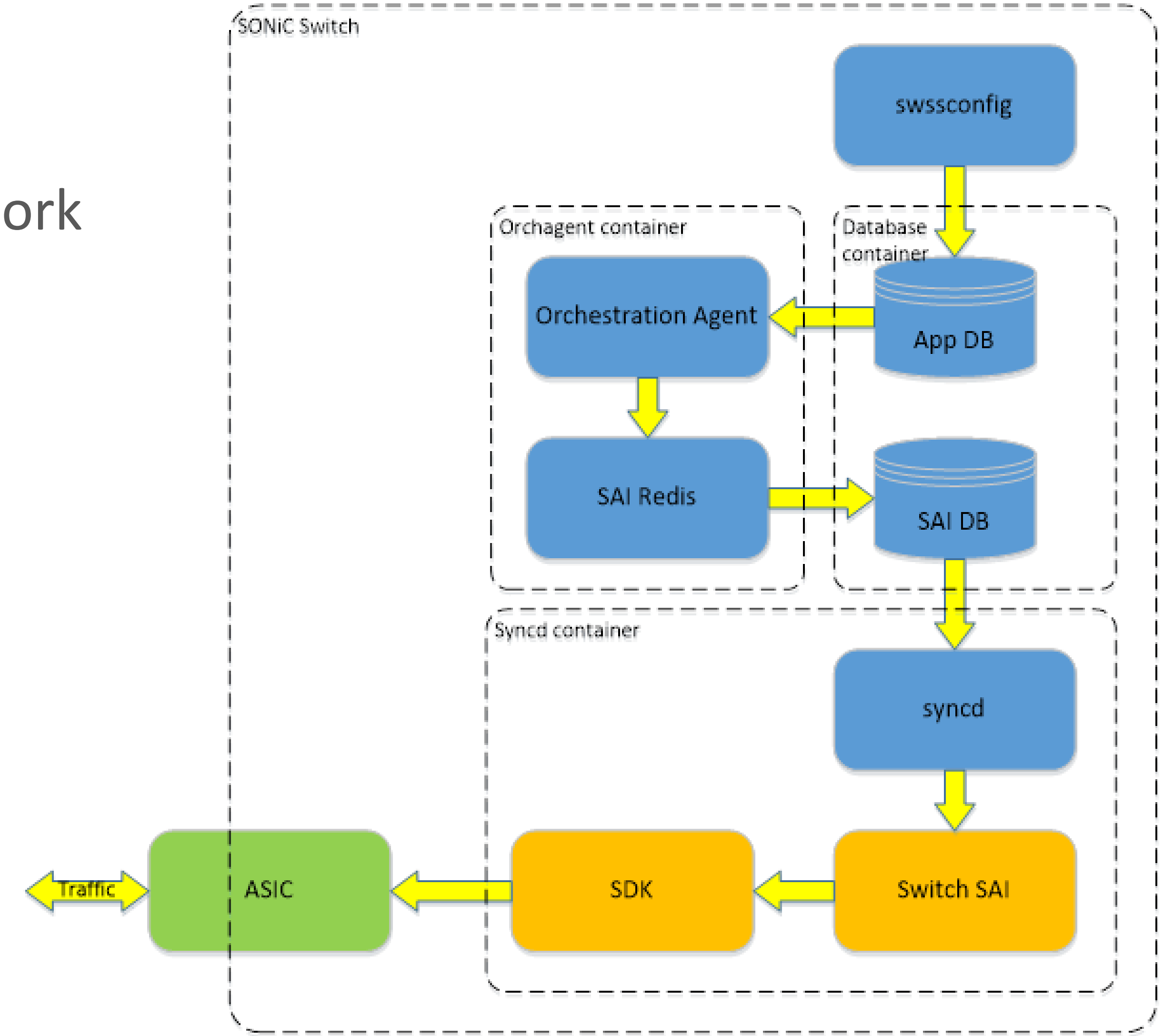
- SONiC is supported by the community and all code is shared in public github
  https://github.com/Azure/SONiC
  https://github.com/Azure/SONiC/wiki/Architecture

- SONiC deployment
  - in Microsoft production datacenters today and in Mellanox IT
  - Alibaba is planning soon as well
  - EMEA

SONiC Architecture cont.

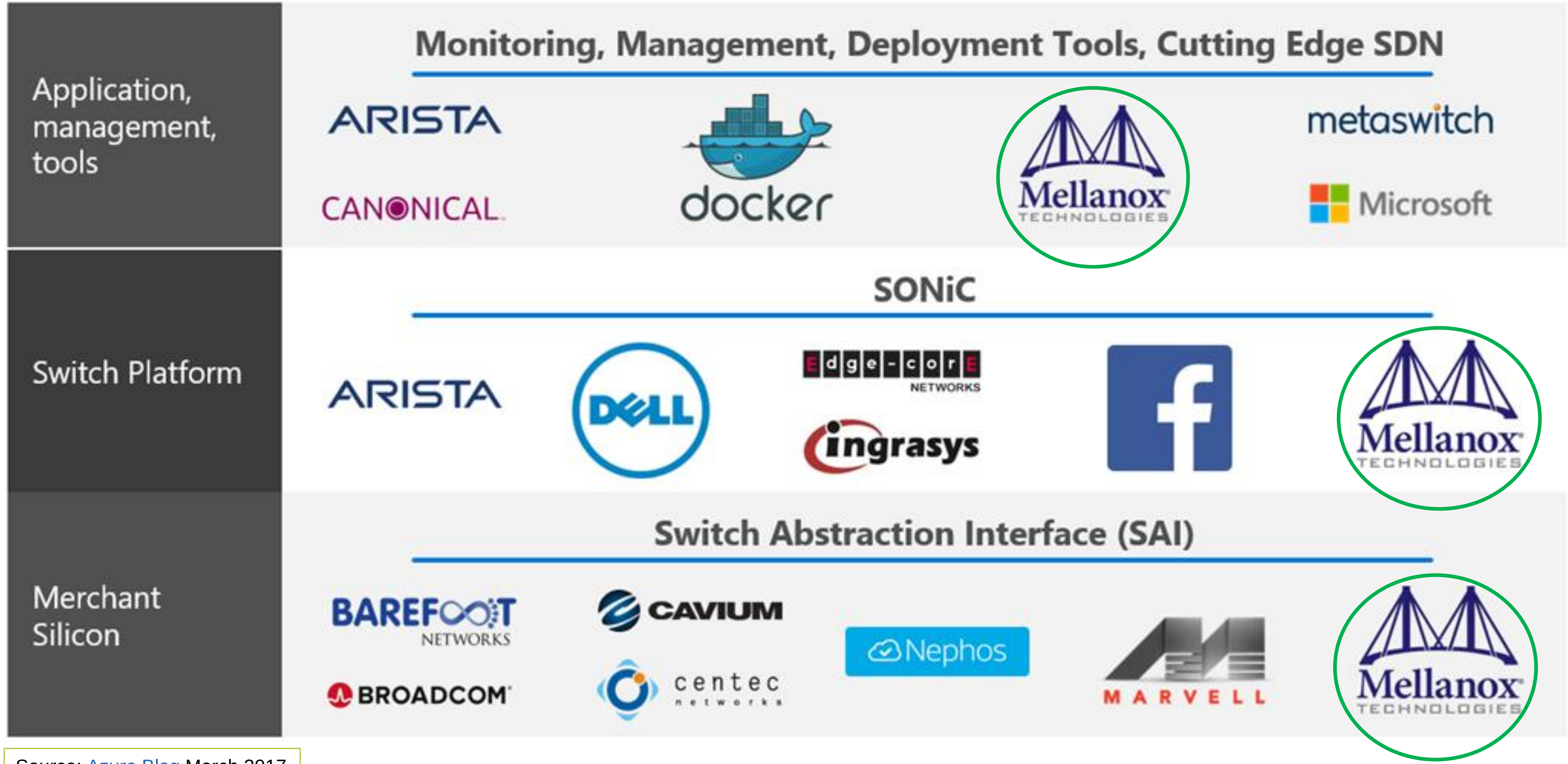
- The Switch State Service (SwSS) is a collection of software that provides a database interface for communication with and state representation of network applications and network switch hardware.





SONiC participants (From Azure blog)

Mellanox is the ONLY vendor to contribute at all levels



Application, management, tools

Monitoring, Management, Deployment Tools, Cutting Edge SDN

Switch Platform

SONiC

Merchant Silicon

Switch Abstraction Interface (SAI)

Source: Azure Blog March 2017

Mellanox Contribution

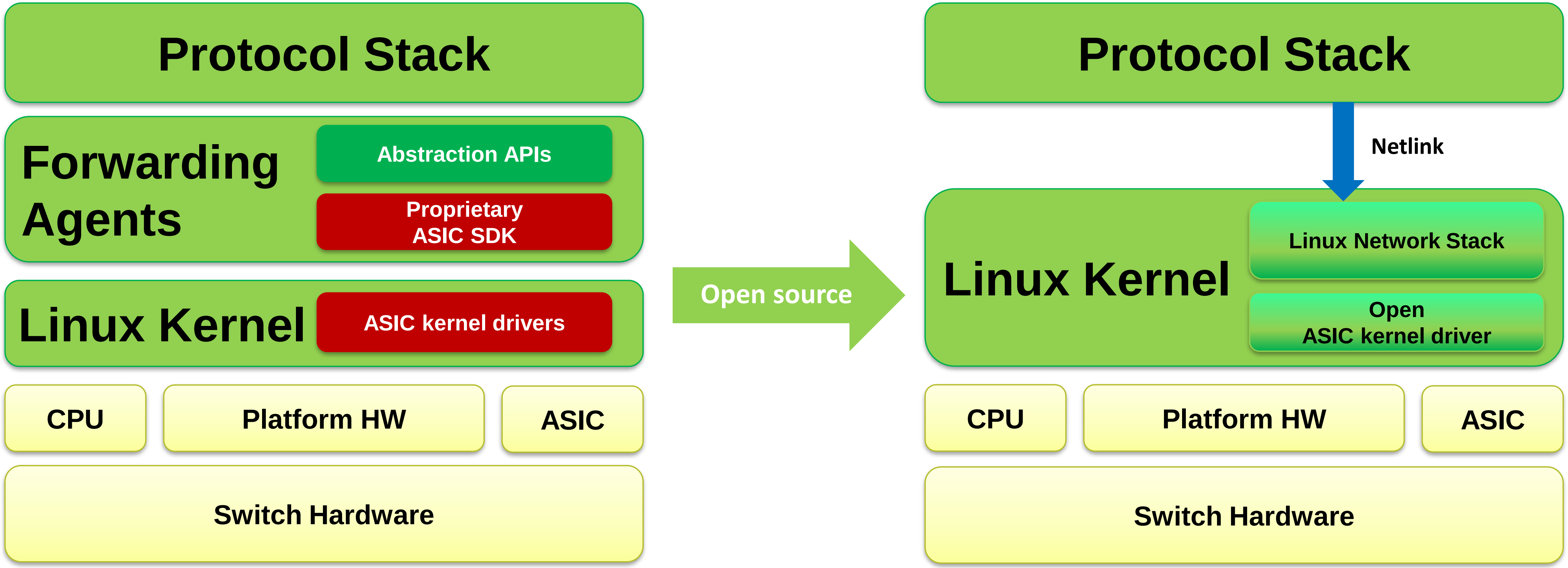
- MAC Aging
- BGP-EVPN support (type 5)
- New Platform APIs
  *For Mellanox platforms
- Virtual path for streaming telemetry
- Performance on Redis DB
- Management
  - AAA
  - TACAS
- L3 RIF counter support
- Transceiver parameter tuning
- Port Split
- Upgrade docker engine to 18.09
- Egress ACL
- Upgrade each docker to stretch version
- Everflow enhancement
- Warm Reboot
- IPinIP v6 for IPv4 and IPv6
- Basic VRF
  Microsoft & Mellanox]
- WHAT JUST HAPPENED
- COPP
- Mirroring
- Debian 9 and Kernel 4.9 upgrade
- FRR as default routing stack
- Management
  - SNMP
  - NTP
- ACL
- L3 VXLAN
  Microsoft & Mellanox]
- PMON Refactoring
- LAG, LLDP, QoS, COS,
- Tunnel decap
- Routing Stack Graceful Restart
- DHCP Relay
- Config DB and basic SONiC CLI
- Asymmetric PFC
- IPinIP v6 including all v4 and v6 combinations
- PFC WD
- Fast Reload
  [Microsoft & Mellanox]
- Basic L2, L3 Quagga
- Incremental Config (IP, LAG, Port admin)

HW offload without SDK = Linux Switch

switchdev

Open NOS based on Linux

Open Linux as a NOS



Protocol Stack

Forwarding Agents

Abstraction APIs

Proprietary ASIC SDK

Linux Kernel

ASIC kernel drivers

CPU

Platform HW

ASIC

Switch Hardware

Open source

Protocol Stack

Netlink

Linux Kernel

Linux Network Stack

Open ASIC kernel driver

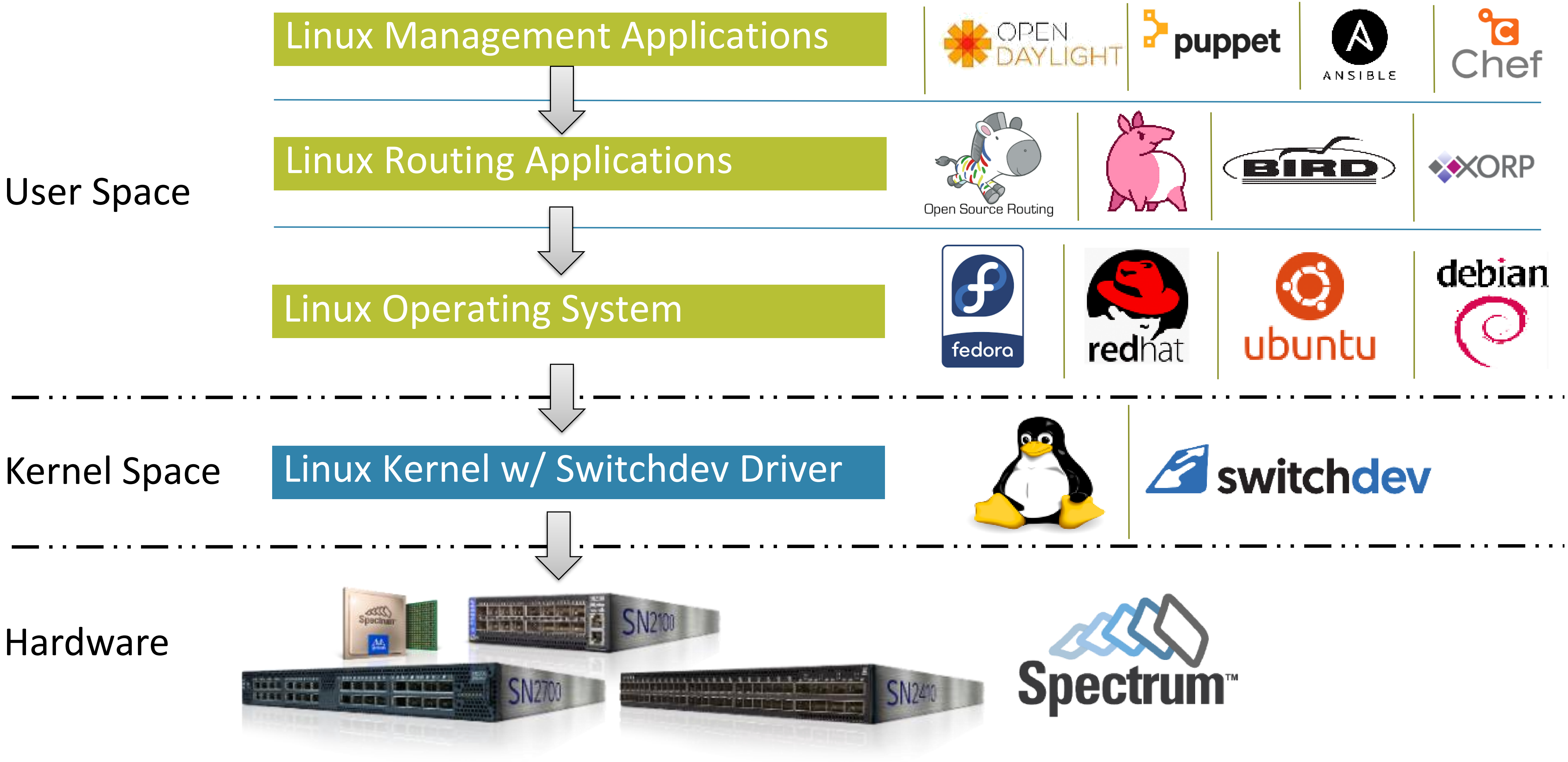
CPU

Platform HW

ASIC

Switch Hardware

Linux Switch architecture



User Space

Linux Management Applications

Linux Routing Applications

Linux Operating System

Kernel Space

Linux Kernel w/ Switchdev Driver

Hardware

Available today [Linux Kernel 5.0] – feature list

Visibility and Maintainability

- [ER]SPAN
- Temperature
- Fans
- LED Control
- ethtool (port counter, FW version, transceiver data)
- Resource queries
- RIF counters

Protocols

- Bridge - 802.1D
- VLAN - 802.1Q
- LAG
- LLDP
- IGMP snooping
- Unicast IPv4/IPv6 router
- ECMP
- DCB
- QoS
- IGMP flood control
- sFlow
- 256 VRFs
- GRE tunnelling
- Multicast IPv4/IPv6 router
- IPv4/IPv6 weighted ECMP
- VRRP
- VxLAN

ACL

- tc-flower offload
- Actions: Drop, Forward, Counters, Trap, TC_ACT_OK
- TC chain template
- Keys: Port, DMAC, SMAC, Ethertype, IP proto, SIP DIP (IPv4/6), TCP/UDP, L4 port, VLAN-ID, PCP, DCSP, VLAN valid, TCP flags

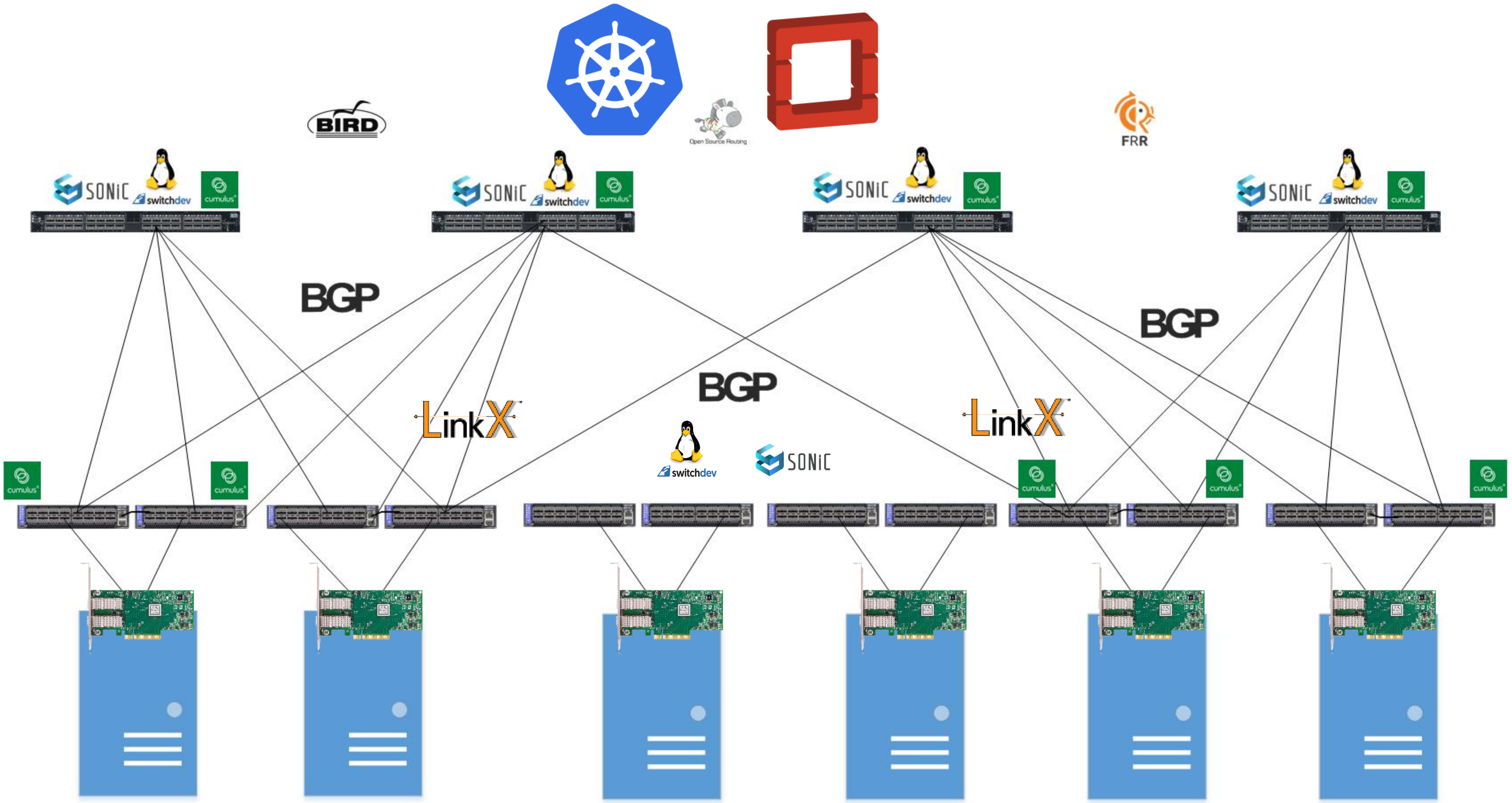
Misc

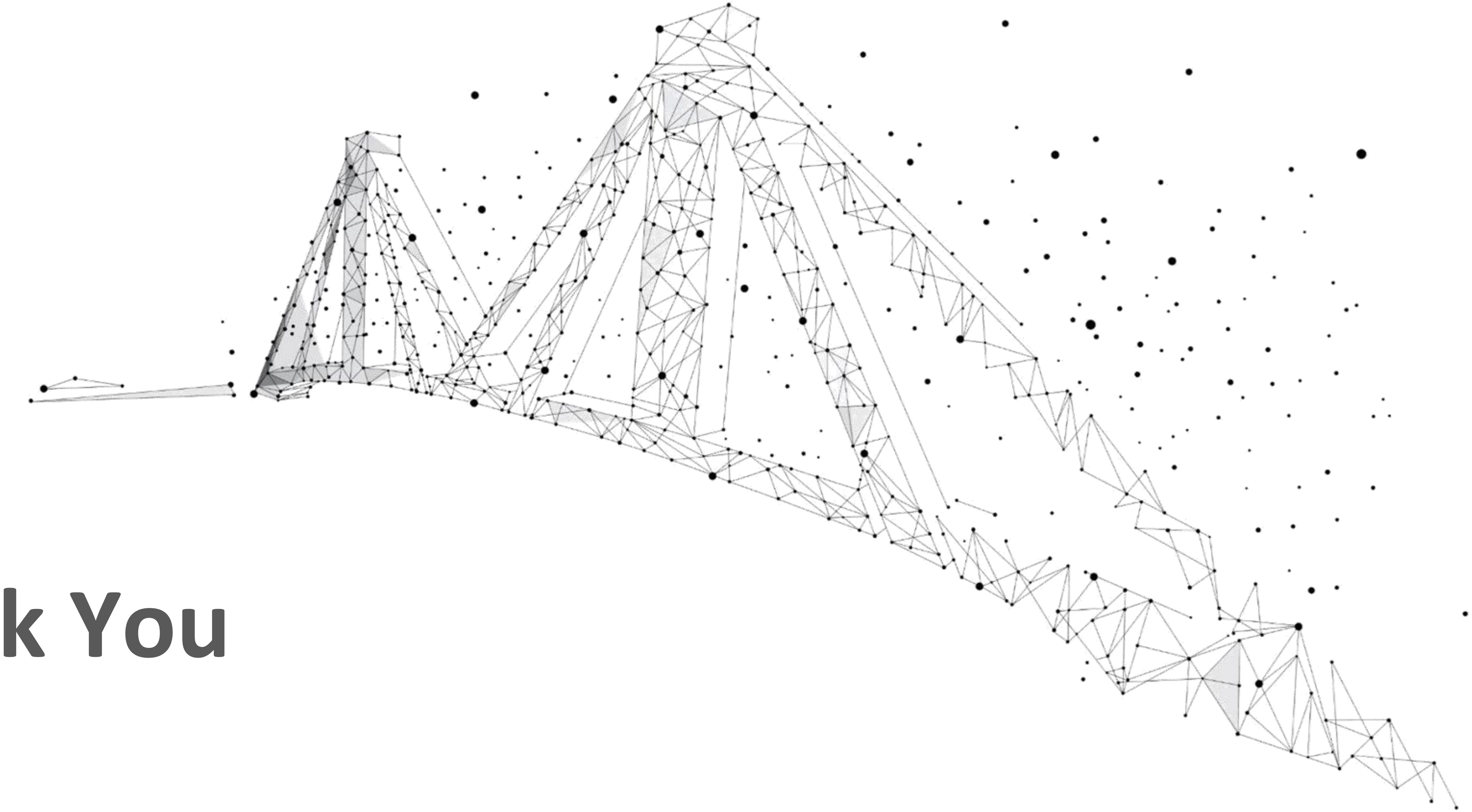
- ‘devlink’ tool
- Port splitter
- Shared buffer configuration
- Internal secured FW upgrade
- ECN: RED and PRIO

Open NOS in production?

- **Why not?**
  - Microsoft, Alibaba - SONiC
  - Facebook - FBOSS
  - Ngenix (CDN) - Linux/switchdev
  - Russia Biggest Bank (Cloud) - Linux/siwtchdev
  - …
- **Vendor Support**
  - SONiC - Mellanox, Dell, Edge-Core, Arista, *Cisco*, … 20+
  - Linux/switchdev - Linux, Mellanox, Cumulus, ALT …

- **Feature Richness**
  - L2 bridging
    - Linux bridge, MSTPd
  - L3 routing
    - Open OSPF/BGP implementations - 20+ yrs
    - Quagga/FRR/Bird/…
  - Tunneling - GRE/VXLAN
    - Linux kernel, OVS
    - EVPN - FRR/GoBGP
  - Security/Isolation - ACL, VRF
    - iptables, Linux TC, Linux NS/VRF
  - Management, monitoring
    - SNMPd, hsFlowd, Grafana


BIRD
Open Source Routing
FRR
SONiC
switchdev
cumulus
BGP
BGP
BGP
LinkX
LinkX



Thank You