CREDO

HiWire
Active Electrical Cables

STORDIS
The Open Networking Expert

Demystifying Dual TOR – Credo's SWITCH AEC and Upstream NOS Support

May 19, 2022

Łukasz Łukowski

Chief Sales and Marketing Officier

Don Barnetson

VP Product, HiWire AECs

Agenda

- TORs : Single Point of Failure for a Rack
- MLAG: A Legacy solution with scalability issues
- HiWire SWITCH AEC + NOS Management Container
- Failure Convergence Scenarios
- Current and Planned NOS Support
- SONiC Integration and Support
- Q&A Section - Conclusions and Questions

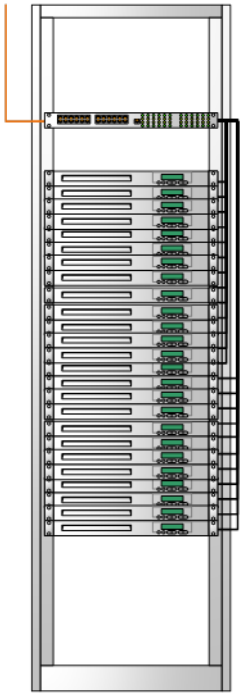
Credo Safe Harbor

We have made statements in this presentation that are forward-looking statements. In some cases, you can identify these statements by forward-looking words such as "may," "might," "will," "should," "expects," "plans," "anticipates," "believes," "estimates," "predicts," "potential" or "continue," the negative of these terms and other comparable terminology. These forward-looking statements, which are subject to risks, uncertainties and assumptions about us, may include projections of our future financial performance, our anticipated growth strategies and anticipated trends in our business and in the industry in which we operate. These statements are only predictions based on our current expectations and projections about future events. There are important factors that could cause our actual results, level of activity, performance or achievements to differ materially from the results, level of activity, performance or achievements expressed or implied by the forward-looking statements, including those factors discussed in our most recent annual report and quarterly reports as filed with the SEC, including (but not limited to) those discussed under the caption entitled "Risk Factors" in those filings.

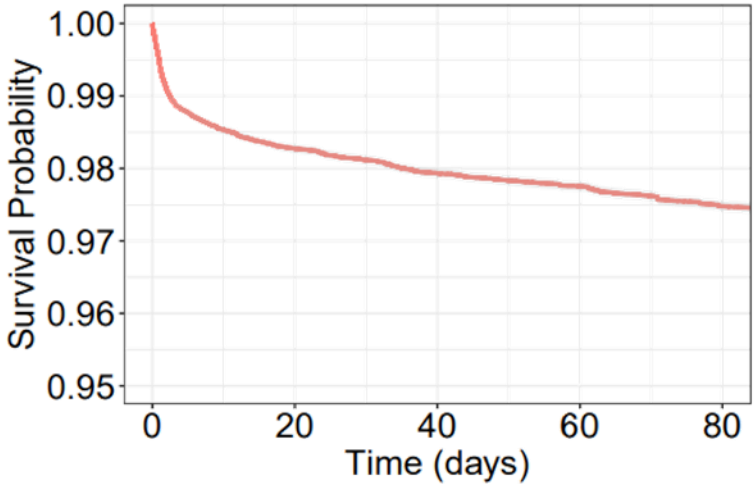
You should not rely upon forward-looking statements as predictions of future events. Although we believe that the expectations reflected in the forward-looking statements are reasonable, we cannot guarantee that the future results, levels of activity, performance or events and circumstances reflected in the forward-looking statements will be achieved or occur. These forward-looking statements speak only as of the date of our presentation. We undertake no obligation to update publicly any forward-looking statements for any reason after the date of this presentation to conform these statements to actual results or to changes in our expectations, except as required by applicable law.

You should read this presentation and the documents that we reference in this presentation with the understanding that our actual future results, levels of activity, performance and events and circumstances may be materially different from what we expect.

Top of Rack Switch : Single Point of Failure for a Rack

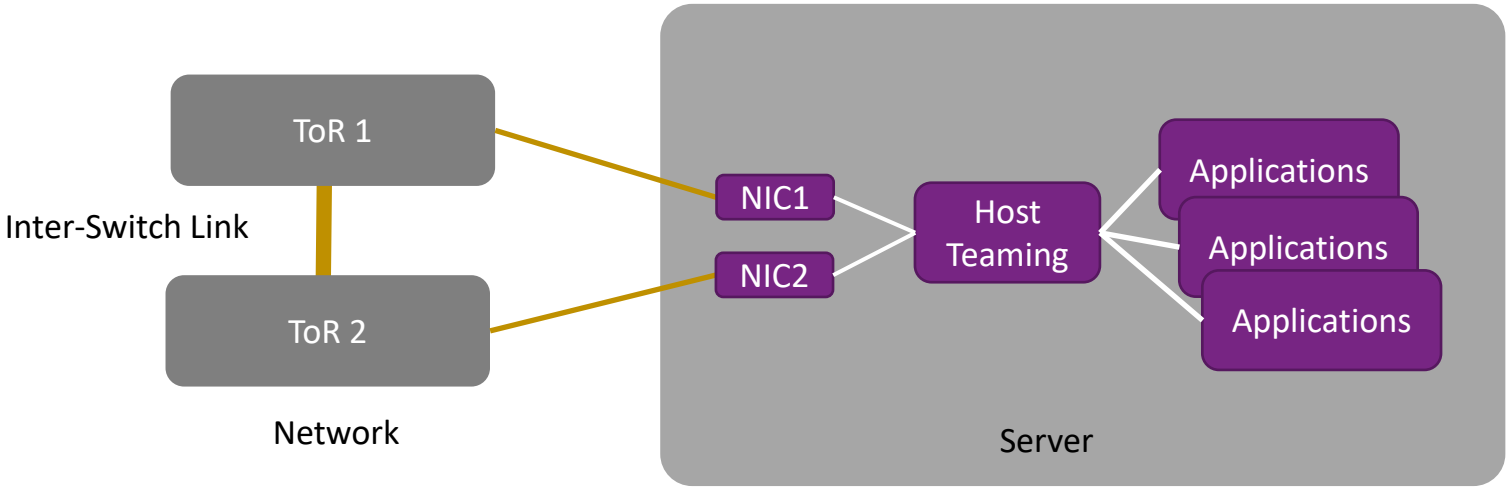


- ToR is a Single-Point-of-Failure (SPOF) for full rack of servers
- And TORs do fail
  - ~2% of switches fail in first 3 months
  - 32% due to hardware failures
  - 27% due to power failures

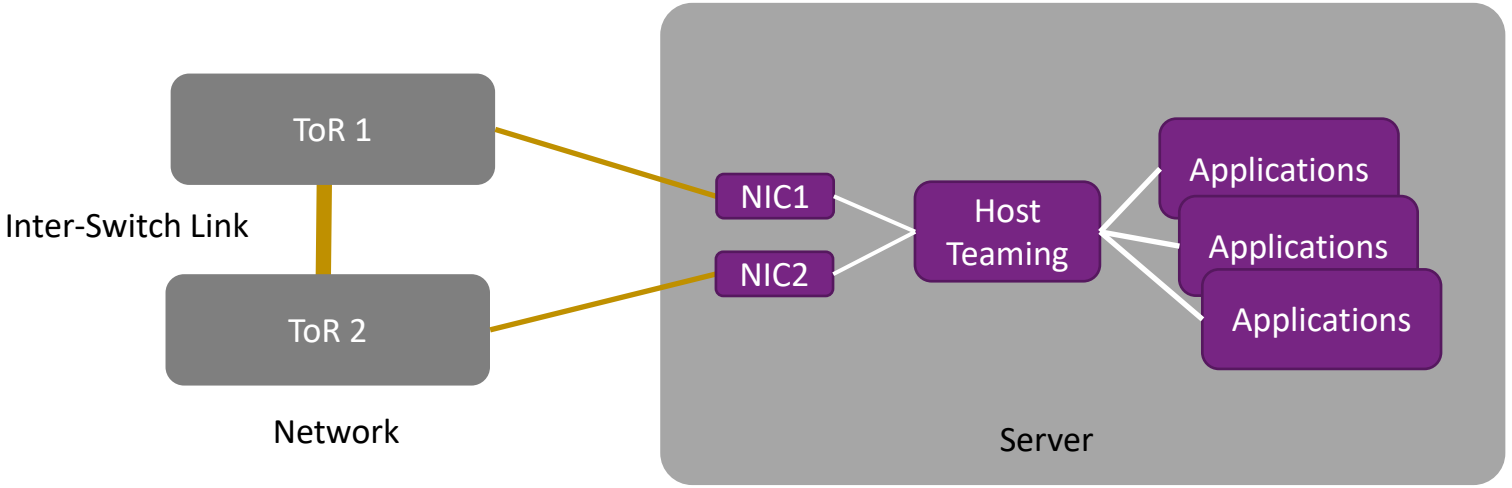


Source : MSFT Azure, April/21

Classic Solution: Multi-Chassis LAG with Dual Uplink

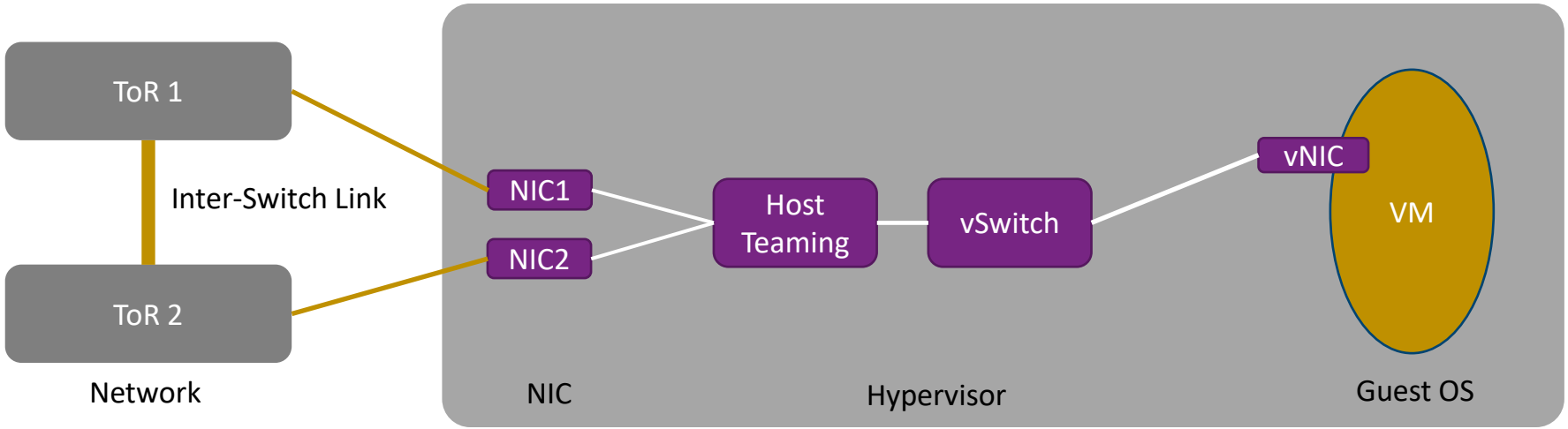


Network: Inter-Switch Link



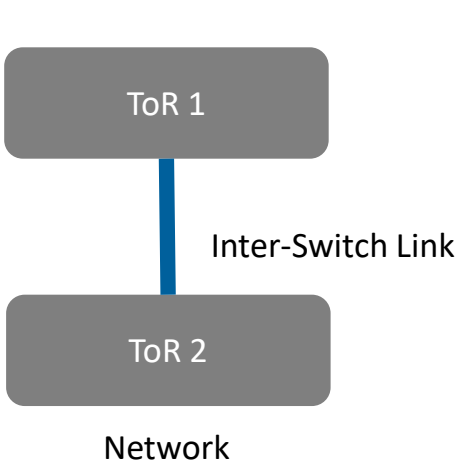
- Inter-Switch Link requires custom design for capacity planning
- Requires complex state sync between ToRs
  - Creates split-brain problem when ISL fails

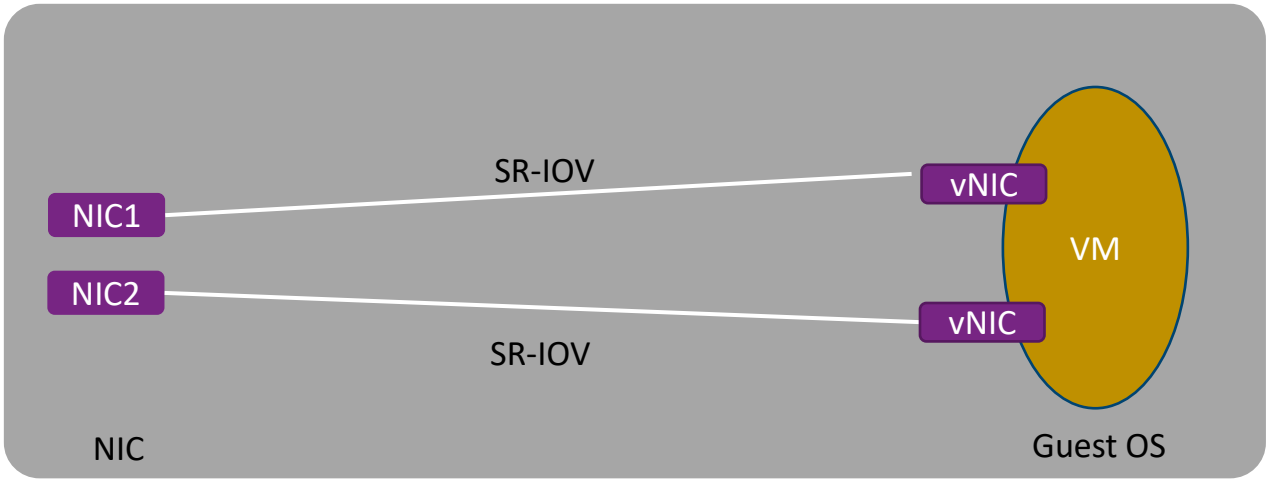
HyperVisor: Performance Limit



- A few Gbps/core
- Unstable latency
- Won't work for RDMA

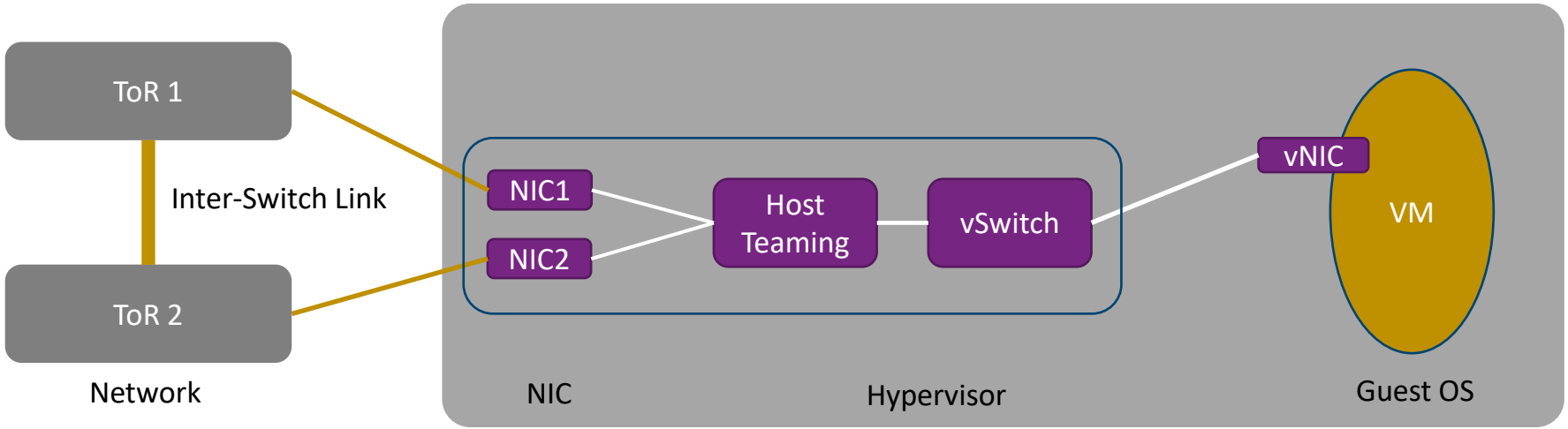
SR-IOV: VM Complexity





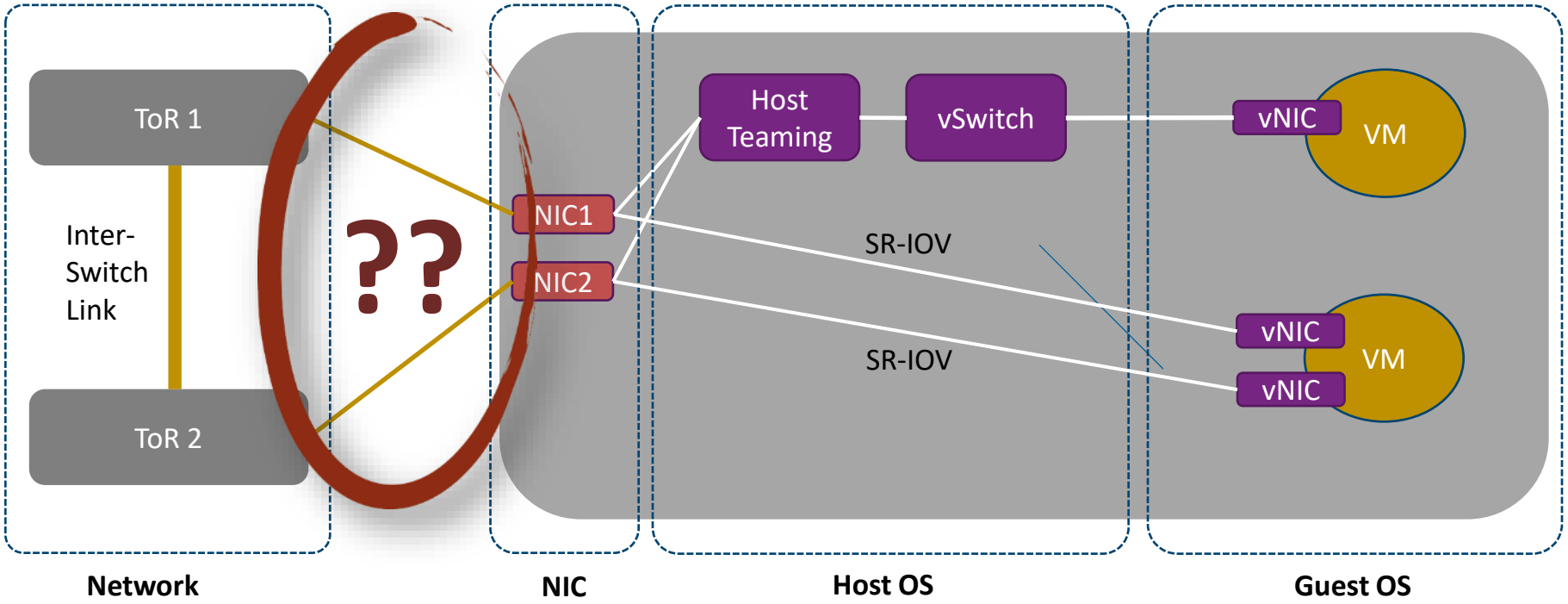
- VM sees failures and must handle
- Hardware-dependent vNIC driver

vSwitch Offloading: Vendor Specific

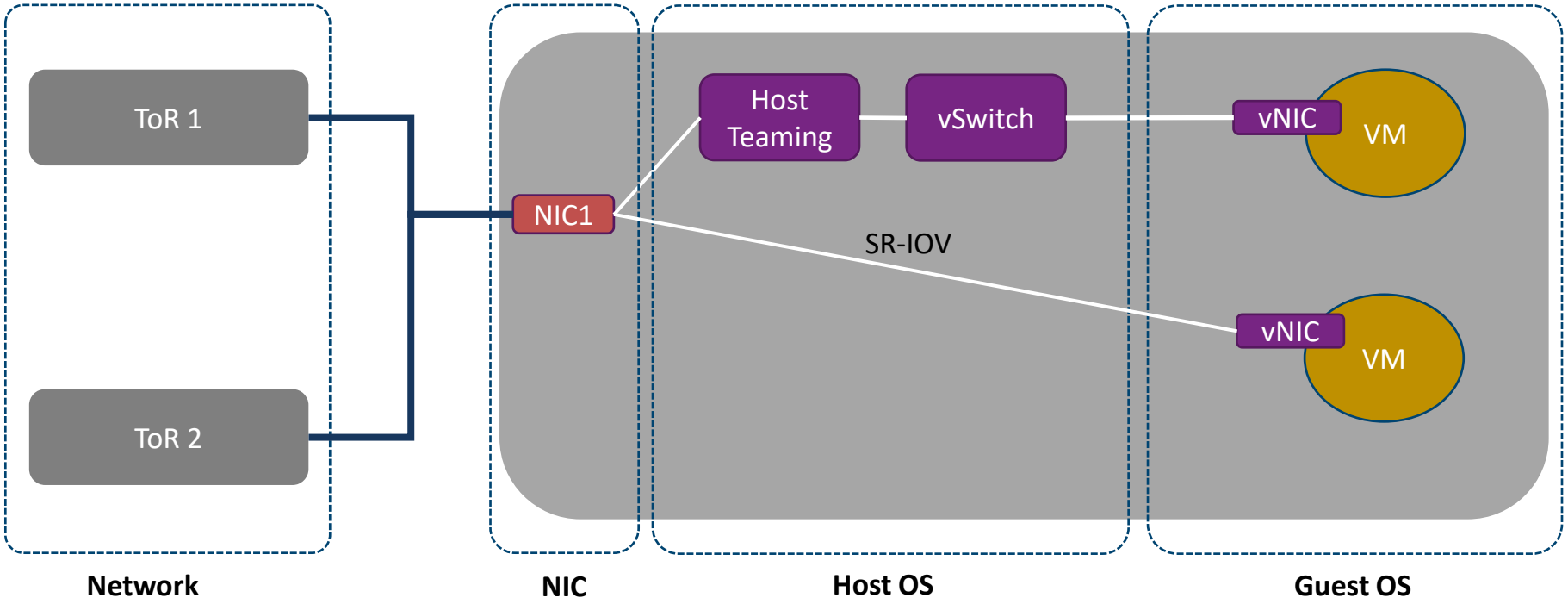


- Variety of solutions to offload Teaming and vSwitch to Smart NICs
- Capabilities and Implementations vary by NIC Vendor
- Require NIC vendor drivers in VM -> HW Dependency

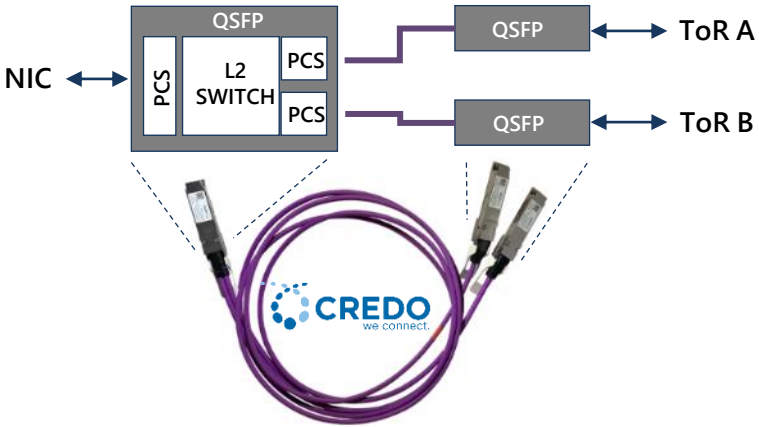
The Solution Space - Options to Manage Redundancy



The Solution Space – Introducing the Switch AEC



New Solution in Collaboration with Microsoft / Azure



Dual TOR Management Container

SONiC

Microsoft Azure

HiWire SWITCH AEC: 50G/100G/200G QSFP AEC Cable

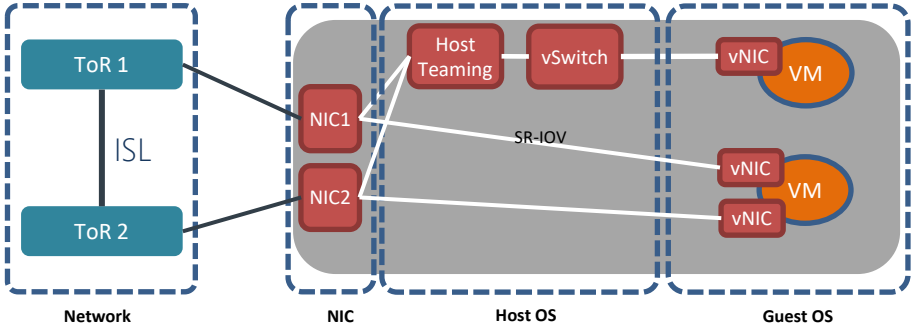
- Active/Standby Layer 2 switch – switching in <1 µs
- Common control plane on all 3 ends
- Fully ToR managed, works in standard QSFP NICs

Dual TOR Management Container on SONiC

- Manages the SWITCH AEC
- Manages convergence in failover conditions using standards based ARP/BGP/Encap & forward
- Free and Open Source

DualTOR Connectivity: Azure Switch Cable

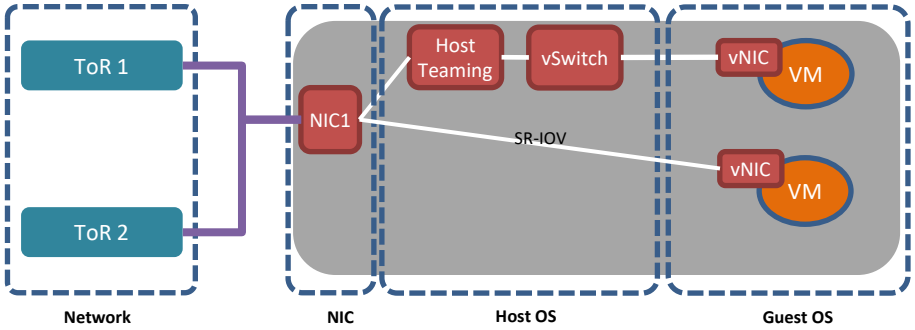
Legacy LAG Solution



Legacy LAG / MLAG Issues

- Capacity Planning on ISL
- Split Brain on TORs
- vSwitch – a Few Gbps/core; no RDMA; Jitter
- SR-IOV – Kicks problem to Guest OS
- Failure convergence time in seconds

Credo / Microsoft HiWire Switch AEC Solution



HiWire Switch / SONiC Solution

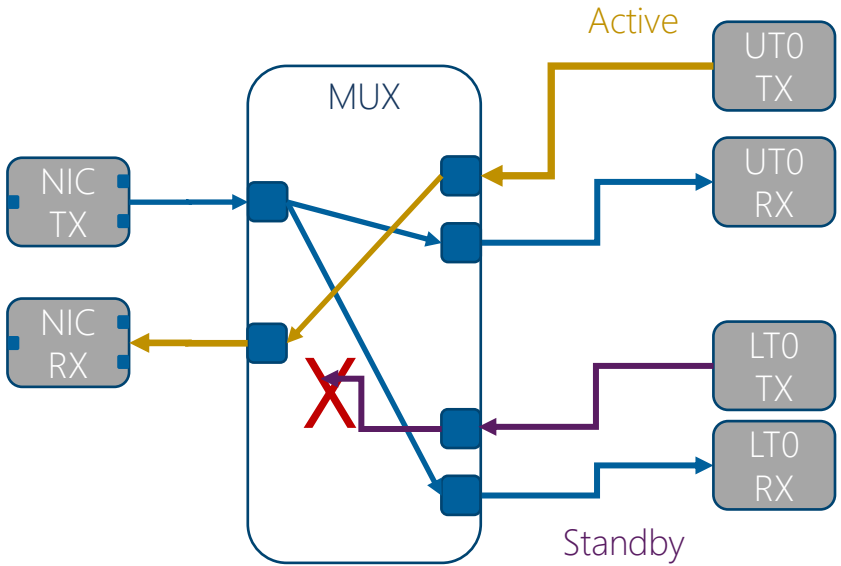
- No ISL needed
- State is held in cable / no split brain
- No server involvement at all; works with any HW/SW
- Failure Convergence in milliseconds

How do we deploy?

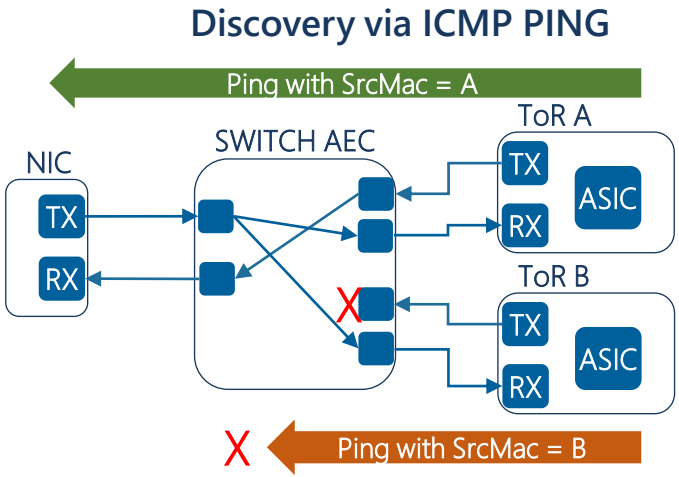
Switch AEC Data Plane Behavior

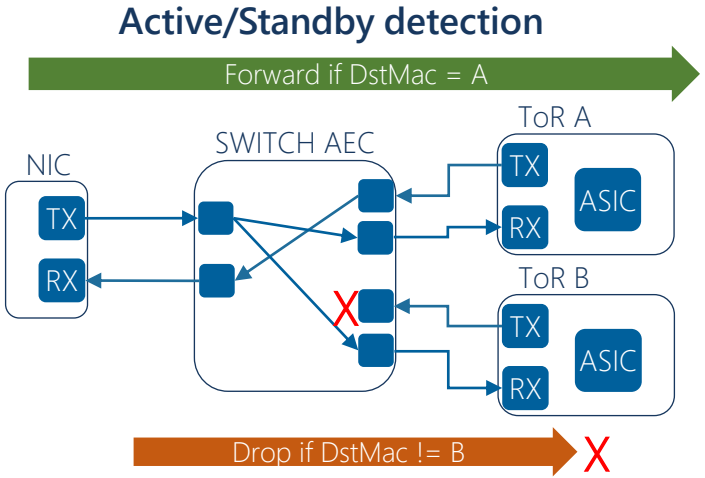
NIC TX is always broadcasted

NIC RX will take active side and only Switch with Loss-of-signal or on-command over time < 1ms



Independent Active/Standby Design





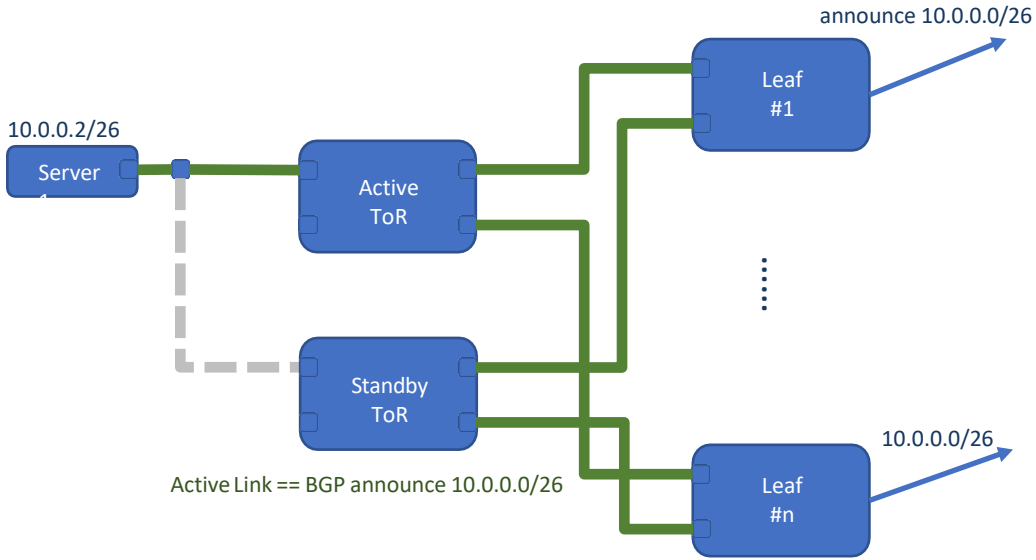
Southbound Traffic is MUX'd

- Both ToRs both ping NIC,
- ToR A ping is forwarded to NIC
- ToR B ping is dropped by SWITCH AEC
- NIC learns ToR A as destination MAC

Northbound traffic broadcasted to both ToRs

- ToR A:
  - Forward north based on DstMac correctness
  - Verifies link integrity
- ToR B:
  - Drop packets due to wrong DstMac
  - Sniffs to verify ToR A's link integrity

Routing Behavior – Both ToRs announce via BGP

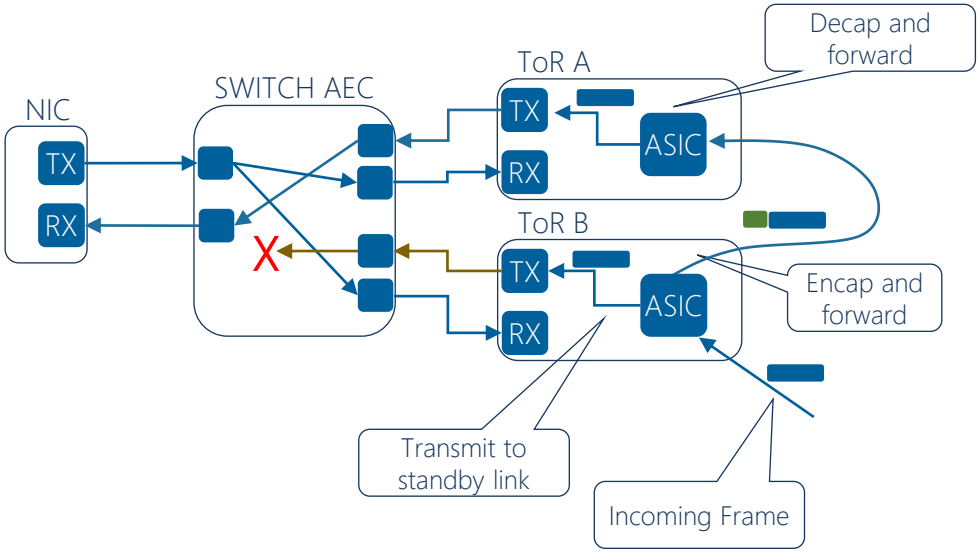


Both ToRs see Server (10.0.0.2/26) and announce to Leaf Nodes

All Leaf nodes see ToR announcements and announce to high level via standard BGP

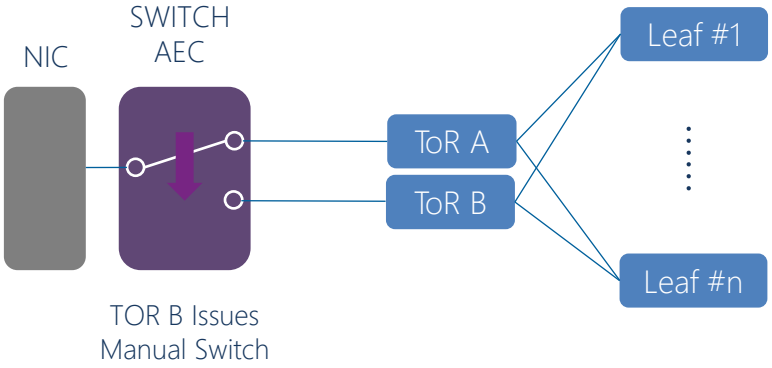
Southbound traffic can arrive at either ToR

Southbound Packet Path



- Southbound traffic arriving on active ToR A is forwarded to NIC
- Southbound traffic arriving on Standby ToR B
  - ToR B encaps and transmits to ToR A
  - Tunneled through leaf instead of ISL
  - ToR A decaps and forwards to NIC
- Encap/decap handled by ASIC with minimum overhead

Failure Scenarios 1: Planned Maintenance



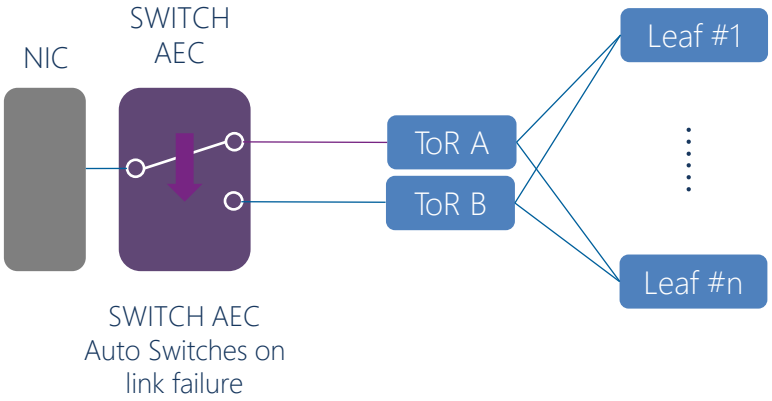
Scenarios:

ToR OS upgrade, hardware replacement

1. ToR B proactively issues command to switch MUX to ToR B
2. ToR B ICMP Ping is now forwarded; ToR A is blocked
3. NIC updates DstMaC to ToR B
4. ToR B becomes active, ToR A become standby

Convergence time < 100ms

Failure Scenarios 2: Hardware Failure for ToR and Link

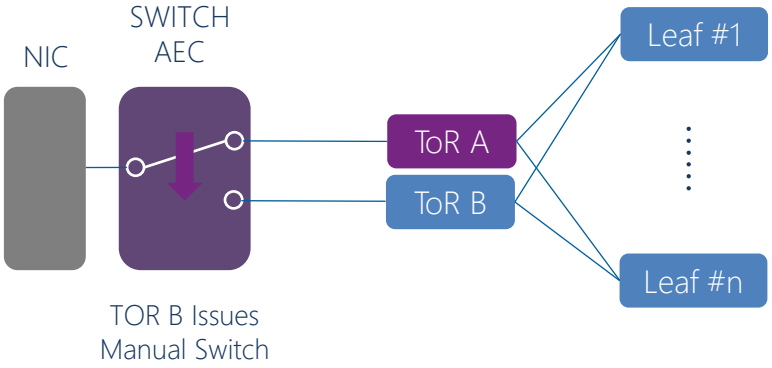


Deployment Scenario:
Cable cut, ToR port failure, ToR power failure, ToR

1. Cable detect Loss-of-signal to ToR A
2. Cable auto switches on link failure in less than 1µs
3. ToR B ICMP is now forwarded; ToR A is blocked
4. NIC changes DstMac address
5. ToR assumes Active role

Convergence time < 100ms

Failure Scenarios 3: ToR Forwarding Failure



Deployment Scenario:

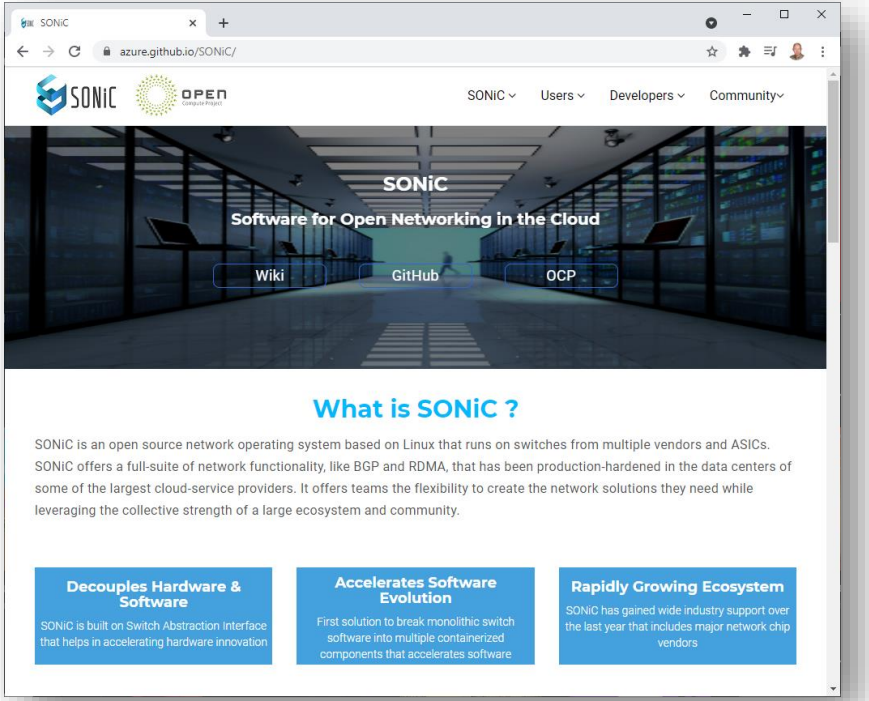
ToR grey failure impacts forwarding, but link remained up. Actually ~26 scenarios here – review SONiC Dual TOR Container docs

1. ToR B times out on sniffed ToR A Pings
2. ToR B Initiates Manual Switch of MUX
3. ToR B ICMP is now forwarded; ToR A is blocked
4. NIC changes DstMac address
5. ToR assumes Active role

Convergence in <100ms

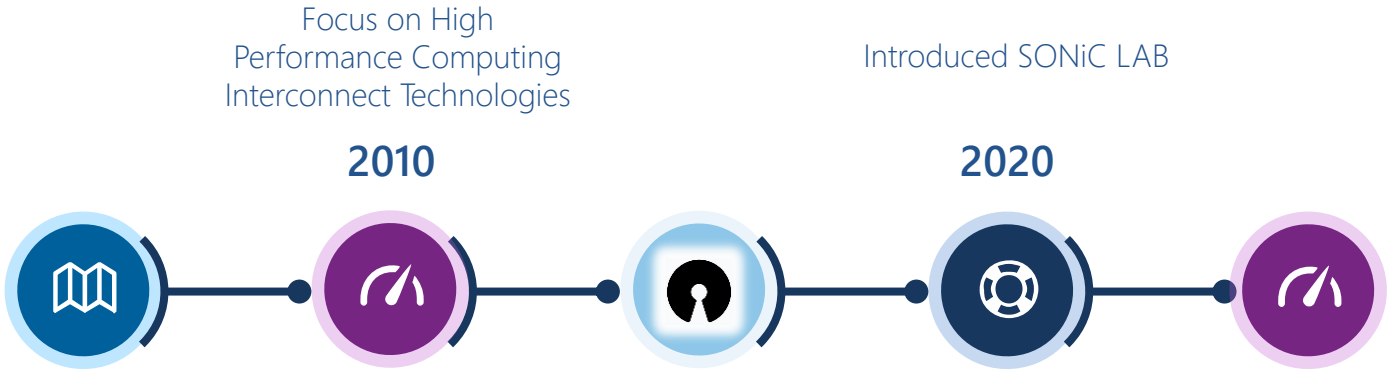
NOS Support for HiWire SWITCH AEC

- 50G and 100G HiWire™ SWITCH AEC GA now! 200G SWITCH AEC samples available new
- SWITCH AEC fully upstreamed and abstracted in SONiC Nov/21 release
  - Support for any SONiC hardware
  - Hitless firmware updates from any cable end
  - Advanced Cable telemetry, loopback, BER and debug
  - Implementation reference in development for June release
- Arrcus OS Support available now.
- Cumulus / NVIDIA Spectrum support in development.



Our Way2Open

STORDIS Mission to Deliver Open Solutions



2007

Founded as specialized Distributor

2010

Focus on High Performance Computing Interconnect Technologies

2014

First to Introduce Open Networking in EU

2020

Introduced SONiC LAB

2021 / 2022

SONiC Container Development, Tool Development, Introduced Open2Support

Our Target

We want to deliver SONiC

STORDIS
The Open Networking Expert

CREDO

Enterprise SONiC

Delivering the most advanced NOS

SONiC Services

Delivering Consulting, Tool Development, Support and Training Services


Q&A
STORDIS
The Open Networking Expert
CREDO
we connect.

*"**Freedom** is nothing else but a chance to be better"* —
Albert Camus

www.credosemi.com I www.stordis.com I www.stordirect.com I www.route2open.com