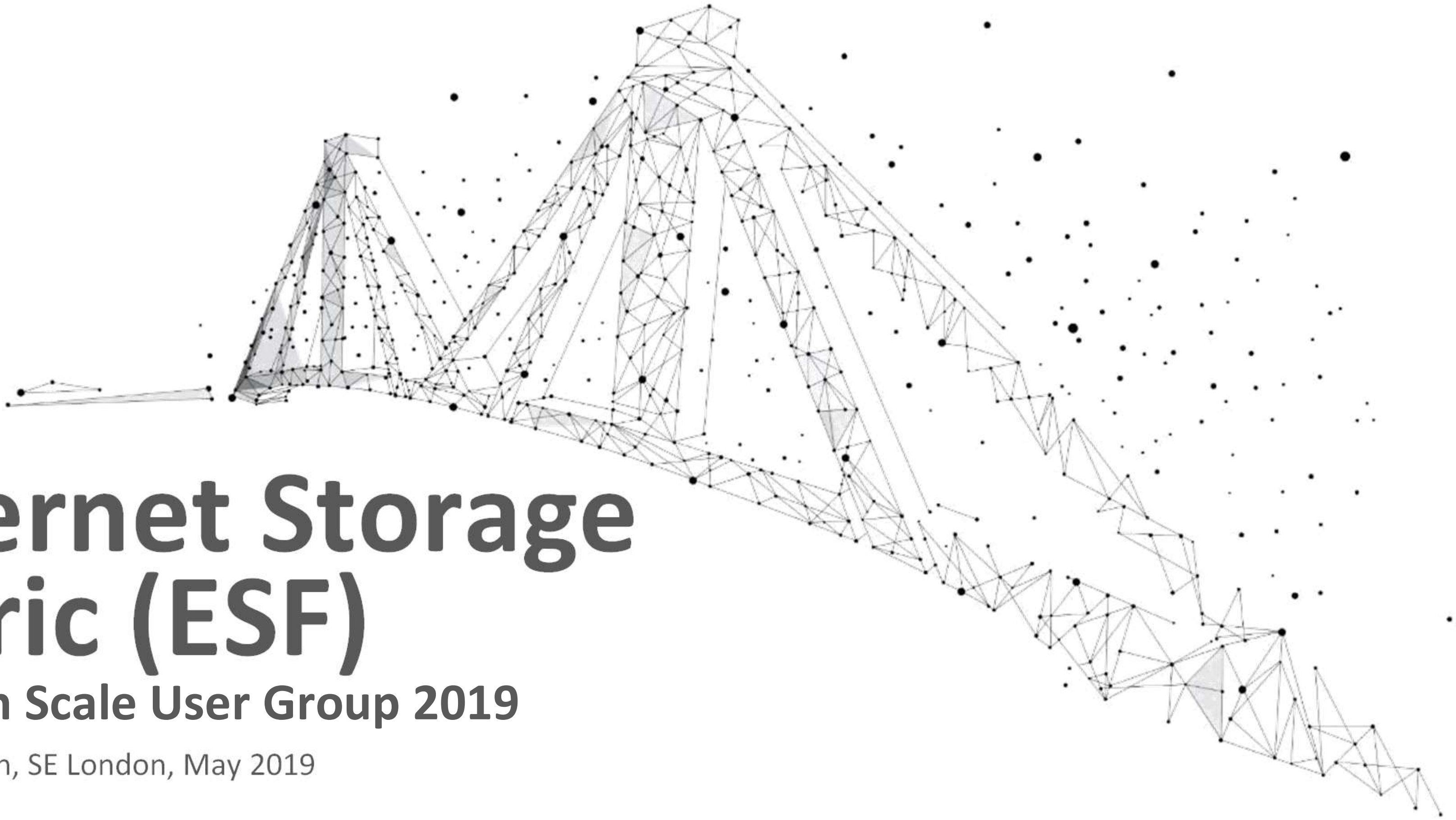
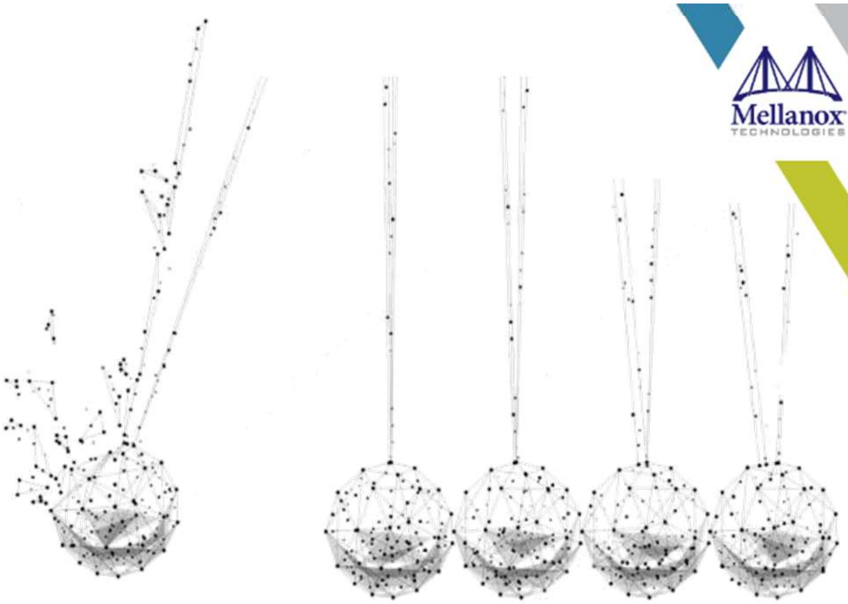


Ethernet Storage Fabric (ESF)

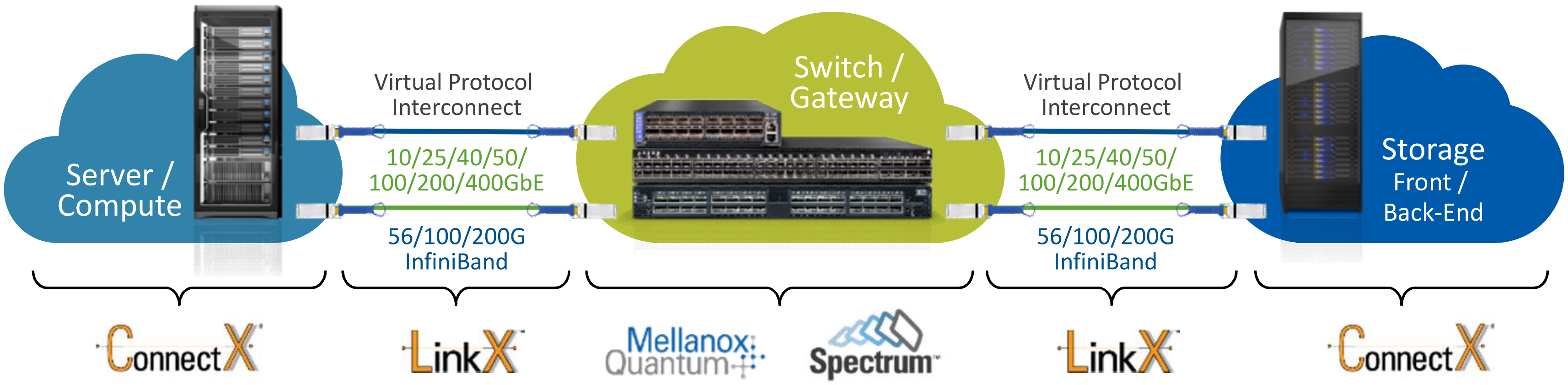
Spectrum Scale User Group 2019

Gadi Godanyan, SE London, May 2019

Leading Supplier of InfiniBand and Ethernet End-to-End Interconnect Solutions



The Smart Choice for Intelligent Compute and Storage Platforms



Ethernet Storage Fabric (ESF)

Best in class Networking for HCI

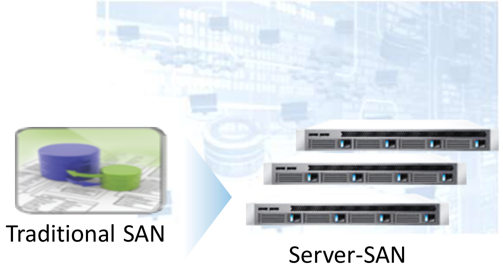
The Storage World is Changing

Changes	Effects
Flash, Faster servers	**Faster networking;** 10/25/40/50/100GbE
Social/Mobile/Video	**Huge data growth;** more file and object content
Hyperconverged	**Distributed “Server-SAN”** on Ethernet
Cloud	**Virtualization**, software-defined, price pressure
Big Data	**File and distributed** storage
Distributed applications	**More** east-west traffic

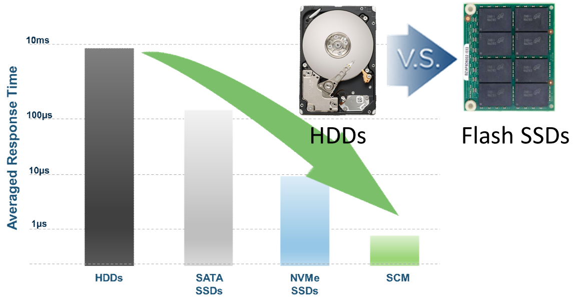
Bottom Line: More Ethernet Storage Traffic

Data Centers Are Changing to Accommodate HCI

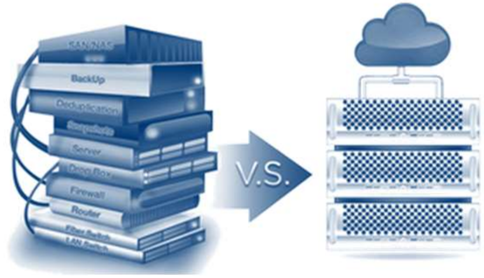
Scale-out Architecture



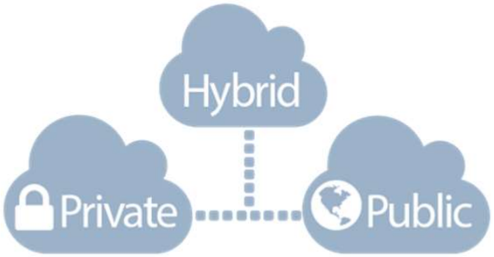
Flash Storage



Hyperconverged Infrastructure

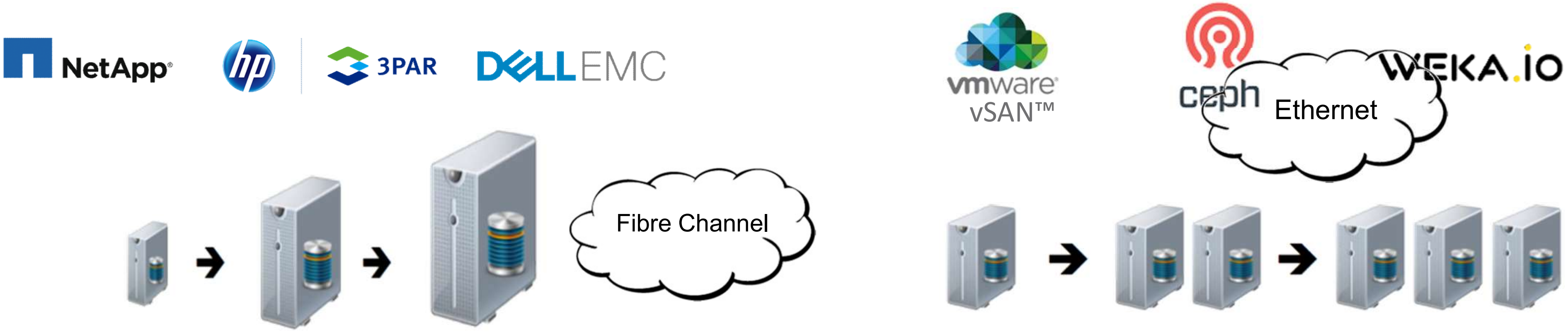


Cloud and Hybrid Cloud



Ethernet the de facto Storage Network

Scale-Out Storage Needs Faster Networks



Scale-Up Storage

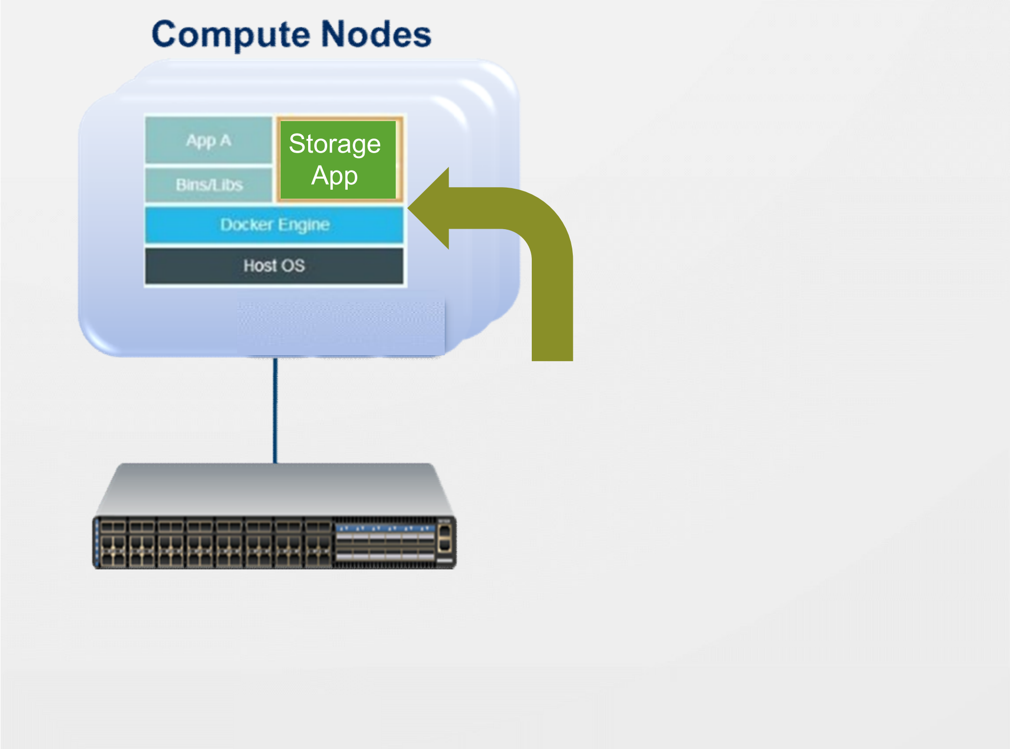
Scale-Out Storage

- Clouds abandoned traditional SAN
- Enterprises migrating to a cloud model
- Storage innovation is using the scale-out architecture

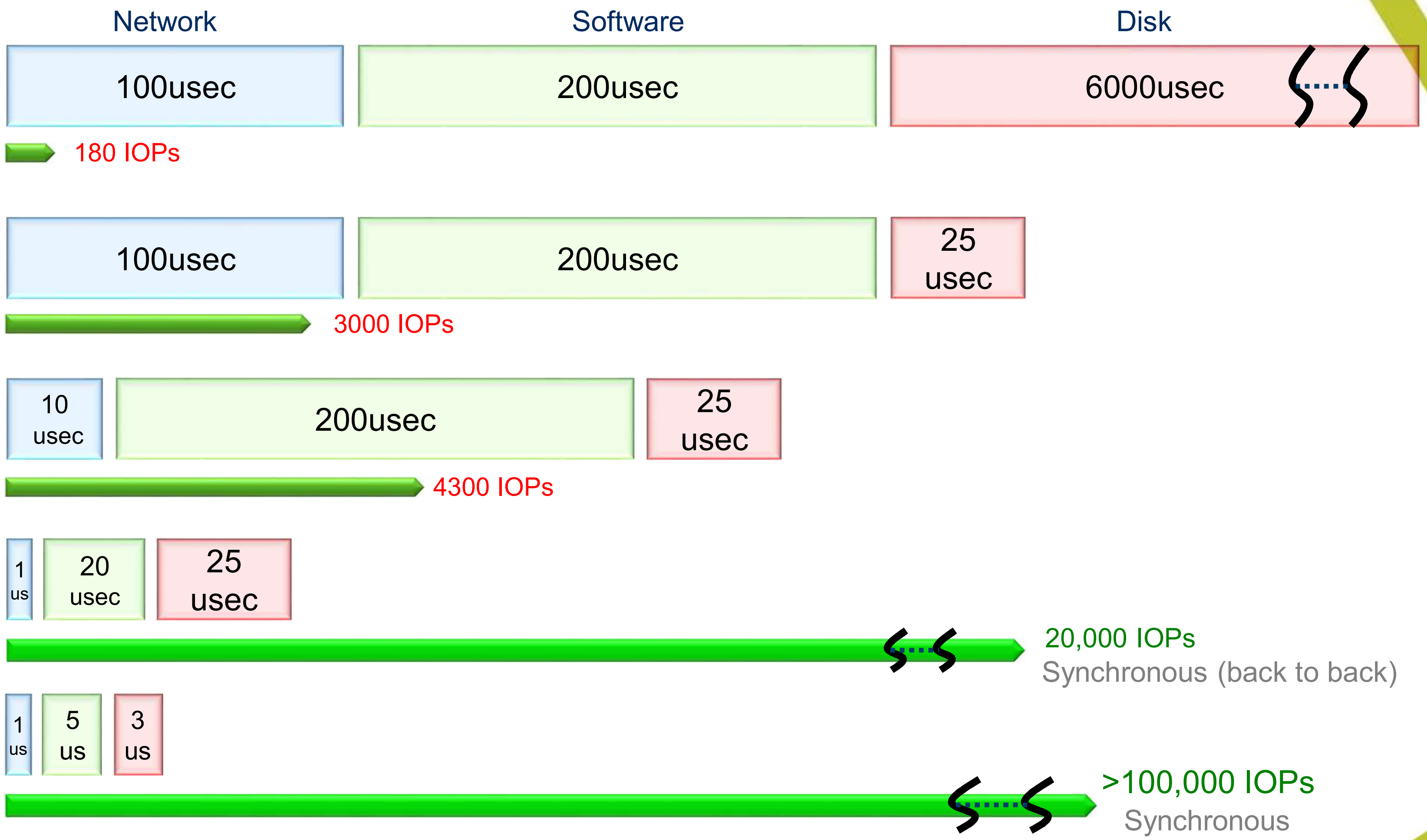
Hyperconverged Storage Needs Faster Networks

- Hyperconverged infrastructure (HCI) collapses compute & scale-out storage into one layer
- Faster networks enable higher performance applications

vmware® Microsoft NetApp™ NUTANIX™ Excelero



Chasing the Storage (synch) IOPs Bottleneck



	Network	Software	Disk	IOPs
Mechanical Disks (~6msec)	100usec	200usec	6000usec	180 IOPs
With SSDs (~0.5msec)	100usec	200usec	25 usec	3000 IOPs
With Fast Network (~0.2msec)	10 usec	200usec	25 usec	4300 IOPs
With RDMA/RoCE (~0.05msec)	1 us	20 usec	25 usec	20,000 IOPs Synchronous (back to back)
With Full OS Bypass & NV-Dimm/Cache (~0.007msec)	1 us	5 us	3 us	>100,000 IOPs Synchronous

Spectrum the Best ESF Switch

Ethernet Storage Fabric needs dedicated storage switches

- Performance
- High Availability
- Simple
- Automated
- Scalable
- Cost Efficient

Spectrum™

- ✓ 2 Switches in 1RU
- ✓ Storage/HCI port count
- ✓ Zero Packet Loss
- ✓ Low Latency
- ✓ RoCE optimized switches (NVMe-oF)
- ✓ NEO for Network automation/visibility
- ✓ Native SDK on a container
- ✓ Cost optimized
- ✓ NOS alternatives

Open Ethernet 25/50/100 Switch Portfolio

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

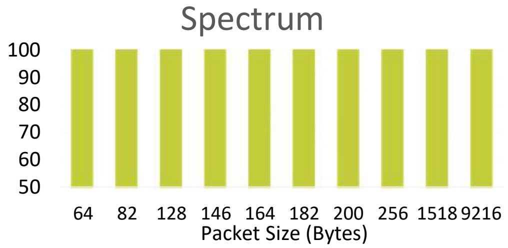
- 32x 40/100GbE
- 64x 10/25/50GbE
- Supports 1GbE ports

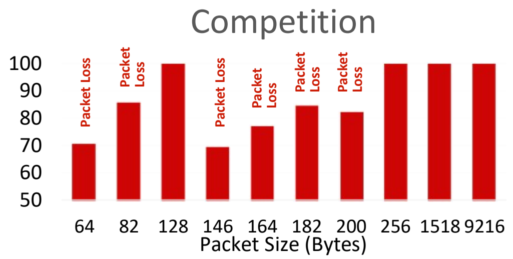
SN3700C
32x100GbE Leaf/Spine
- Up to 128x 25GbE with split cables

SN3800
64x100GbE Spine/Super-Spine

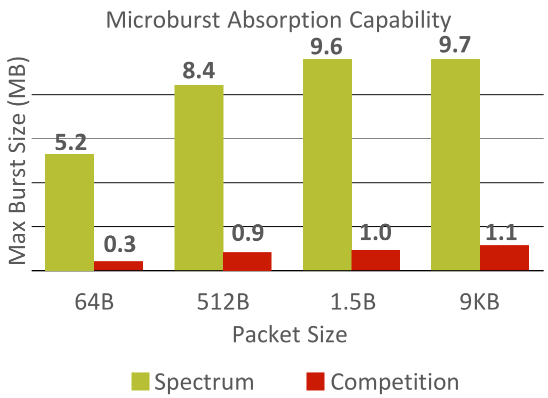
Spectrum is Purpose-Built for ESF

Avoidable Packet Loss

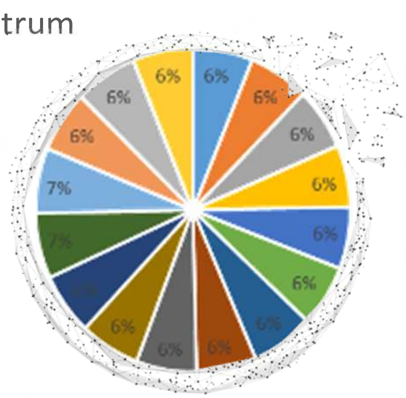


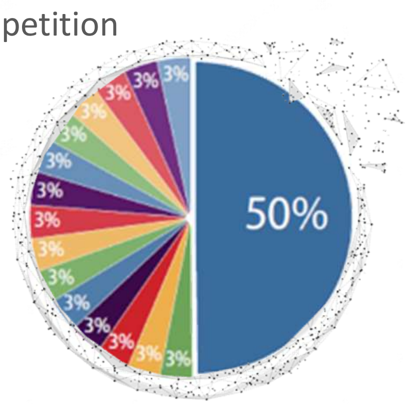


Congestion Management

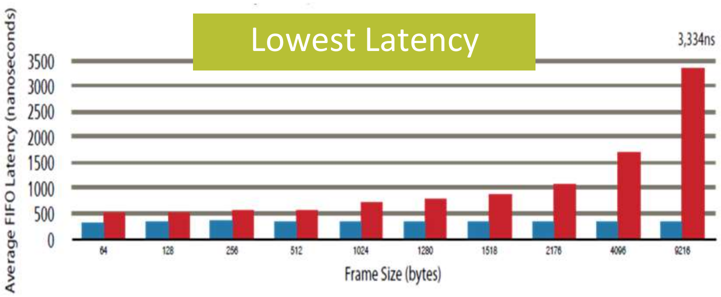


Fairness & QoS





Lowest Latency



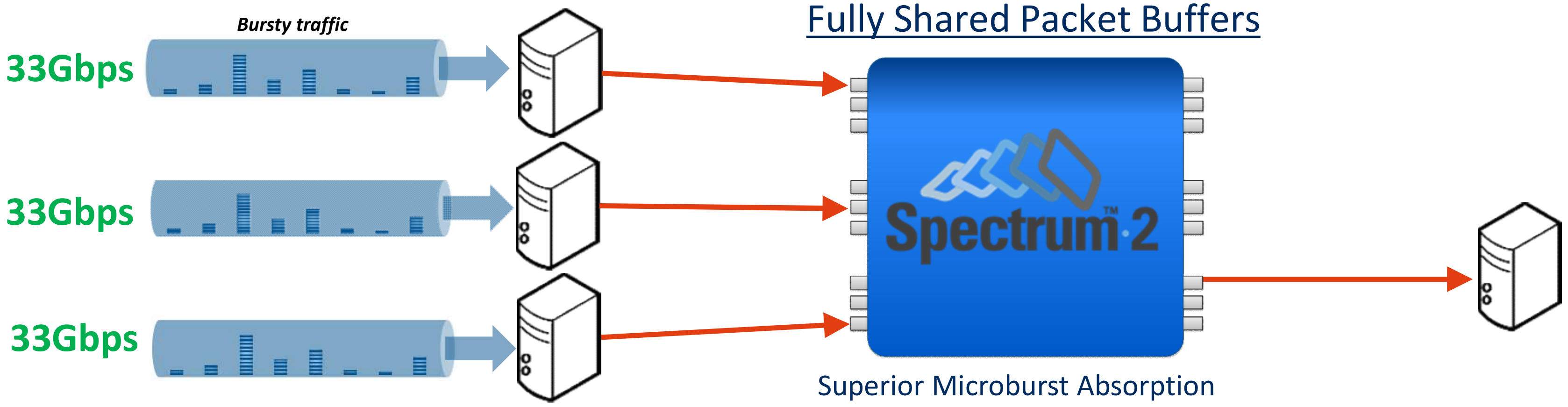
TEST REPORT
Tolly.

www.zeropacketloss.com

www.Mellanox.com/tolly

Fully Shared Buffers

Fully Shared Packet Buffers



Bursty traffic

33Gbps

33Gbps

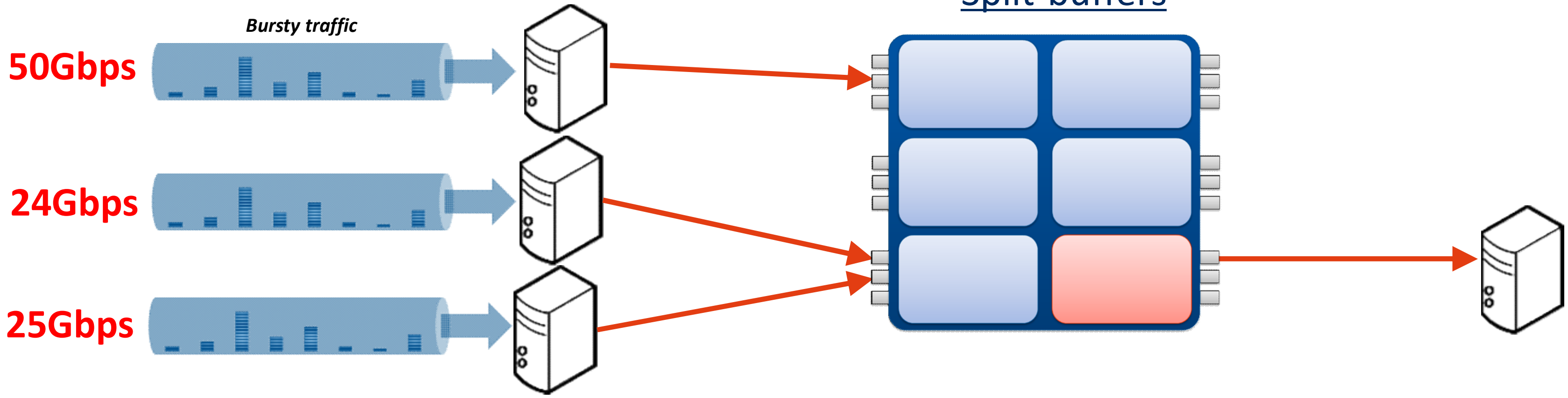
33Gbps

Spectrum™-2

Superior Microburst Absorption

4-6X larger effective buffer size!

Split-buffers



Bursty traffic

50Gbps

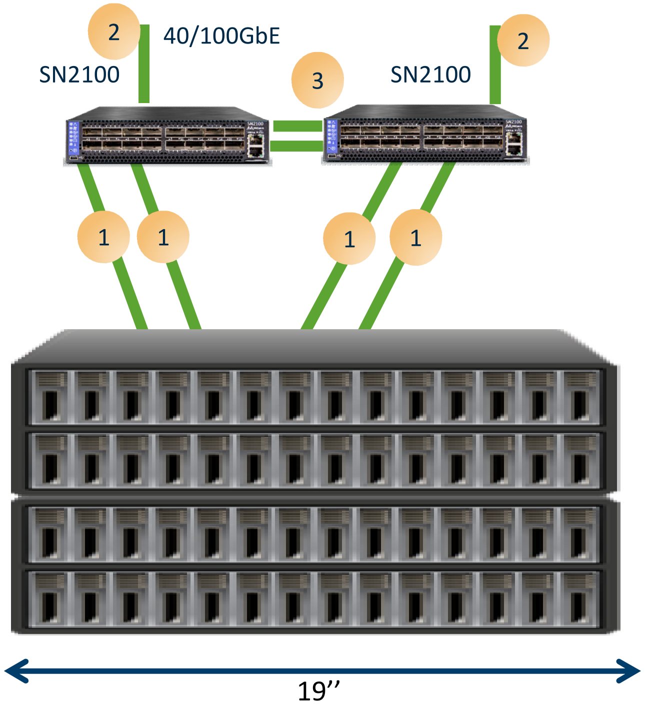
24Gbps

25Gbps

A Typical HCI/NVMe-oF Deployment

Mellanox Storage Rack Design

(1/10/25 & 40/50/100GbE supported)



1 rack = 18 nodes

1. 10/25 or 40/50/100GbE link: QSFP or SFP (using QSA, DAC or fiber)

2. 1/10/25 or 40/50/100GbE link: QSFP or SFP (using QSA, DAC or fiber)

3. 100GbE mLAG Links: QSFP28 to QSFP28

- **½ 19" width, 1RU height**
- **95W max power**
- **16x100GbE**
- **Any speed from 1-100GbE**

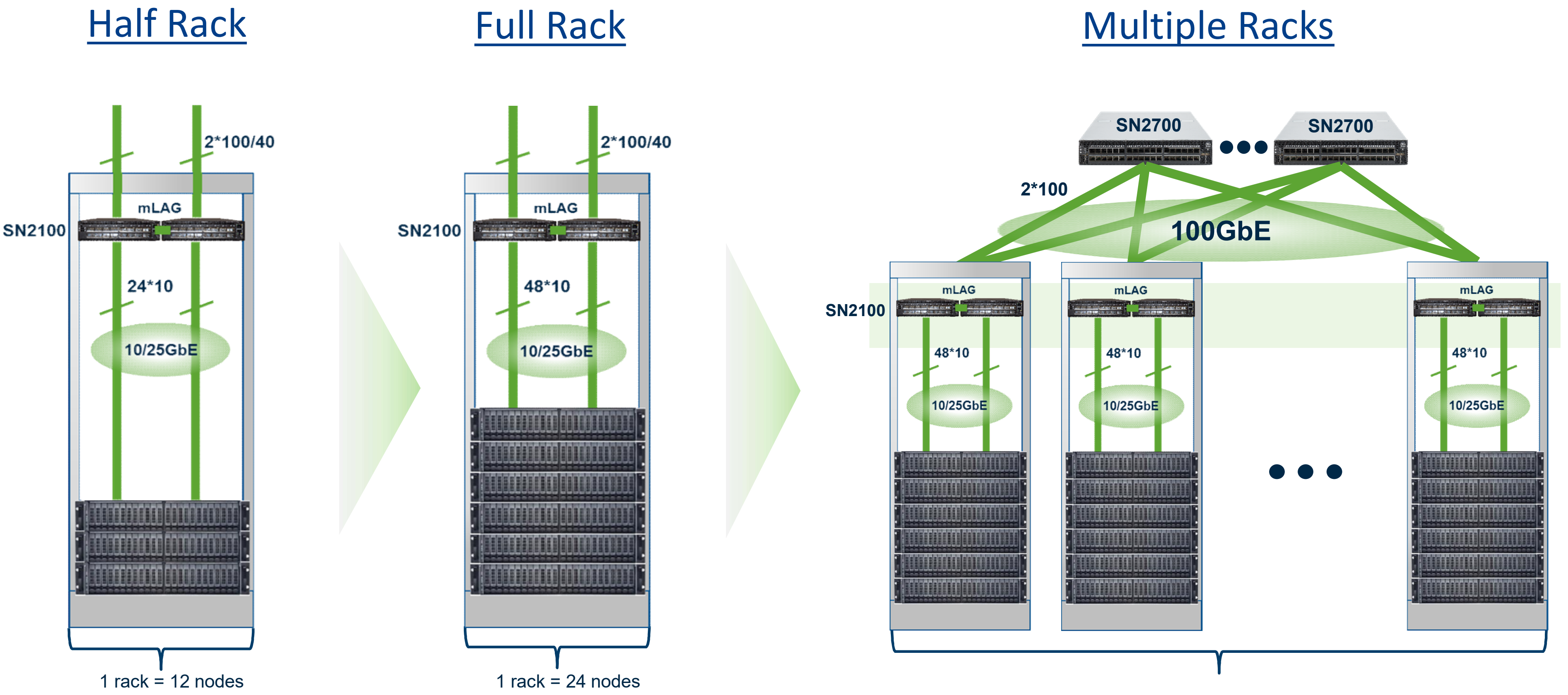
Performance

100GbE Optimized

Best SWaP
Size Weight and Power

Best $/Gb/s

Mellanox ESF Scale

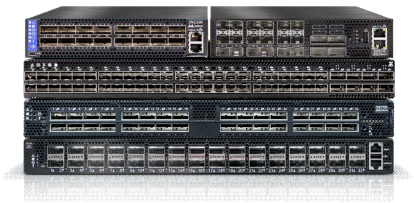


Mellanox NEO’s Value Proposition

Simplified management, automation and orchestration for Mellanox end-to-end Ethernet portfolio

ConnectX®

Adapters



Spectrum™

Switches

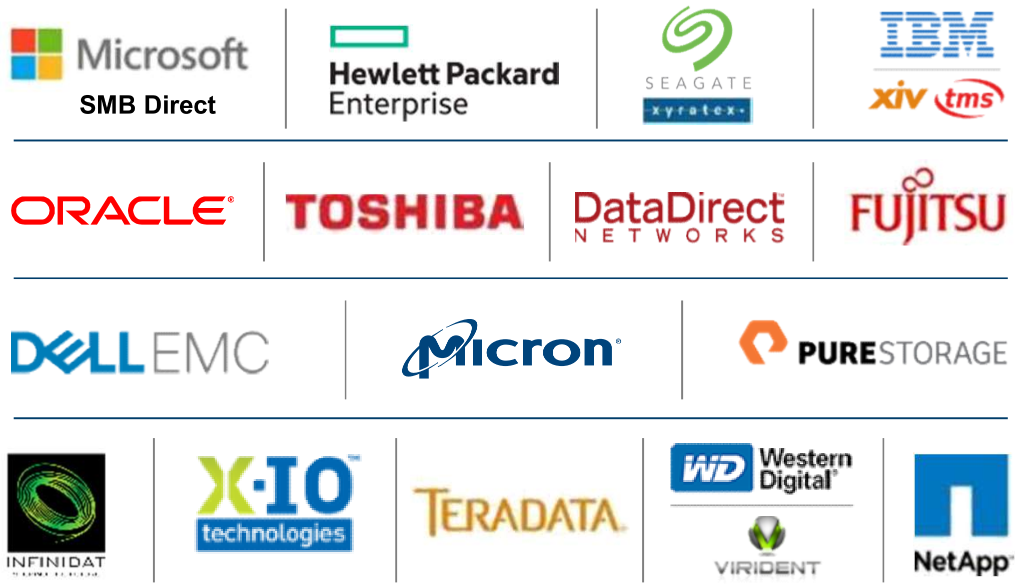
LinkX™

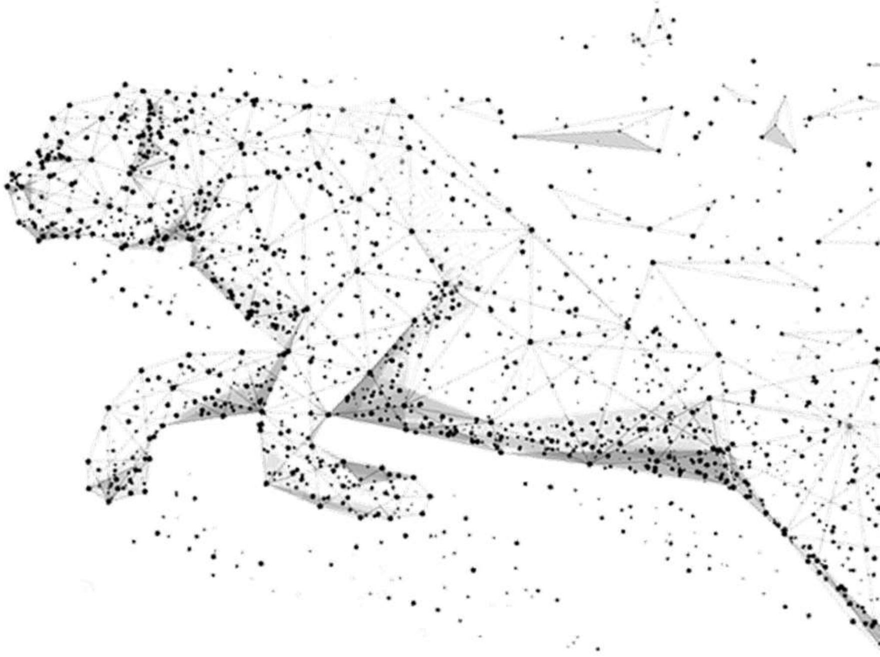
Cables & Transceivers

Management Software

Partner Storage Solution

Mellanox Empowers Leading Storage Platforms





RDMA – Remote Direct Memory Access

Why RDMA is needed?

- Traditional network processing has data copy, utilizes TCP/IP and has significant CPU overhead
- Packets stored in OS memory vs host memory

What is RDMA?

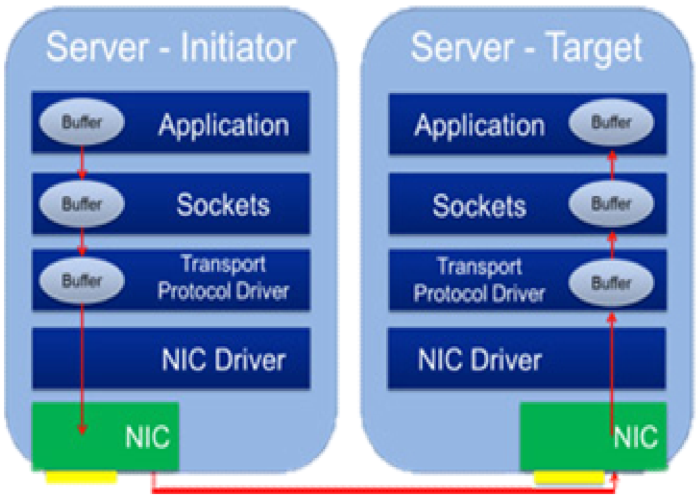
- RDMA is the direct read from or write to an application’s memory by use of HW
- RDMA NIC writes data directly to application host memory
- Enables the movement of data between servers with no CPU involvement

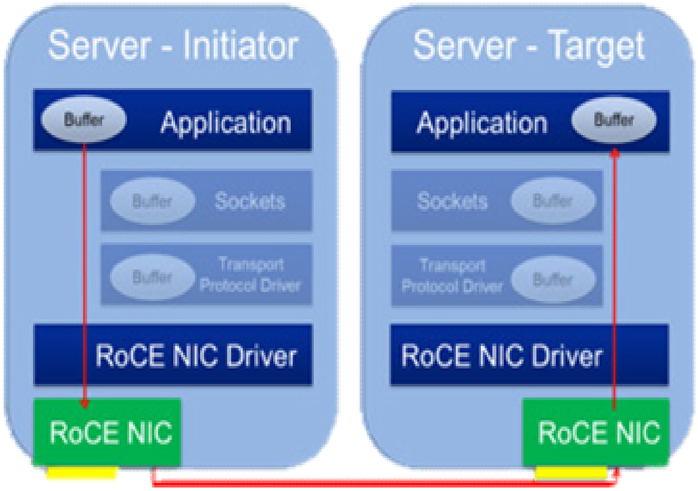
Benefits of RDMA

- Higher server productivity cost and power saving
- Provides low latency, high throughput, low CPU usage.
- Offloads CPU network processing (OS TCP/IP stack)
- Avoids data copy between user space and kernel space

RoCE - RDMA over Converged ETH

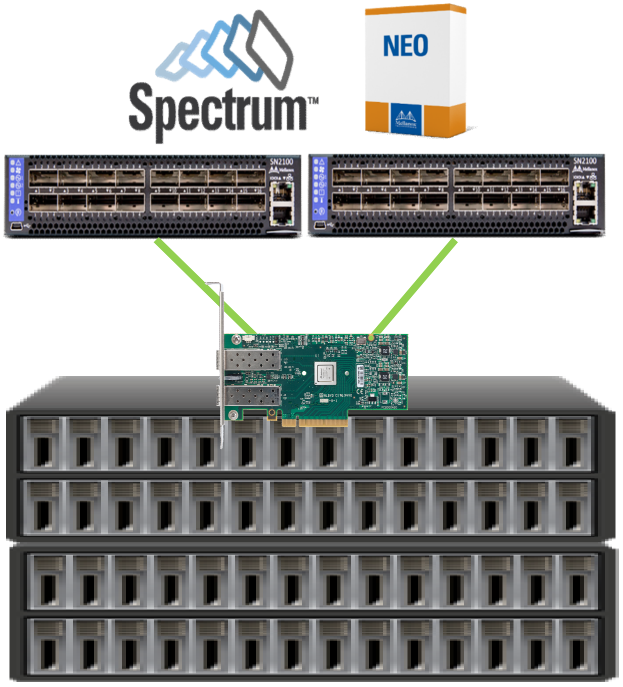
- Benefits of RDMA in standard ETH environment
- Open source and formal standard in IBTA
- Available in: Linux, Windows, Vmware and inFreeBSD





Mellanox ESF for RoCE-based Storage

RoCE | Mellanox TECHNOLOGIES



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

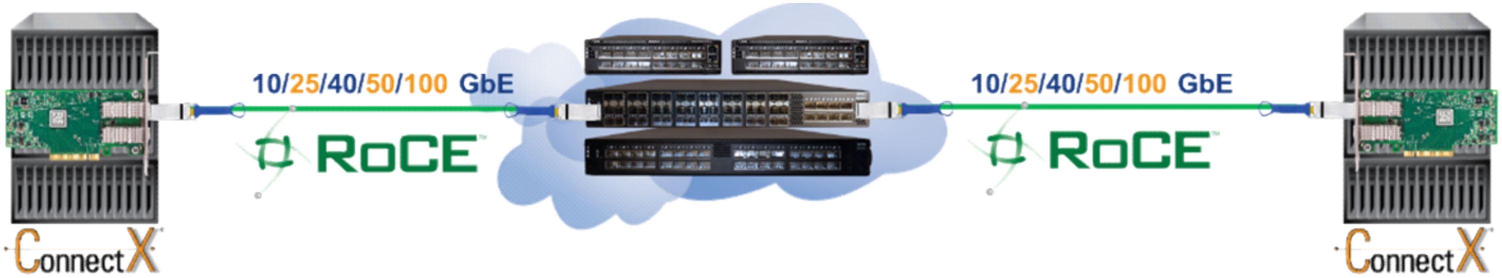
Low Latency

Easy Configuration

Guaranteed QoS

Automated Mgmt

Mellanox ESF Provides E2E RoCE Acceleration



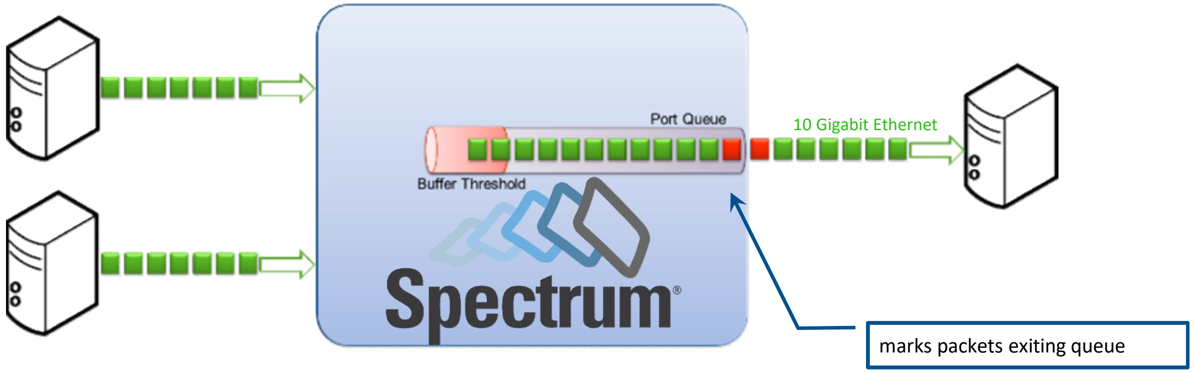
- Zero packet loss, line-rate performance at all packet sizes and port combination
- Predictable buffer allocation to any port & packet sizes
- Low latency, up to 90% latency in a typical TOR deployment scenario

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

RoCE Enabled Storage

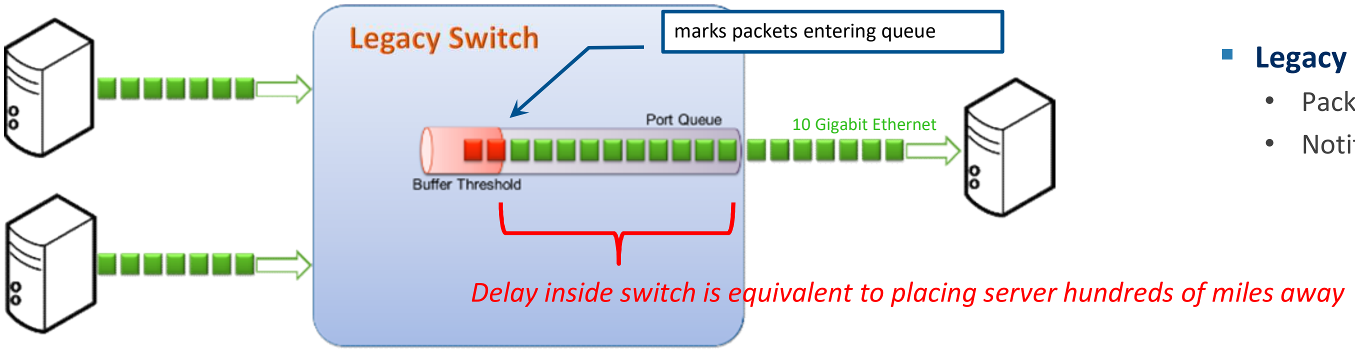
- **iSER**
- **NVMe-oF**
- **Microsoft SMB 3.0**
- **vSphere 6.5**
- **Ceph**
- **Spark**

Better RoCE with Explicit Congestion Notification



marks packets exiting queue

- **Fast Congestion Notification**
  - Packets marked as they leave queue
  - Up to 10ms faster alerts
  - Servers react faster
  - Reduces average queue depth
    - Lowers real world latency
  - ***Improves application performance***



marks packets entering queue

Delay inside switch is equivalent to placing server hundreds of miles away

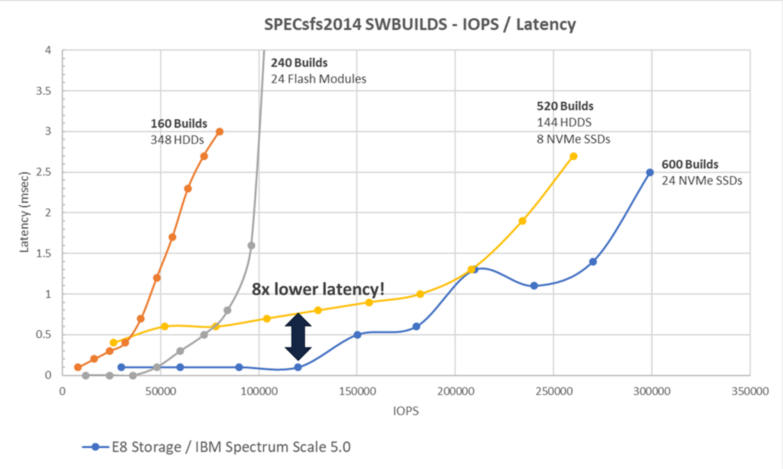
- **Legacy Congestion Notification**:
  - Packets marked as they enter queue
  - Notification delayed until queue empties

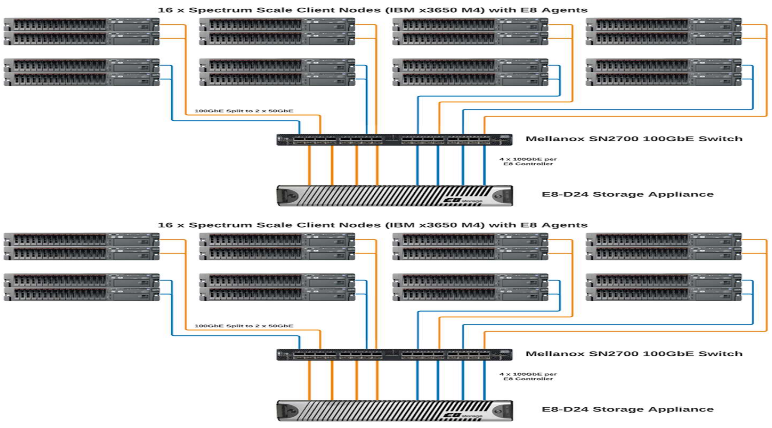
E8 High-Performance NVMe Storage

1st Sub-millisecond Overall Response Time (ORT) – a World Record!

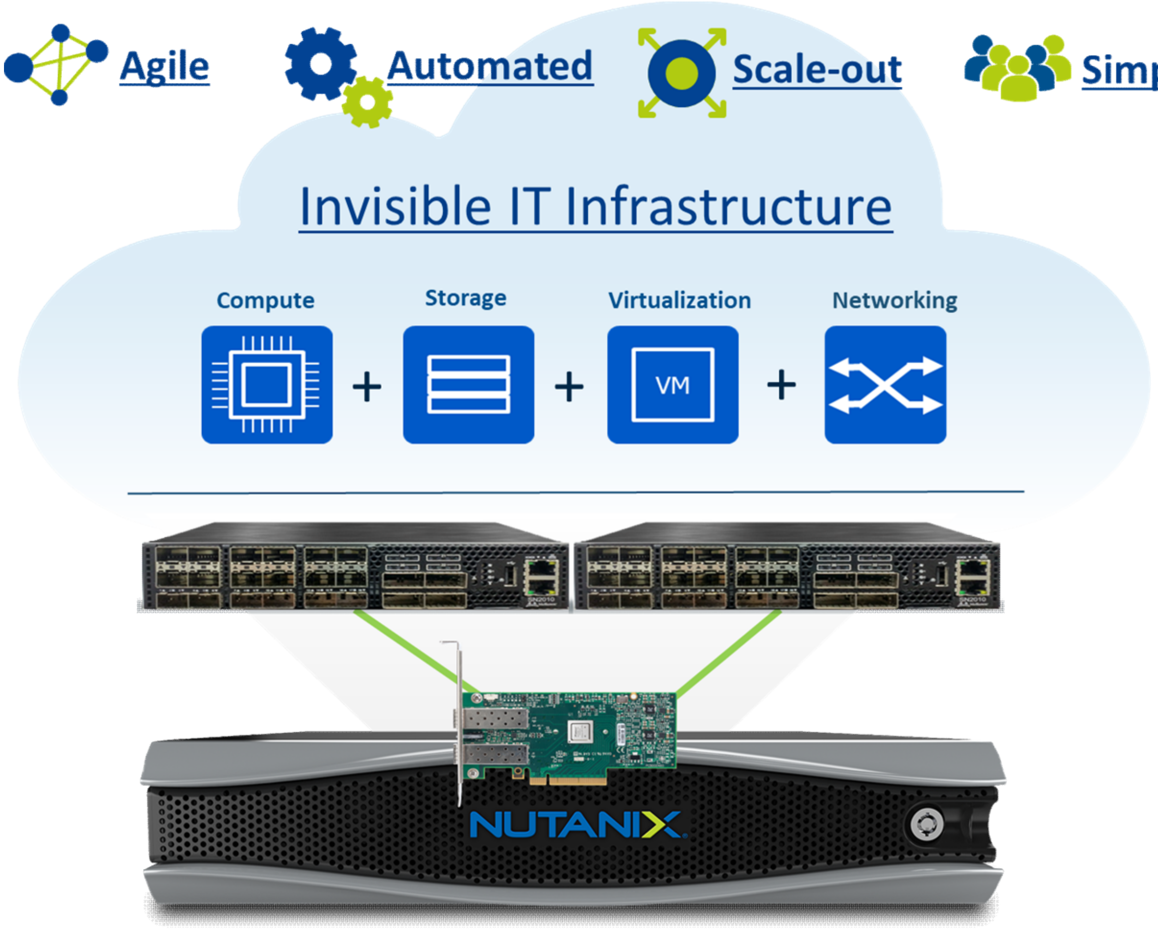
- SPEC SFS2014 Software build benchmark results
  - **600 builds and 0.69ms ORT**
  - Previous record – 520 builds and 1.04ms ORT
- E8 Storage configuration
  - 16 IBM Spectrum Scale client nodes simulate software build workloads
    - Mellanox ConnectX-4 100GbE adapter
    - E8 can support up to 96 host servers providing compute services
  - **Mellanox SN2700 32-port 100GbE switch**
  - E8-D24, high availability NVMe enclosure
    - 4 Mellanox ConnectX-4 100GbE adapters
    - 24 dual port NVMe SSDs from leading vendors
  - End-to-end RoCE

References:
- SPEC SFS 2014 benchmark smashed by storage newbie
- E8 solution brief




16 x Spectrum Scale Client Nodes (IBM x3650 M4) with E8 Agents
100GbE Split to 2 x 50GbE
Mellanox SN2700 100GbE Switch
4 x 100GbE per E8 Controller
E8-D24 Storage Appliance
16 x Spectrum Scale Client Nodes (IBM x3650 M4) with E8 Agents
100GbE Split to 2 x 50GbE
Mellanox SN2700 100GbE Switch
4 x 100GbE per E8 Controller
E8-D24 Storage Appliance

Nutanix + Mellanox



Agile

Automated

Scale-out

Simple

Invisible IT Infrastructure

Compute + Storage + Virtualization + Networking

VM

Nutanix Elevate Partner of the Year

2018 for Calm Blueprint • 2017 for Nutanix Ready

> We have seen worldwide deployment and great customer experience together with Mellanox. Our customers and channel partners are realizing the value of Mellanox as the perfect complement to Nutanix enterprise cloud solutions.

Venugopal Pai
VP, Customer Success, Nutanix

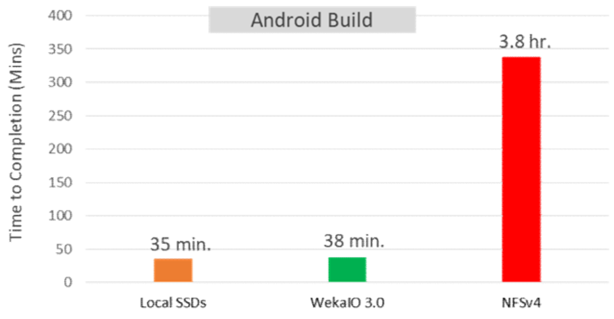
WekaIO Cloud-native High Performance File System

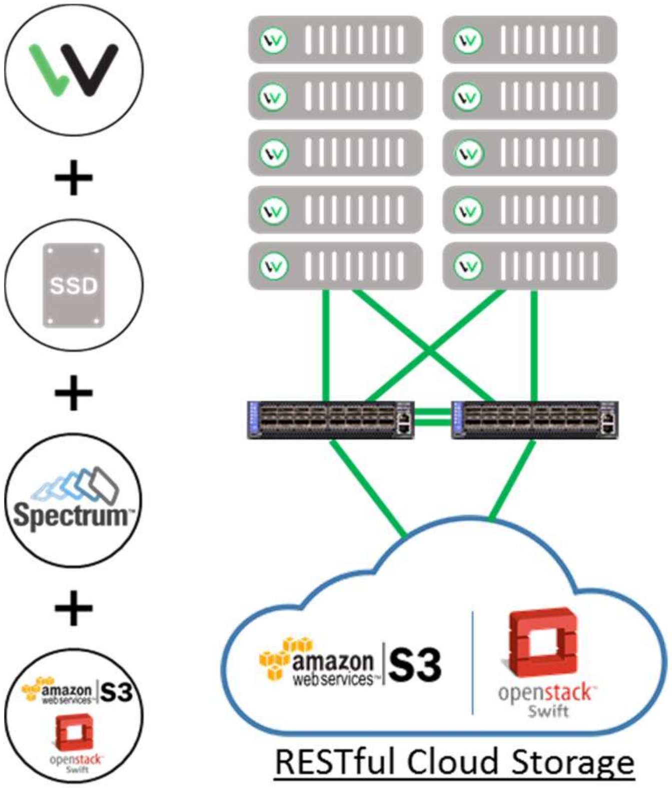
Scale-out, shared file system

- Local FS performance with linear scalability
  - 4K random file R/W performance – 50k / 11K IOPS per core
  - 1M sequential file R/W performance – 980 / 370 MB/sec per core
  - Latency – ~150 microseconds (QD=1)
  - Linearly scale with the number of cores
- Flexible usage - hyperconverged, dedicated storage server, or a mixed topology in bare metal, virtualized, or private cloud environments
- Designed for demanding workloads: Machine Learning, Genomics, M&E and EDA

Example Use Case: Compiling Android

- Performance comparable to local file system
- No copying of data needed anymore
- 6x Faster than Oracle ZFS





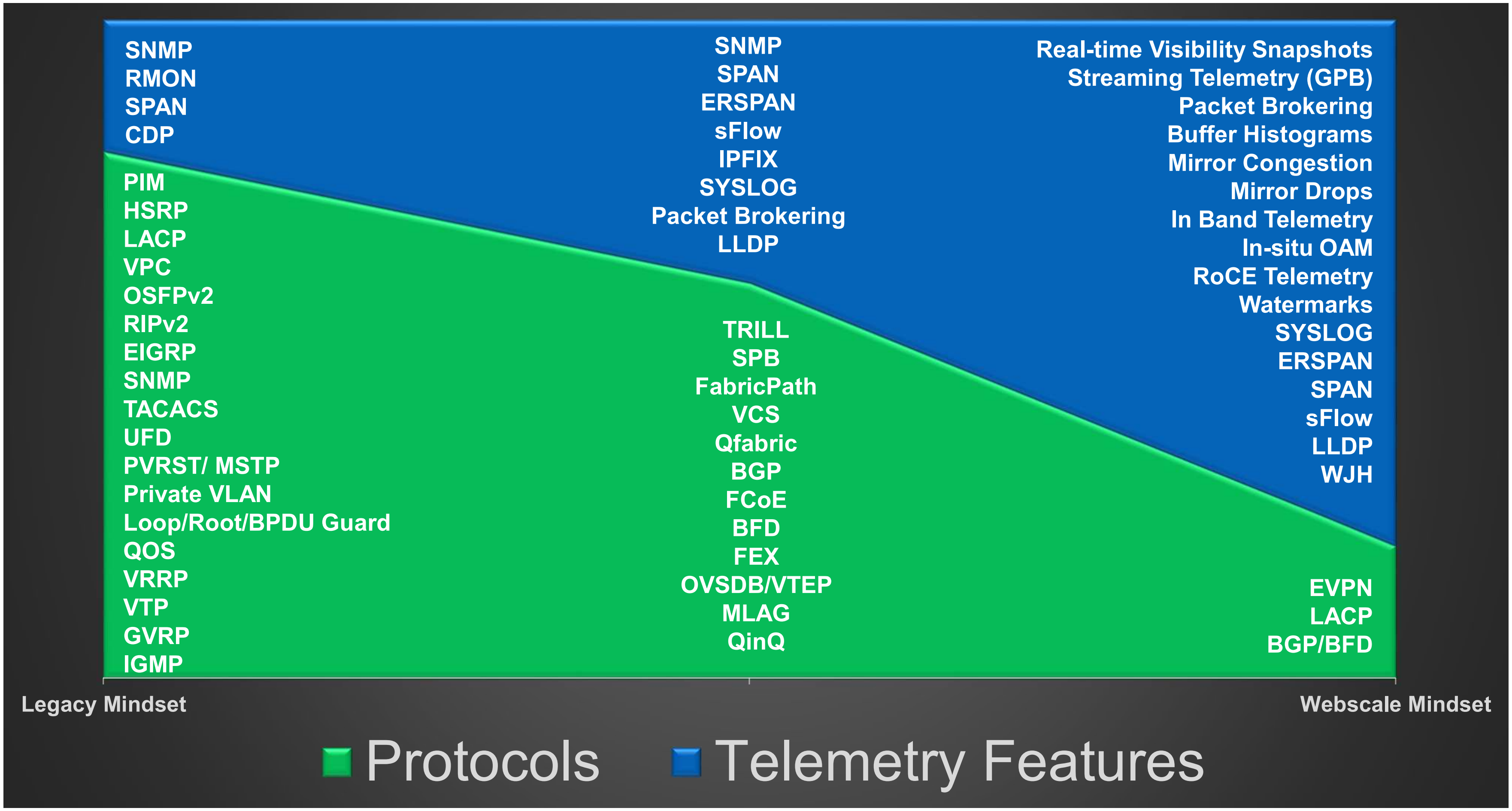
RESTful Cloud Storage

Source: Solution Brief

What Just Happened (WJH)

Best in class Telemetry

Data Center Evolution



SNMP
RMON
SPAN
CDP

PIM
HSRP
LACP
VPC
OSFPv2
RIPv2
EIGRP
SNMP
TACACS
UFD
PVRST/ MSTP
Private VLAN
Loop/Root/BPDU Guard
QOS
VRRP
VTP
GVRP
IGMP

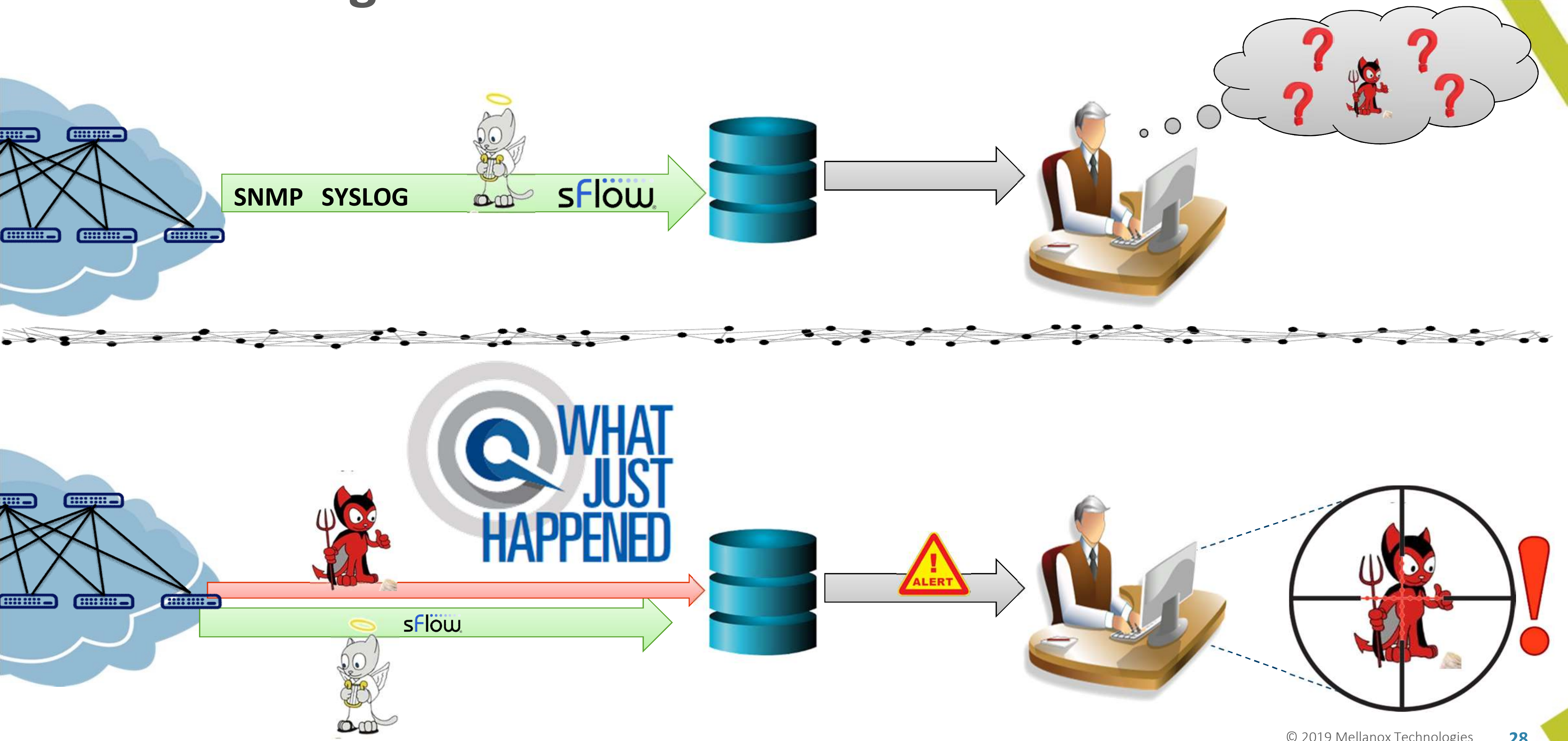
SNMP
SPAN
ERSPAN
sFlow
IPFIX
SYSLOG
Packet Brokering
LLDP

TRILL
SPB
FabricPath
VCS
Qfabric
BGP
FCoE
BFD
FEX
OVSDB/VTEP
MLAG
QinQ

Real-time Visibility Snapshots
Streaming Telemetry (GPB)
Packet Brokering
Buffer Histograms
Mirror Congestion
Mirror Drops
In Band Telemetry
In-situ OAM
RoCE Telemetry
Watermarks
SYSLOG
ERSPAN
SPAN
sFlow
LLDP
WJH

EVPN
LACP
BGP/BFD

Accelerating the Time to Root-Cause

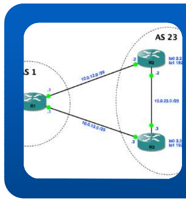


What Does WJH Monitor?

Packet Drop

L1

- Bad CRC
- Flaky cable



L2/L3

- BGP
- VLAN

Buffer

- Incast
- Rate Limit

ACLs

- Deny based on IP
- Deny based on VLAN

No Packet Drop

Congestion

- Incast
- Busy storage device

Latency

- Pause frames
- Congestion → latency

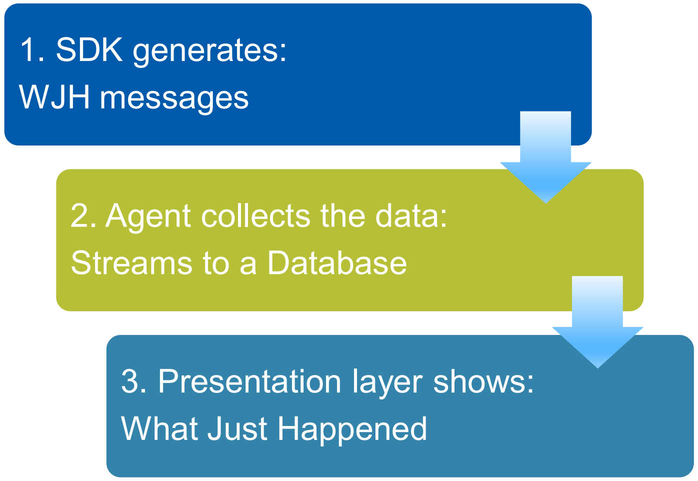
Route Validation

- Packet doesn’t reach firewall
- Packet takes suboptimal path

Load Balance Validation

- Suboptimal ECMP
- Suboptimal LAG

WJH – How Does It Work?



1. SDK generates: WJH messages

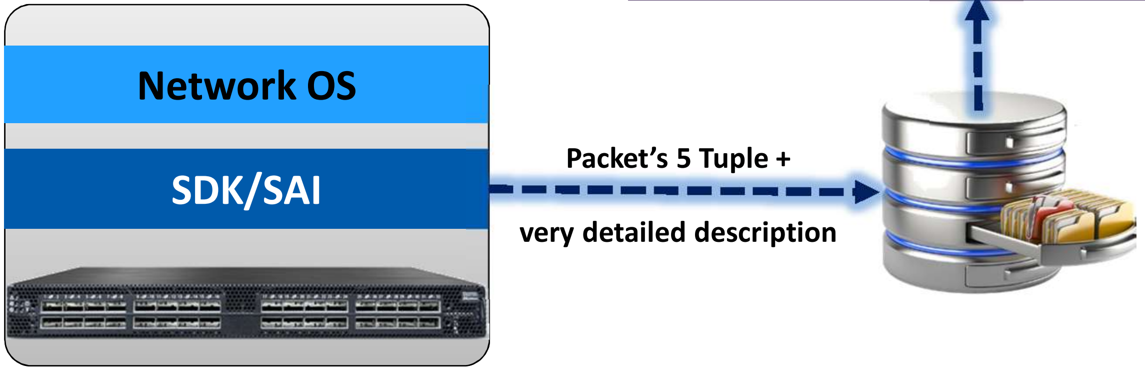
2. Agent collects the data: Streams to a Database

3. Presentation layer shows: What Just Happened

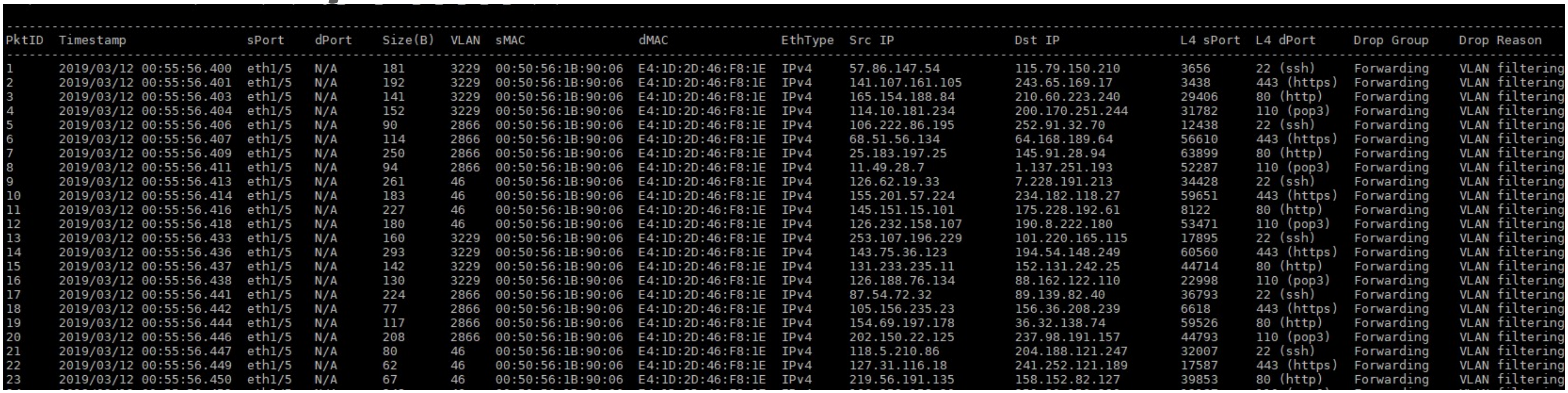
The Important Questions

✓WHO	is being impacted
✓WHAT	is causing the problem
✓WHERE	is the problem
✓WHEN	it happened
✓WHY	it is happening

Root Cause + how to fix it

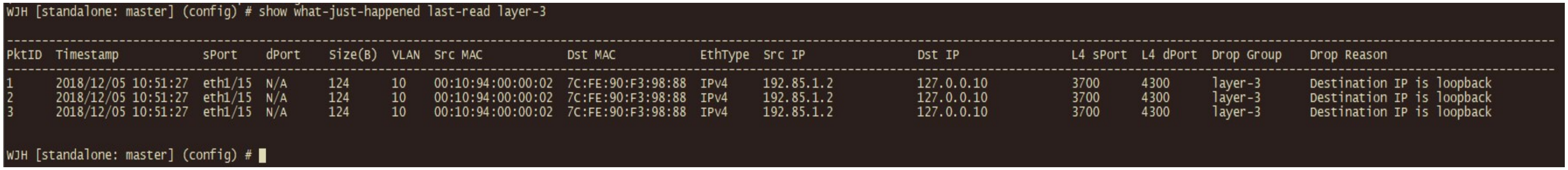


WJH on Onyx CLI - Show Commands



```
----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
PktID  Timestamp                sPort    dPort    Size(B)  VLAN  sMAC               dMAC               EthType  Src IP              Dst IP              L4 sPort  L4 dPort     Drop Group   Drop Reason
----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
1      2019/03/12 00:55:56.400  eth1/5   N/A      181      3229  00:50:56:1B:90:06  E4:1D:2D:46:F8:1E  IPv4     57.86.147.54        115.79.150.210      3656      22 (ssh)     Forwarding   VLAN filtering
2      2019/03/12 00:55:56.401  eth1/5   N/A      192      3229  00:50:56:1B:90:06  E4:1D:2D:46:F8:1E  IPv4     141.107.161.105     243.65.169.17       3438      443 (https)  Forwarding   VLAN filtering
3      2019/03/12 00:55:56.403  eth1/5   N/A      141      3229  00:50:56:1B:90:06  E4:1D:2D:46:F8:1E  IPv4     165.154.188.84      210.60.223.240      29406     80 (http)    Forwarding   VLAN filtering
4      2019/03/12 00:55:56.404  eth1/5   N/A      152      3229  00:50:56:1B:90:06  E4:1D:2D:46:F8:1E  IPv4     114.10.181.234      200.170.251.244     31782     110 (pop3)   Forwarding   VLAN filtering
5      2019/03/12 00:55:56.406  eth1/5   N/A      90       2866  00:50:56:1B:90:06  E4:1D:2D:46:F8:1E  IPv4     106.222.86.195      252.91.32.70        12438     22 (ssh)     Forwarding   VLAN filtering
6      2019/03/12 00:55:56.407  eth1/5   N/A      114      2866  00:50:56:1B:90:06  E4:1D:2D:46:F8:1E  IPv4     68.51.56.134        64.168.189.64       56610     443 (https)  Forwarding   VLAN filtering
7      2019/03/12 00:55:56.409  eth1/5   N/A      250      2866  00:50:56:1B:90:06  E4:1D:2D:46:F8:1E  IPv4     25.183.197.25       145.91.28.94        63899     80 (http)    Forwarding   VLAN filtering
8      2019/03/12 00:55:56.411  eth1/5   N/A      94       2866  00:50:56:1B:90:06  E4:1D:2D:46:F8:1E  IPv4     11.49.28.7          1.137.251.193       52287     110 (pop3)   Forwarding   VLAN filtering
9      2019/03/12 00:55:56.413  eth1/5   N/A      261      46    00:50:56:1B:90:06  E4:1D:2D:46:F8:1E  IPv4     126.62.19.33        7.228.191.213       34428     22 (ssh)     Forwarding   VLAN filtering
10     2019/03/12 00:55:56.414  eth1/5   N/A      183      46    00:50:56:1B:90:06  E4:1D:2D:46:F8:1E  IPv4     155.201.57.224      234.182.118.27      59651     443 (https)  Forwarding   VLAN filtering
11     2019/03/12 00:55:56.416  eth1/5   N/A      227      46    00:50:56:1B:90:06  E4:1D:2D:46:F8:1E  IPv4     145.151.15.101      175.228.192.61      8122      80 (http)    Forwarding   VLAN filtering
12     2019/03/12 00:55:56.418  eth1/5   N/A      180      46    00:50:56:1B:90:06  E4:1D:2D:46:F8:1E  IPv4     126.232.158.107     190.8.222.180       53471     110 (pop3)   Forwarding   VLAN filtering
13     2019/03/12 00:55:56.433  eth1/5   N/A      160      3229  00:50:56:1B:90:06  E4:1D:2D:46:F8:1E  IPv4     253.107.196.229     101.220.165.115     17895     22 (ssh)     Forwarding   VLAN filtering
14     2019/03/12 00:55:56.436  eth1/5   N/A      293      3229  00:50:56:1B:90:06  E4:1D:2D:46:F8:1E  IPv4     143.75.36.123       194.54.148.249      60560     443 (https)  Forwarding   VLAN filtering
15     2019/03/12 00:55:56.437  eth1/5   N/A      142      3229  00:50:56:1B:90:06  E4:1D:2D:46:F8:1E  IPv4     131.233.235.11      152.131.242.25      44714     80 (http)    Forwarding   VLAN filtering
16     2019/03/12 00:55:56.438  eth1/5   N/A      130      3229  00:50:56:1B:90:06  E4:1D:2D:46:F8:1E  IPv4     126.188.76.134      88.162.122.110      22998     110 (pop3)   Forwarding   VLAN filtering
17     2019/03/12 00:55:56.441  eth1/5   N/A      224      2866  00:50:56:1B:90:06  E4:1D:2D:46:F8:1E  IPv4     87.54.72.32         89.139.82.40        36793     22 (ssh)     Forwarding   VLAN filtering
18     2019/03/12 00:55:56.442  eth1/5   N/A      77       2866  00:50:56:1B:90:06  E4:1D:2D:46:F8:1E  IPv4     105.156.235.23      156.36.208.239      6618      443 (https)  Forwarding   VLAN filtering
19     2019/03/12 00:55:56.444  eth1/5   N/A      117      2866  00:50:56:1B:90:06  E4:1D:2D:46:F8:1E  IPv4     154.69.197.178      36.32.138.74        59526     80 (http)    Forwarding   VLAN filtering
20     2019/03/12 00:55:56.446  eth1/5   N/A      208      2866  00:50:56:1B:90:06  E4:1D:2D:46:F8:1E  IPv4     202.150.22.125      237.98.191.157      44793     110 (pop3)   Forwarding   VLAN filtering
21     2019/03/12 00:55:56.447  eth1/5   N/A      80       46    00:50:56:1B:90:06  E4:1D:2D:46:F8:1E  IPv4     118.5.210.86        204.188.121.247     32007     22 (ssh)     Forwarding   VLAN filtering
22     2019/03/12 00:55:56.449  eth1/5   N/A      62       46    00:50:56:1B:90:06  E4:1D:2D:46:F8:1E  IPv4     127.31.116.18       241.252.121.189     17587     443 (https)  Forwarding   VLAN filtering
23     2019/03/12 00:55:56.450  eth1/5   N/A      67       46    00:50:56:1B:90:06  E4:1D:2D:46:F8:1E  IPv4     219.56.191.135      158.152.82.127      39853     80 (http)    Forwarding   VLAN filtering
```

recently - read and clear data **from HW**



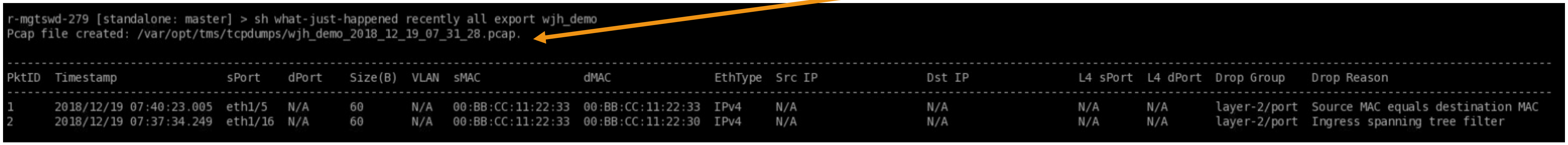
```
WJH [standalone: master] (config) # show what-just-happened last-read layer-3

----------------------------------------------------------------------------------------------------------------------------------------------------------------------------
PktID  Timestamp            sPort    dPort    Size(B)  VLAN  Src MAC            Dst MAC            EthType  Src IP               Dst IP               L4 sPort  L4 dPort  Drop Group   Drop Reason
----------------------------------------------------------------------------------------------------------------------------------------------------------------------------
1      2018/12/05 10:51:27  eth1/15  N/A      124      10    00:10:94:00:00:02  7C:FE:90:F3:98:88  IPv4     192.85.1.2           127.0.0.10           3700      4300      layer-3      Destination IP is loopback
2      2018/12/05 10:51:27  eth1/15  N/A      124      10    00:10:94:00:00:02  7C:FE:90:F3:98:88  IPv4     192.85.1.2           127.0.0.10           3700      4300      layer-3      Destination IP is loopback
3      2018/12/05 10:51:27  eth1/15  N/A      124      10    00:10:94:00:00:02  7C:FE:90:F3:98:88  IPv4     192.85.1.2           127.0.0.10           3700      4300      layer-3      Destination IP is loopback

WJH [standalone: master] (config) # █
```

last-read - read info **from memory cache.** Can be executed several times in order to receive the same historical info until recently is executed again

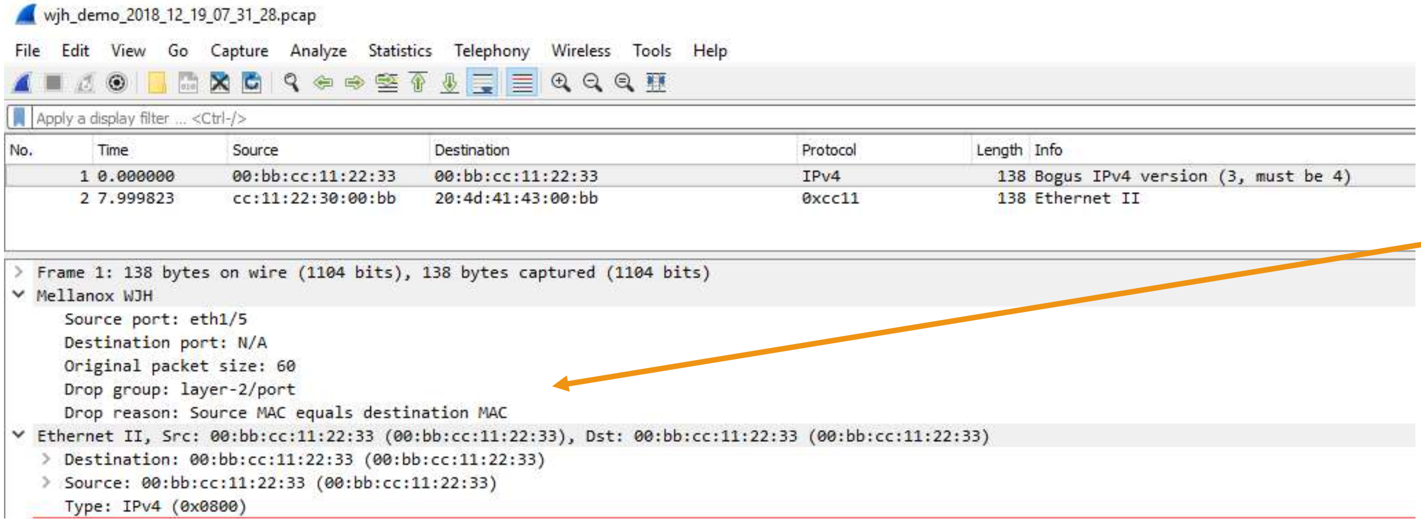
WJH – Create pcap file

- By default, Show recently command creates a pcap file with all the packets and the metadata
- On packet loss, or a critical system failure, the system will autogenerate a .pcap file – user configurable

Pcap file location



```
r-mgtswd-279 [standalone: master] > sh what-just-happened recently all export wjh_demo
Pcap file created: /var/opt/tms/tcpdumps/wjh_demo_2018_12_19_07_31_28.pcap.

------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
PktID  Timestamp                sPort   dPort   Size(B)  VLAN  sMAC               dMAC               EthType  Src IP               Dst IP               L4 sPort  L4 dPort  Drop Group    Drop Reason
------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
1      2018/12/19 07:40:23.005  eth1/5  N/A     60       N/A   00:BB:CC:11:22:33  00:BB:CC:11:22:33  IPv4     N/A                  N/A                  N/A       N/A       layer-2/port  Source MAC equals destination MAC
2      2018/12/19 07:37:34.249  eth1/16 N/A     60       N/A   00:BB:CC:11:22:33  00:BB:CC:11:22:30  IPv4     N/A                  N/A                  N/A       N/A       layer-2/port  Ingress spanning tree filter
```

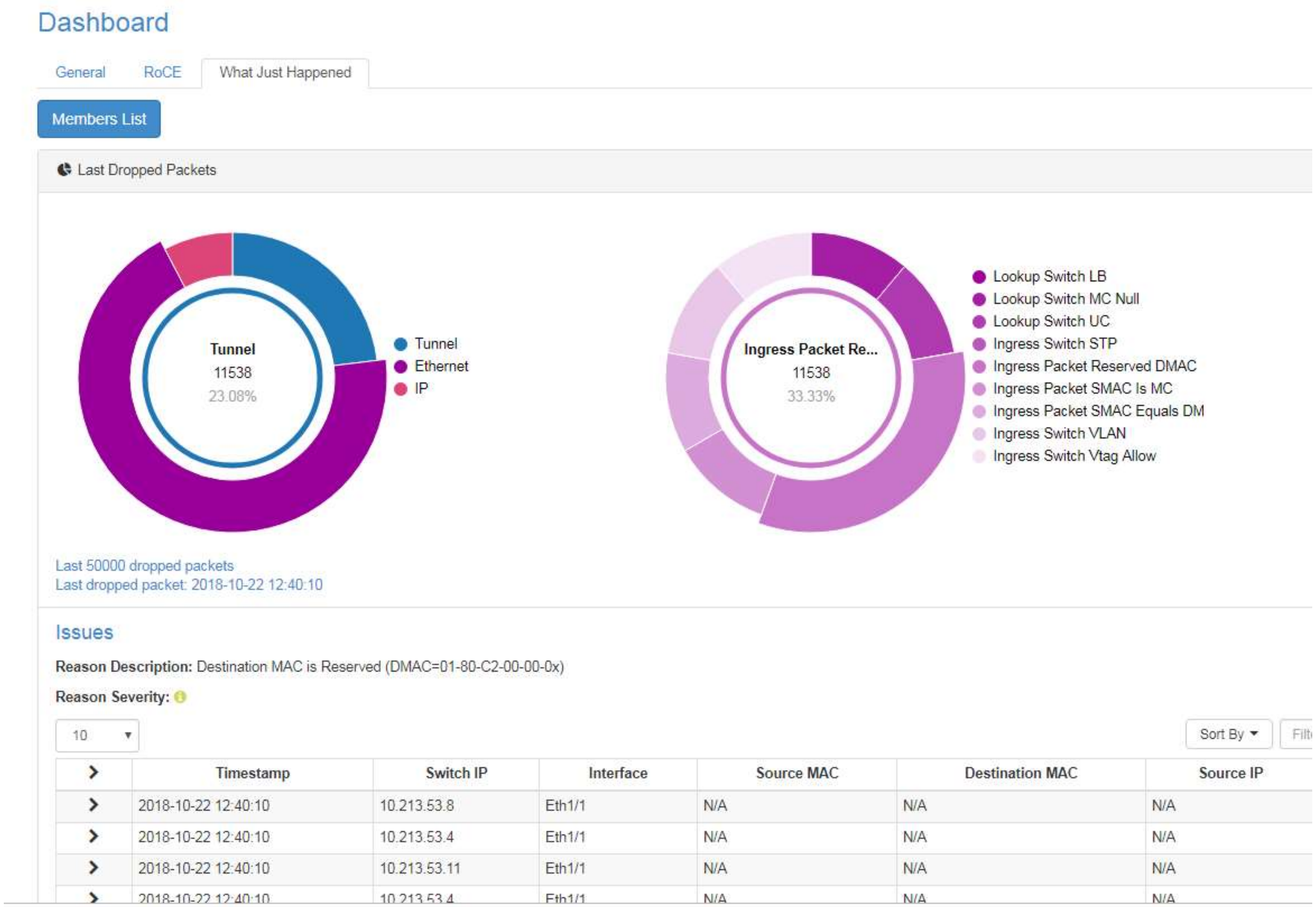

wjh_demo_2018_12_19_07_31_28.pcap
File
Edit
View
Go
Capture
Analyze
Statistics
Telephony
Wireless
Tools
Help
Apply a display filter … <Ctrl-/>

```
No.   Time        Source              Destination         Protocol   Length  Info
    1 0.000000    00:bb:cc:11:22:33   00:bb:cc:11:22:33   IPv4          138  Bogus IPv4 version (3, must be 4)
    2 7.999823    cc:11:22:30:00:bb   20:4d:41:43:00:bb   0xcc11        138  Ethernet II

> Frame 1: 138 bytes on wire (1104 bits), 138 bytes captured (1104 bits)
v Mellanox WJH
    Source port: eth1/5
    Destination port: N/A
    Original packet size: 60
    Drop group: layer-2/port
    Drop reason: Source MAC equals destination MAC
v Ethernet II, Src: 00:bb:cc:11:22:33 (00:bb:cc:11:22:33), Dst: 00:bb:cc:11:22:33 (00:bb:cc:11:22:33)
  > Destination: 00:bb:cc:11:22:33 (00:bb:cc:11:22:33)
  > Source: 00:bb:cc:11:22:33 (00:bb:cc:11:22:33)
    Type: IPv4 (0x0800)
```

Packet metadata

Added wireshark dissector, which enables users to analyze WJH pcap files. It displays the packets' added metadata

NEO – Fabric Level View



Last 50000 dropped packets
Last dropped packet: 2018-10-22 12:40:10

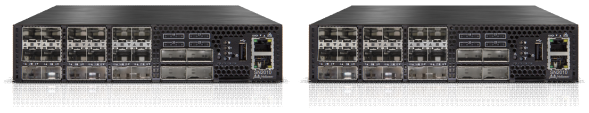
Issues

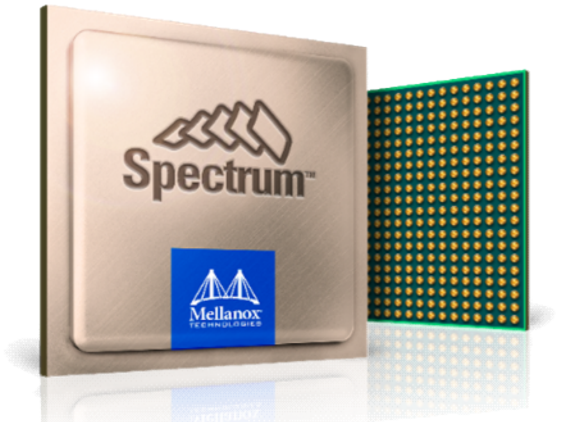
Reason Description: Destination MAC is Reserved (DMAC=01-80-C2-00-00-0x)

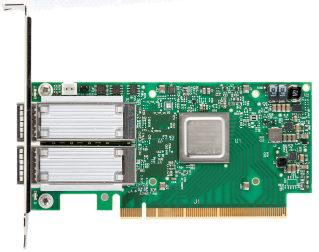
Reason Severity:

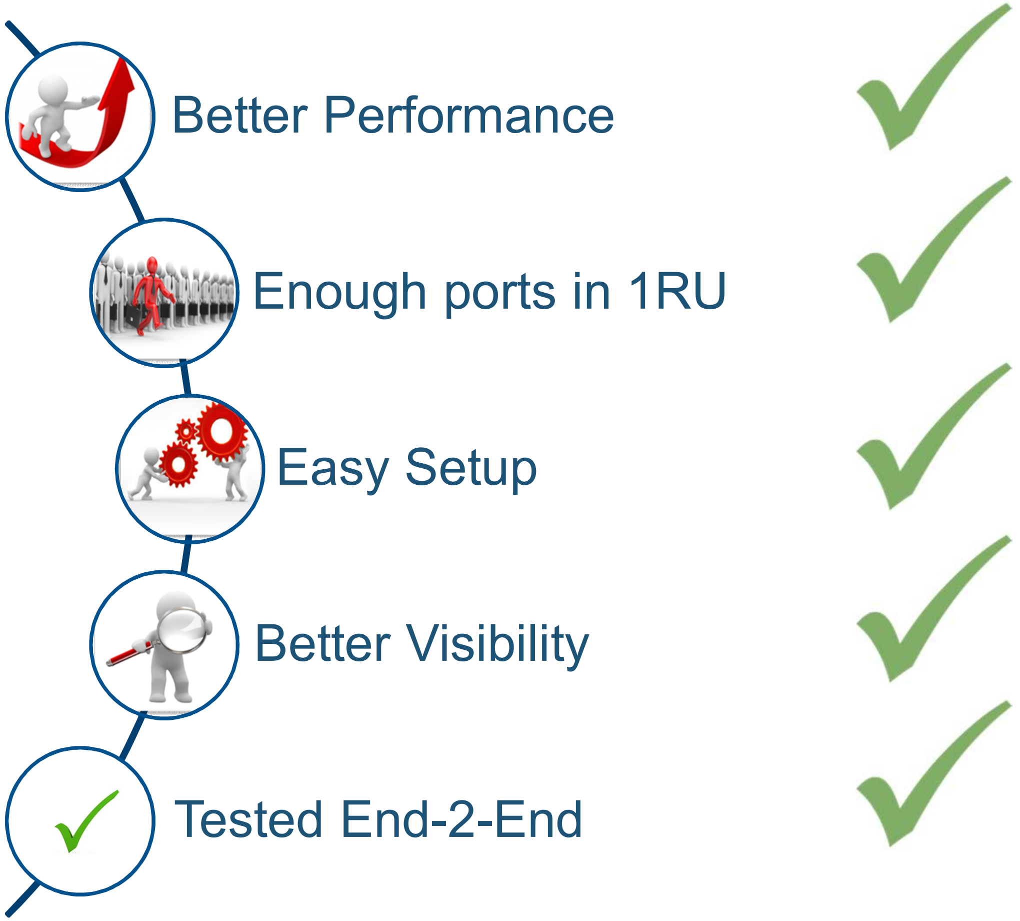
>	Timestamp	Switch IP	Interface	Source MAC	Destination MAC	Source IP
>	2018-10-22 12:40:10	10.213.53.8	Eth1/1	N/A	N/A	N/A
>	2018-10-22 12:40:10	10.213.53.4	Eth1/1	N/A	N/A	N/A
>	2018-10-22 12:40:10	10.213.53.11	Eth1/1	N/A	N/A	N/A
>	2018-10-22 12:40:10	10.213.53.4	Eth1/1	N/A	N/A	N/A

Summary: Mellanox ESF Switches

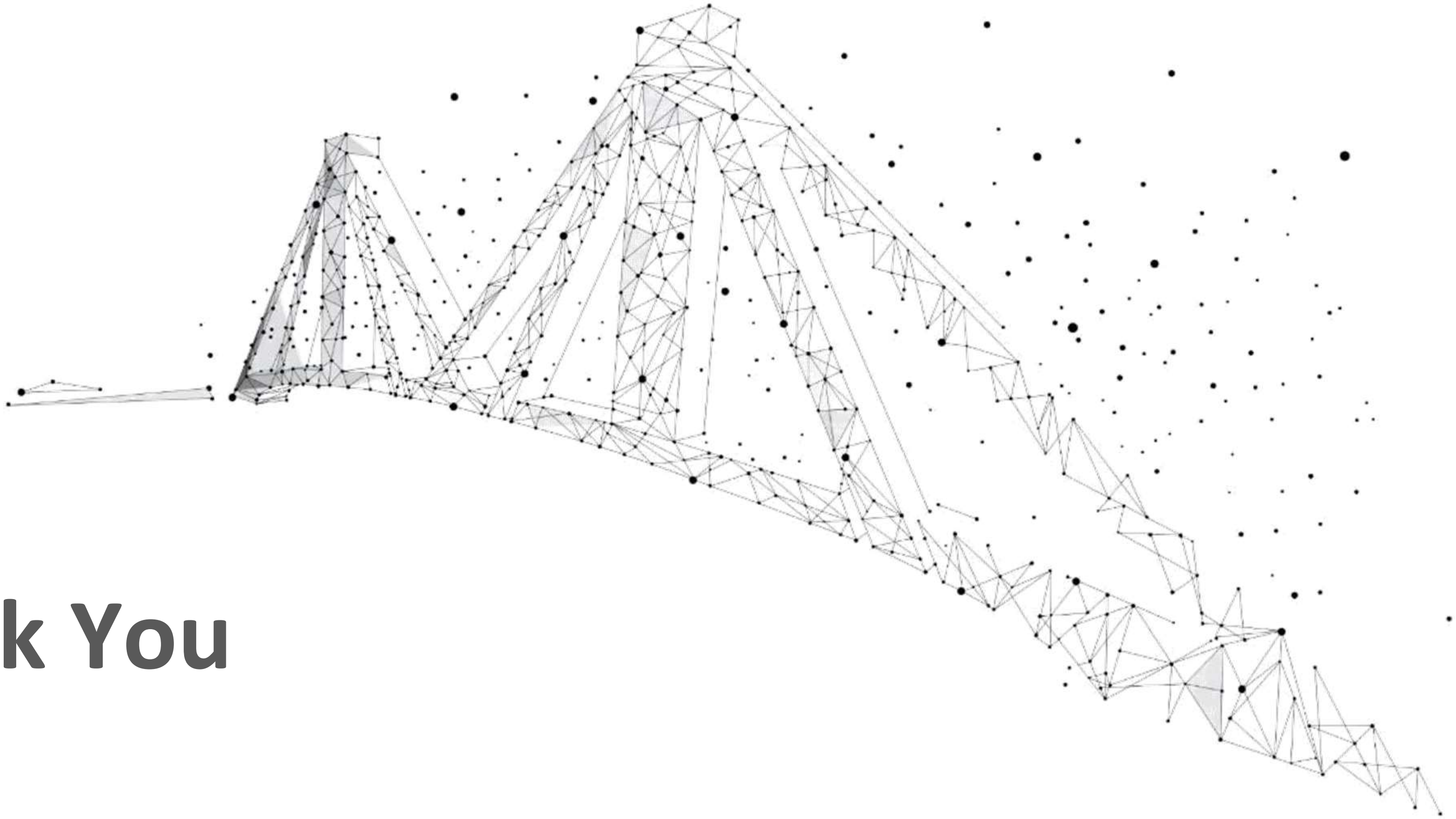








- Better Performance ✓
- Enough ports in 1RU ✓
- Easy Setup ✓
- Better Visibility ✓
- Tested End-2-End ✓



Thank You