# SRv6 uSID

## Deterministic AI Backend Networking and Unified DC Frontend Design

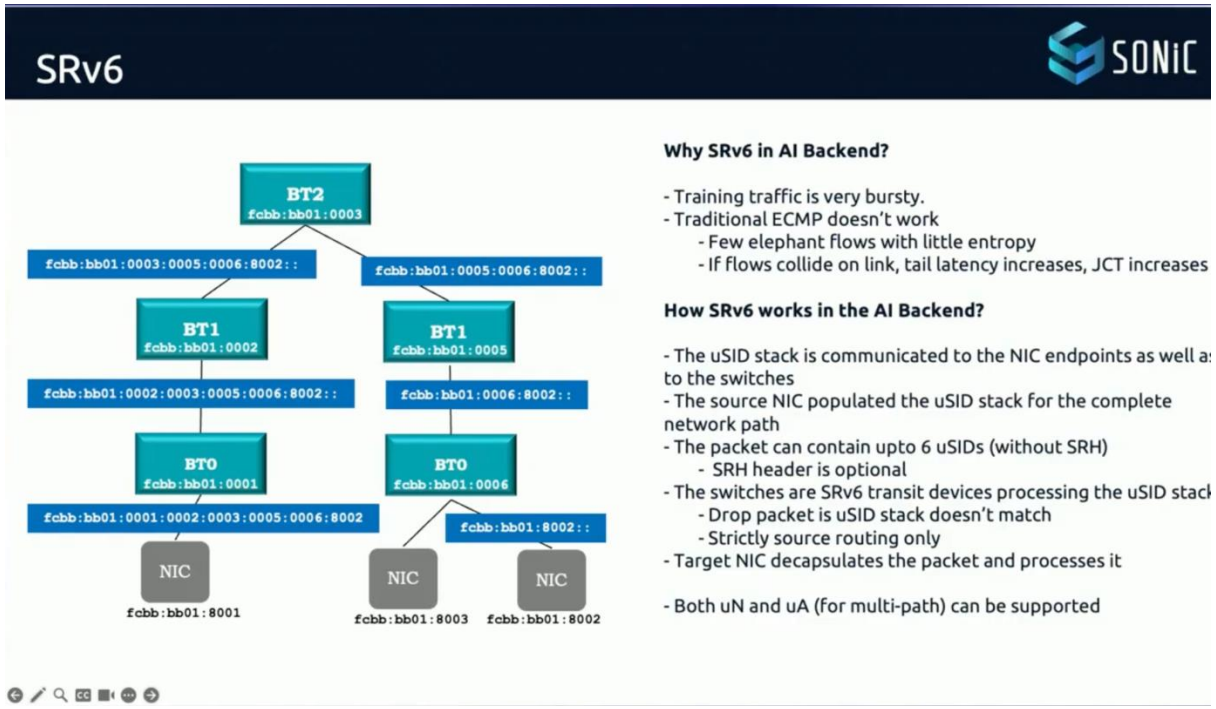
Ahmed Abdelsalam

Cisco Systems

# SRv6 — Powering the World’s Most Powerful AI Data Center

- Sept 2025: Microsoft announced Fairwater DC in Wisconsin. The world’s Most Powerful AI Data Center.
- House hundreds of thousands of GPUs, operating in seamless clusters.
- Microsoft Blog [Link]

# SRv6 — Powering the World's Most Powerful AI Data Center




- Sept 2025: Microsoft announced Fairwater DC in Wisconsin. The world's Most Powerful AI Data Center.
- House hundreds of thousands of GPUs, operating in seamless clusters.
- Microsoft Blog [Link]

- Oct 2025: Microsoft explained the key enabling technology for the Fairwater DC. SRv6!!
- SRv6 used for deterministic placement of AI flows in the fabric to achieve optimal load-balancing.
- Guohan Lu, Microsoft. OCP Global Summit [Link]

# SRv6 uSID

# IP is the Heart of our Digital Economy

# IP Best-Effort is > 40-year-old

# IP - Best-Effort for > 40 years

- Destination Address only localize an endpoint
- Development of per-domain shim layers to enrich IP
  - MPLS in the core
  - VxLAN in the DC
  - GTP for mobility
- Low scale
  - MPLS cannot summarize => BGP used as an IGP
- End-to-End Services challenges and High Operation Cost
  - Neither TE nor stateless service chaining in VxLAN
  - Translation gateways between domain (VxLAN/MPLS)

# Impossible to change?

# SRv6 uSID - Fundamental transformation of IP

# SRv6 uSID

- IPv6 becomes Programmable
- 128-bit DA becomes:
  - ordered list of uSID instructions (uSID)
  - uSID instruction can bound to anything .. Underlay TE, overlay, service chaining
  - 6 uSID instructions in DA
  - Most deployments requires less than 6 uSID
  - SRH is there for ultra-scale TE
- SRv6 uSID outperforms any IPv4 shim layer: MPLS, VxLAN, GTP

DA = FCBB:BBBB:0000:0000:0000:0000:0000:0000

IPv6 Block | uSID1 | uSID2 | uSID3 | uSID4 | uSID5 | uSID6

# SRv6 uSID - Fully Standardized

**Architecture**

- SR Architecture – RFC 8402
- SRTE Policy Architecture – RFC 9256

**Data Plane**

- SRv6 Network Programming – RFC 8986
- IPv6 SR header – RFC 8754
- SRv6 uSID – RFC9800

**Control Plane**

- SRv6 BGP Services – RFC 9252
- SRv6 ISIS – RFC 9352
- SR Flex-Algo – RFC 9350
- BGP-LS – RFC 9514
- PCEP – RFC 9603

**Operation & Management**

- SRv6 OAM – RFC 9259
- Performance Management – RFC 5357
- Performance Management – RFC 9503
- SRv6 SID Block – RFC 9602

# SRv6 uSID - Rich Ecosystem

## Network Equipment Manufacturers

Cisco, Huawei, Nokia, Ciena, Juniper Networks, H3C, ZTE, Arrcus, Arista, Ericsson

## Merchant Silicon

Access: A100 Edge: K100
Core: Q200, P100
ToR/Spine: G100, G200

Broadcom

DNX – Jericho
XGS – Tomahawk

Marvell, Nvidia

## Open-Source Applications

Pyroute2, IPTABLES SERA, GoBGP, exa, BGP, eBPF, cilium, Wireshark, FRRouting, Snort, GoBMP

SRv6 uSID

## Open-Source Networking Stacks

SONiC, Linux, fd.io, P4, Open Compute Project, Kubernetes, OvS Open vSwitch, VyOS

## Smart NIC / DPU

Intel, Nvidia

## Partners

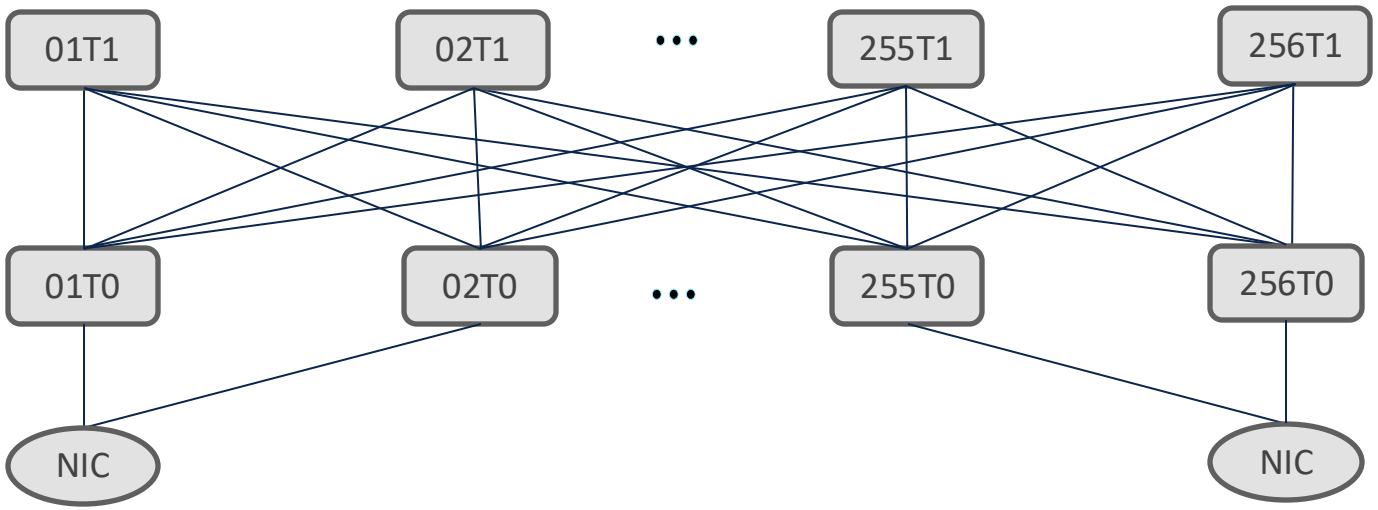
kaloom, Trend Micro, ENEA Qosmos Division, NoviFlow, ixia A Keysight Business, spirent Promise. Assured., Red Hat, f5, 6WIND, Keysight Technologies

# SRv6 uSID - Deterministic AI Backend Networking

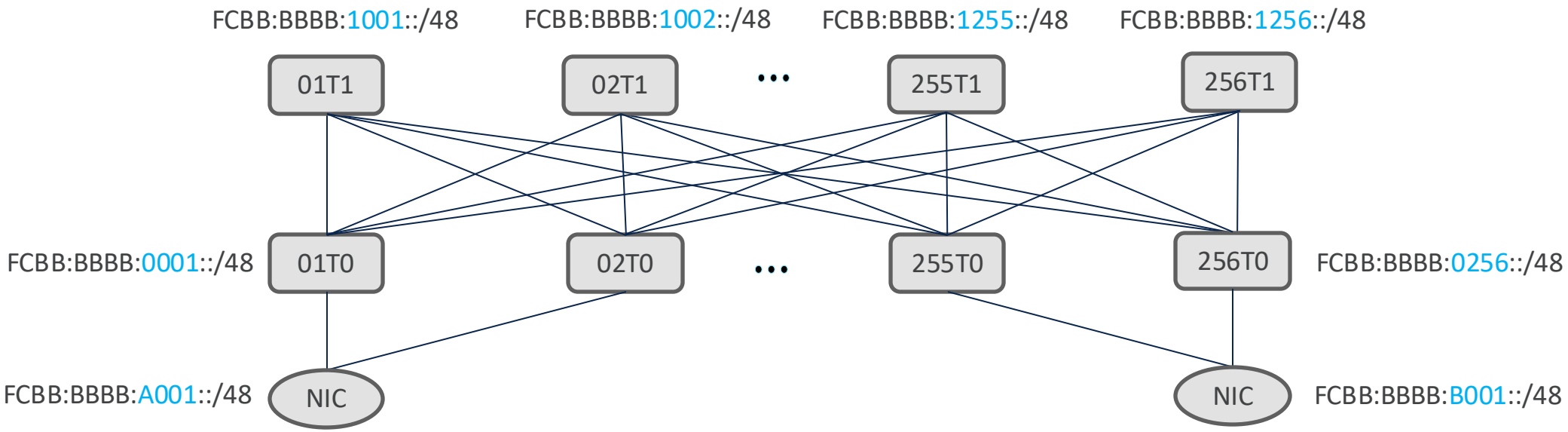
# Problem Statement & Why SRv6

- AI backend
  - the training traffic is very bursty and each of the flow is elephant flow.
  - Traditional ECMP does not work for the AI backend.
  - Particularly bad because the training job is highly synchronized.
  - If flows collide on a link and it gets congested it drags the whole job and delay the whole job significantly.
- SRv6
  - Provides fined-grained network control based on source routing
  - Enables path enumeration for traffic management
  - Integration with AI workloads flow scheduling provides optimal network performance
  - Allow source to quickly reroute upon path failures or congestion
  - Standard, Open, vendor-agnostic

# SRv6 uSID – Deterministic Path Placement

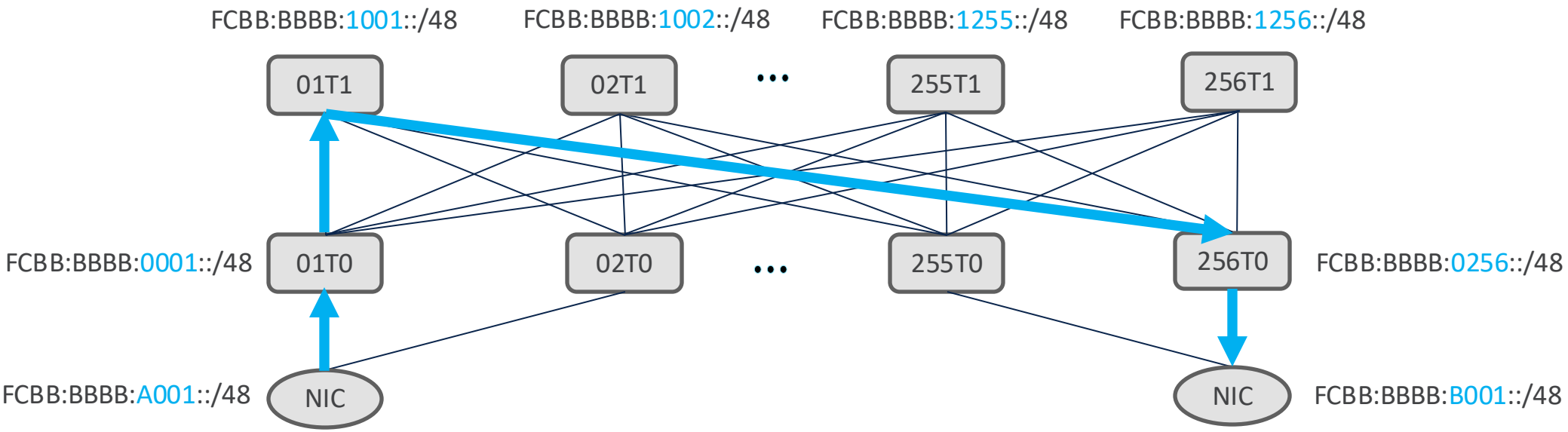


# SRv6 uSID – Deterministic Path Placement



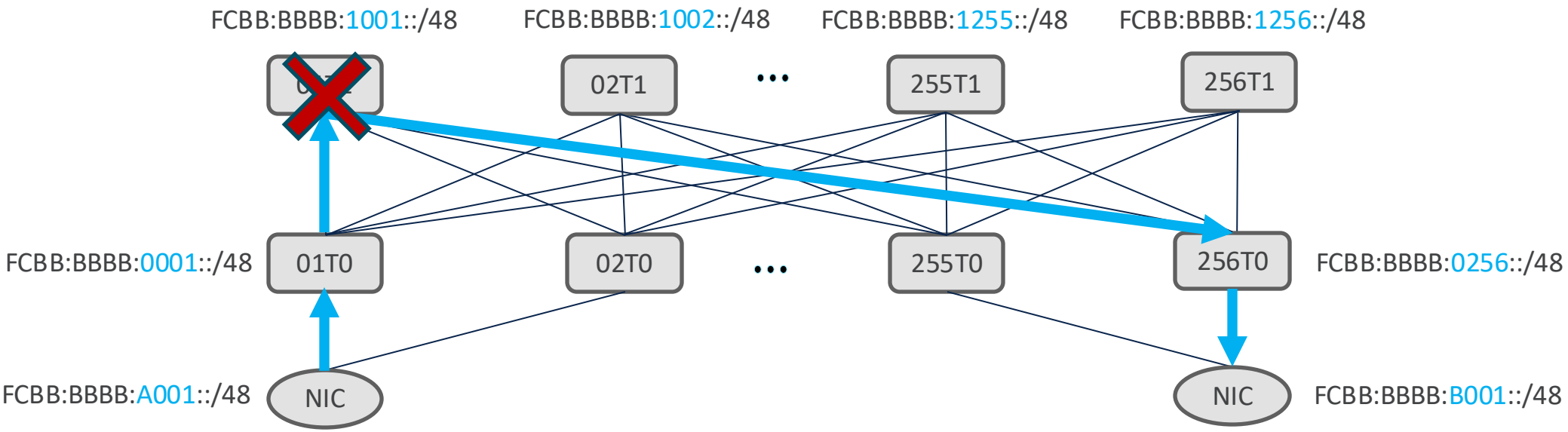

- SRv6 Block - FCBB:BBBB::/32
- SRv6 Locator - FCBB:BBBB:NNNN::/48
  - 01T0 => FCBB:BBBB:0001::/48

# SRv6 uSID – Deterministic Path Placement



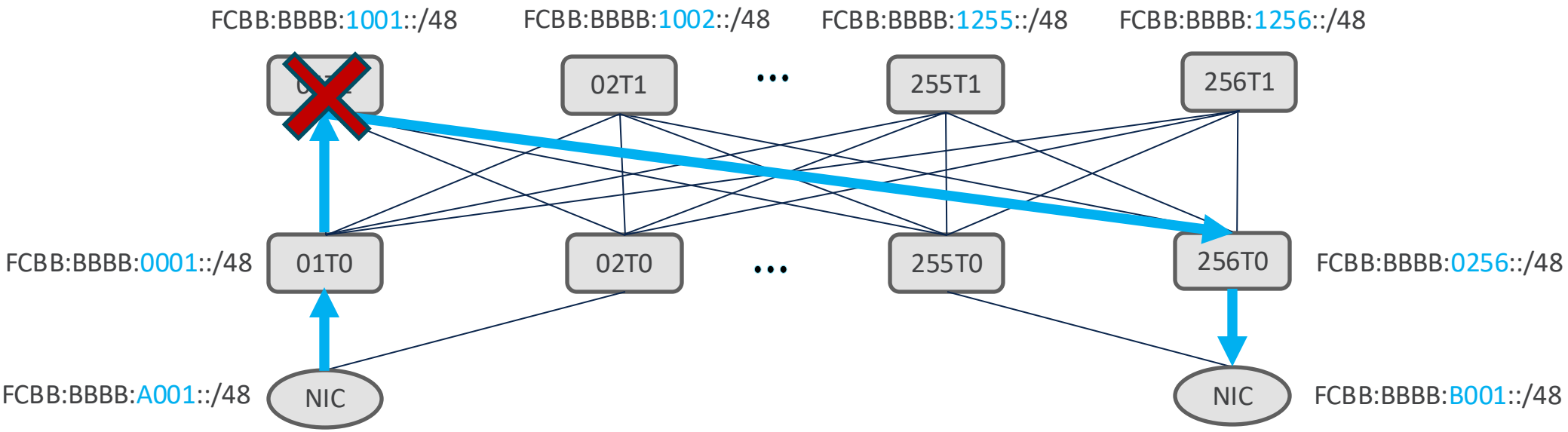

- NIC A to NIC B
  - Via FCBB:BBBB:0001:1001:0256:B001::
  - Go to 01T0 then 01T1 then 256T0 then NIC B

# SRv6 uSID – Deterministic Path Placement




- NIC A to NIC B
  - Via FCBB:BBBB:0001:1001:0256:B001::
  - Go to 01T0 then 01T1 then 256T0 then NIC B

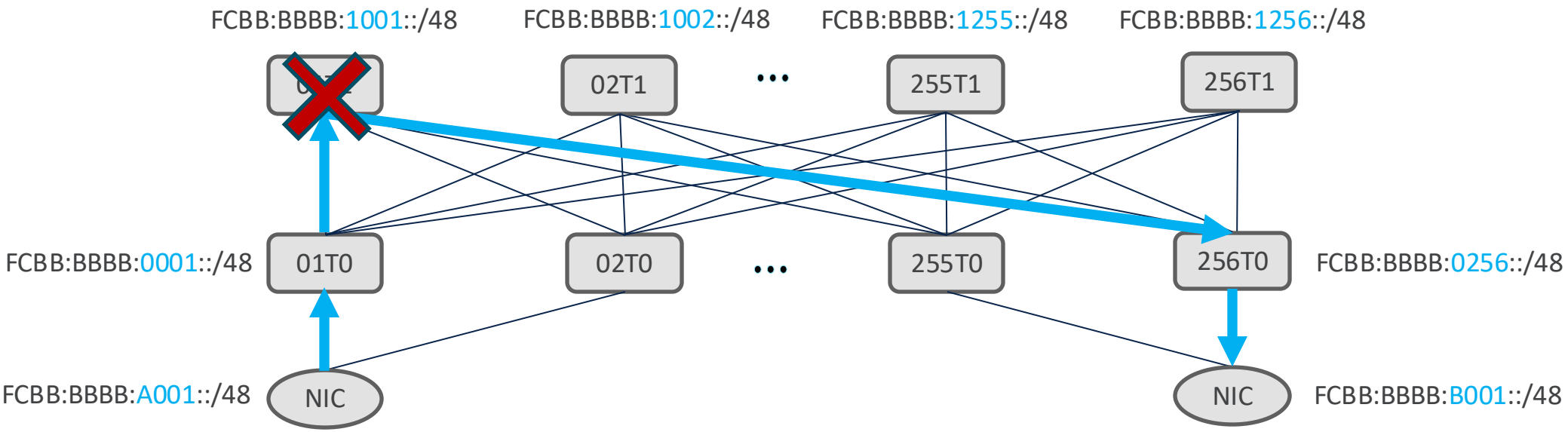
# SRv6 uSID – Deterministic Path Placement




- NIC A to NIC B
  - Via FCBB:BBBB:0001:1001:0256:B001::
  - Go to 01T0 then 01T1 then 256T0 then NIC B

Congestion/Failure " 01T1"

# SRv6 uSID – Deterministic Path Placement

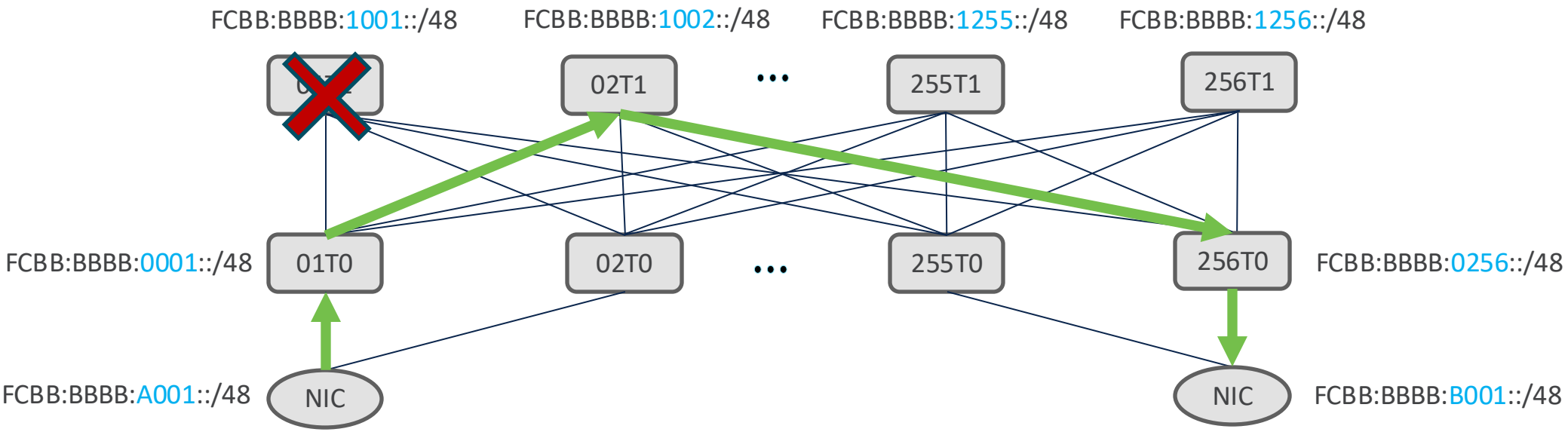



- NIC A to NIC B
  - Via FCBB:BBBB:0001:1001:0256:B001::
  - Go to 01T0 then 01T1 then 256T0 then NIC B

Congestion/Failure " 01T1"

- NIC A to NIC B
  - via FCBB:BBBB:0001:1002:0256:B001::
  - Go to 01T0 then 01T1 then 256T0 then NIC B

# SRv6 uSID – Deterministic Path Placement




- NIC A to NIC B
  - via FCBB:BBBB:0001:1001:0256:B001::
  - Go to 01T0 then 01T1 then 256T0 then NIC B

Congestion/Failure " 01T1"

- NIC A to NIC B
  - via FCBB:BBBB:0001:1002:0256:B001::
  - Go to 01T0 then 01T1 then 256T0 then NIC B

# SRv6 uSID for AI Backed


Workgroup: SRv6 Operations
Published: 18 August 2025
Intended Status: Informational
Expires: 19 February 2026

C. Filsfils
Cisco Systems
C. Martin
Oracle Cloud
K. Pillai
IBM
P. Camarillo, Ed.
Cisco Systems
A. Abdelsalam
Cisco Systems
J. Tantsura
NVIDIA
K. Patel
Arrcus, Inc.

**SRv6 for Deterministic Path Placement in AI Backends**

**Abstract**

This document describes the use of SRv6 to enable deterministic path placement in AI backends, optimizing load balancing and congestion control for predictable GPU workloads.

Cisco, Oracle Cloud, IBM, NVIDIA, Arrcus

**Benefits**

* **Deterministic Path Placement**: SRv6 allows the NIC to control the path of each flow through the fabric.

* **Minimum-MTU**: A plain outer IPv6 encapsulation allows to encode 6 uSIDs in the outer DA. This implies that without the need of additional extension headers, only with 40Bytes of IPv6 encapsulation, we can encode up to 6 intermediate waypoints allowing to enforce a path in a 3-tier Clos network. This is sufficient to control a path hop-by-hop (link by link) through a leaf, spine, super-spine, spine, leaf.

* **Congestion Feedback Loop**: Instant rerouting at the source based on ECN, in-band measured One-Way and Two-Way latency, Packet Trimming feedback and in-band packet loss, without any dependency of routing protocols. There is neither any control-plane signaling involved between the GPU and the fabric, nor between the AI orchestrator and the fabric devices.

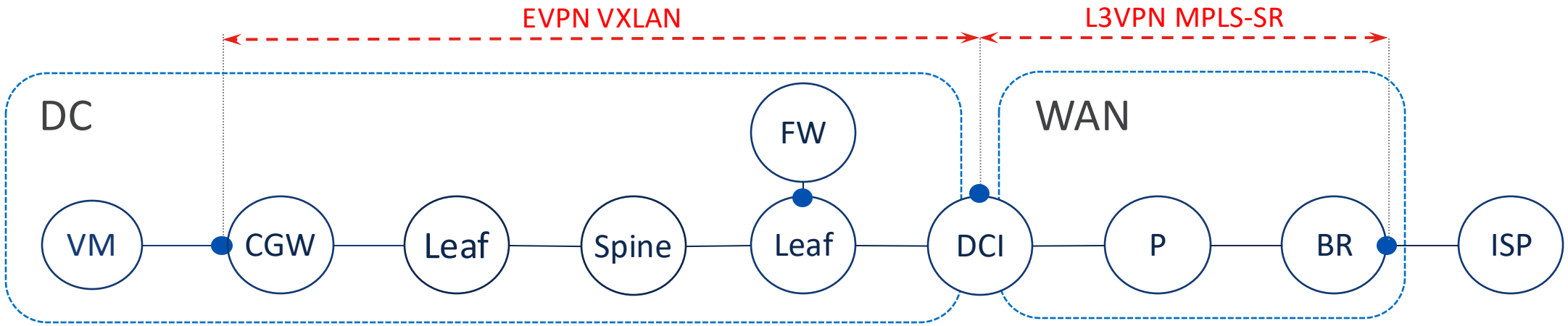
* **Standardization**: Open, vendor-agnostic implementation

* **Ease of operation**: as opposed to black-box proprietary solution which packs opaque layer-2 optimization, the SRv6 solution is minimalistic, IP based, fully standardized and a rich ecosystem (vendor, merchant and open source). The deterministic and open nature of the solution simplifies troubleshooting.

https://datatracker.ietf.org/doc/draft-filsfils-srv6ops-srv6-ai-backend/

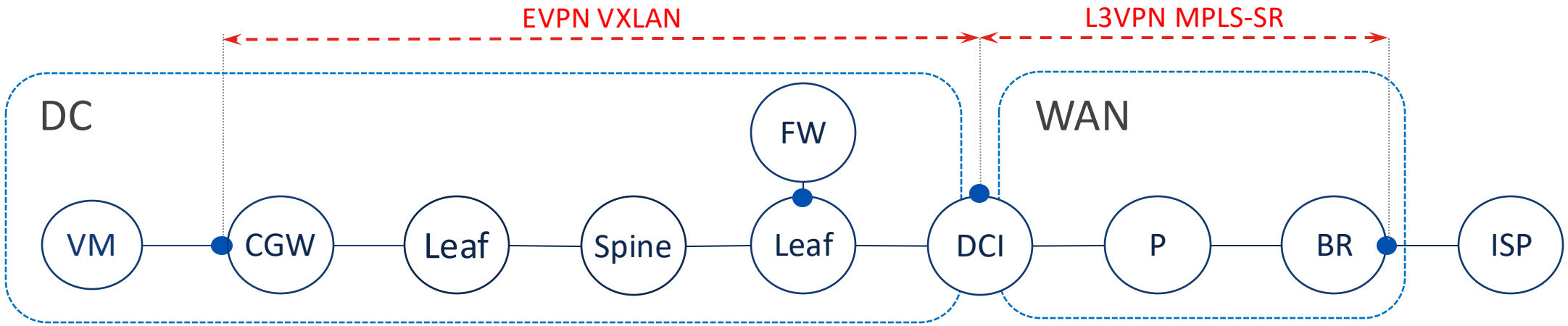
# SRv6 uSID - Unified DC Frontend Design

# Legacy architecture of the frontend network



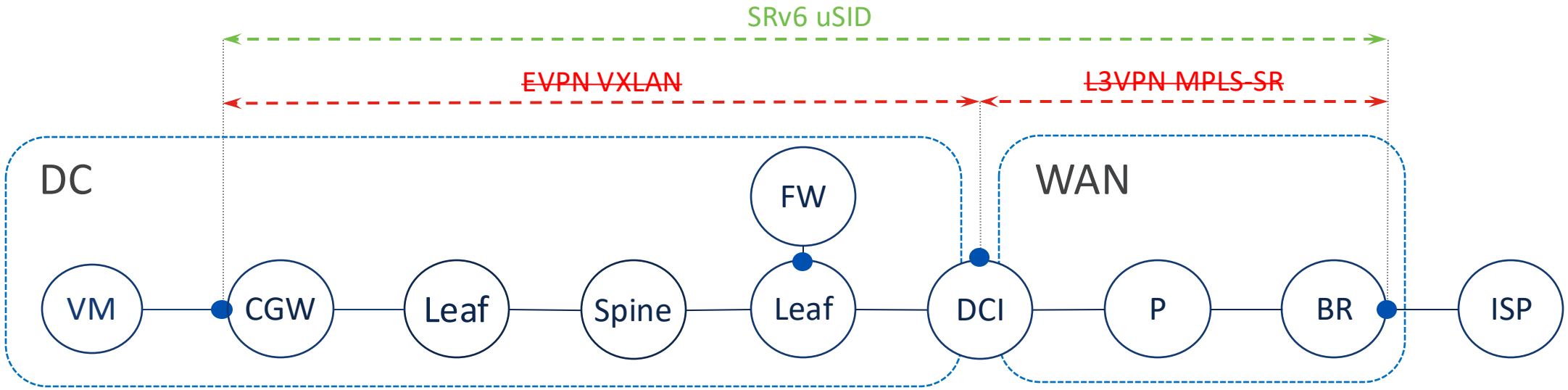

- IPv6 only infrastructure in the Data Center
- CGW (Cloud Gateway) and FW are NFV's running on hosts.
- VXLAN based overlay between CGW and DCI
- DCI does "stitching" between EVPN VXLAN and L3VPN MPLS-SR

# Legacy architecture of the frontend network – Limitation




- VXLAN: No traffic engineering capabilities inside the Data Center
  - Service chaining with VXLAN requires specific routing design (PBR, Default GW, VRF/VLAN hand-off, etc.)
  - Majority vendor implementations of VXLAN still require IPv4 loopbacks in the Underlay
- SR-MPLS: Lack native Data Center optimisations and not applicable in the DC domain
- Requires ”stitching” gateway functionality at the DCI routers to interconnect WAN and DC domains

# SRv6 in the Front-End DC




- SRv6 uSID
  - Unified End-to-End design from the DC Front End, through the metro, up to the Peering
  - Replaces the legacy design with VXLAN in the DC and SR-MPLS in the metro network.
  - Firewall Service insertion – stateless
  - Native TE

Alexey Gorovoy, Nebius
MPLS & SRv6 World Congress. Paris, March 2025 [Link]

# SRv6 uSID for AI Backed – SRv6 Benefits


Workgroup: SRv6 Operations
Internet-Draft:
draft-filsfils-srv6ops-srv6-e2e-dc-frontend-wan-00
Published: 4 July 2025
Intended Status: Informational
Expires: 5 January 2026

C. Filsfils
Cisco Systems
P. Camarillo, Ed.
Cisco Systems
K. Michielsen
Cisco Systems
A. Gorovoy
Nebius

**SRv6 End-to-End DC Frontend and WAN**

**Abstract**

The SRv6 Network Programming architecture allows an application to control the end-to-end journey of traffic flows through different network domains in a unified and stateless manner, meaning intermediate network devices do not store per-flow information.¶

This document covers its application to the integration of data center (DC) and wide area network (WAN) architectures using SRv6 with uSID (NEXT-CSID). This eliminates the complexities and inefficiencies associated with traditional fragmented network designs. The solution enhances scalability and enables flexible stateless service insertion by unifying the DC and WAN under a single SRv6 domain.

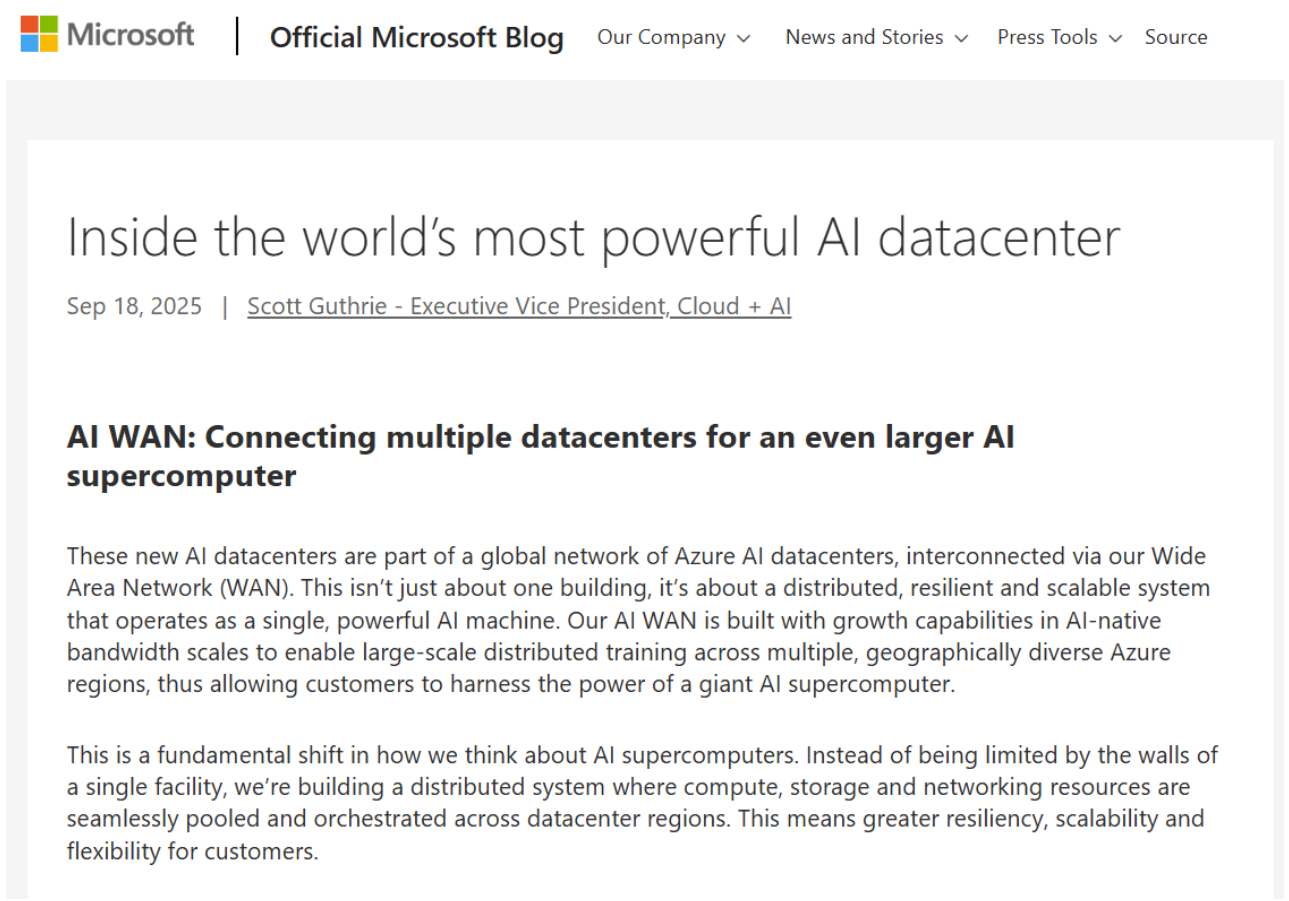
https://datatracker.ietf.org/doc/html/draft-filsfils-srv6ops-srv6-e2e-dc-frontend-wan

**Benefits**

* **Gateway Removal**: Eliminates stateful DCI translation, removing a critical bottleneck in traditional architectures. This leads to better performance, reliability, scalability, and lower operational costs.¶
* **End-to-End**: Unified underlay and overlay from DC to WAN.¶
* **Scalability**: The intermediate nodes do not maintain any per-flow state.¶
* **Lower MTU Tax**: SRv6 uSID (NEXT-CSID) allows encoding up to six uSIDs within the IPv6 Destination Address field. This eliminates the VXLAN overhead, improving transport efficiency.¶
* **Flexible Service Insertion**: Dynamically steer selected traffic through services (e.g., security appliances) without complex policy-based routing mechanisms. Insert services anywhere in the network and apply the policy on a per-vm, per-prefix, or per-customer basis.¶
* **Smaller Clusters**: Due to the flexible traffic steering capabilities provided by SRv6, the operator may choose to deploy smaller service clusters (e.g., firewall clusters) distributed along the network. This drastically reduces the scaling challenge of clusters.¶
* **Egress Peer Engineering (EPE)**: SRv6 enables precise outbound traffic control, optimizing interconnection with external networks.¶

# AI DCI & AI WAN

- DCI is crucial for AI
- TE is a MUST for AI DCI
- SRv6 is the Perfect Fit for AI DCI
  - AI => TE
  - DCI => IP
  - IP + TE = SRv6



Microsoft | Official Microsoft Blog  Our Company  News and Stories  Press Tools  Source

## Inside the world's most powerful AI datacenter

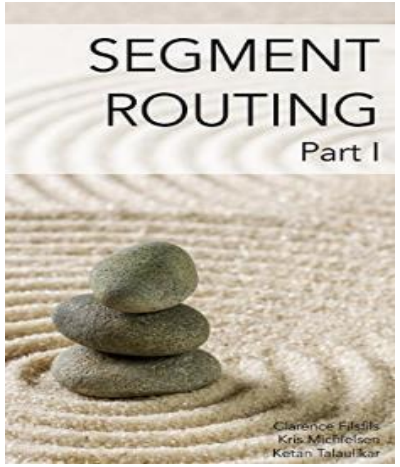
Sep 18, 2025 | Scott Guthrie - Executive Vice President, Cloud + AI

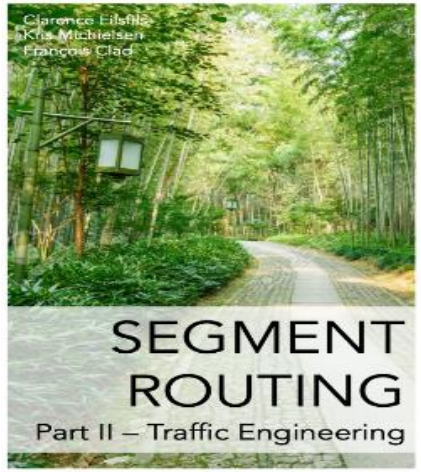
**AI WAN: Connecting multiple datacenters for an even larger AI supercomputer**

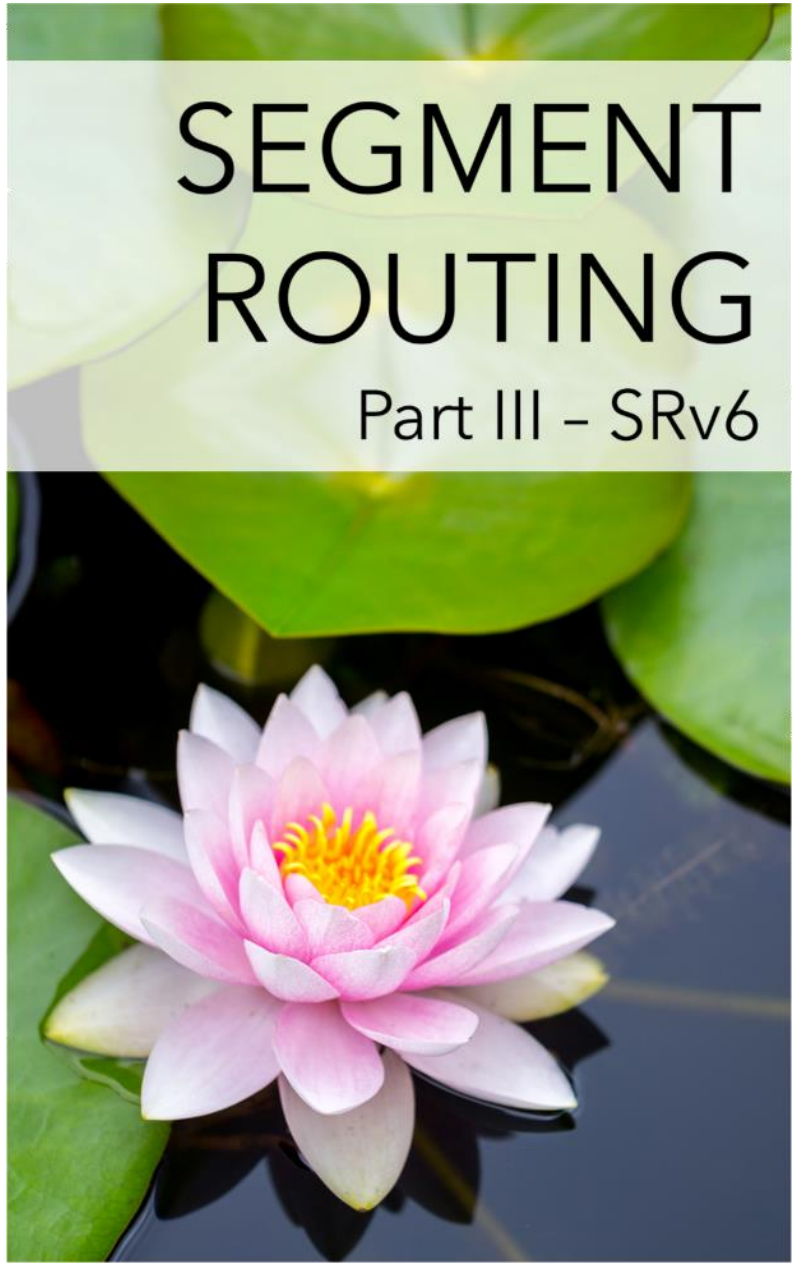
These new AI datacenters are part of a global network of Azure AI datacenters, interconnected via our Wide Area Network (WAN). This isn't just about one building, it's about a distributed, resilient and scalable system that operates as a single, powerful AI machine. Our AI WAN is built with growth capabilities in AI-native bandwidth scales to enable large-scale distributed training across multiple, geographically diverse Azure regions, thus allowing customers to harness the power of a giant AI supercomputer.

This is a fundamental shift in how we think about AI supercomputers. Instead of being limited by the walls of a single facility, we're building a distributed system where compute, storage and networking resources are seamlessly pooled and orchestrated across datacenter regions. This means greater resiliency, scalability and flexibility for customers.

# Stay up-to-date




segment-routing.net


CISCO

# References -1

- OpenAI tutorial at Sigcomm (Sept 2025)
  - https://github.com/craiciu/sigcomm25-ethernet-ai-tutorial
- MSFT Blogs
  - https://blogs.microsoft.com/blog/2025/09/18/inside-the-worlds-most-powerful-ai-datacenter/
  - https://blogs.microsoft.com/on-the-issues/2025/09/18/made-in-wisconsin-the-worlds-most-powerful-ai-datacenter/
- MSFT SRv6 Presentations:
  - https://www.youtube.com/watch?v=WzBI4sW2tzg
  - https://www.segment-routing.net/conferences/2025-10-16-ocp-summit25-microsoft-srv6-ai-backend/
  - https://www.segment-routing.net/conferences/2025-08-20-hoti-symposium-2025/
  - https://www.segment-routing.net/conferences/2025-04-30-ocp-emea-microsoft-srv6-ai-backend/
  - https://www.segment-routing.net/conferences/Paris25-Microsoft-Rita-Hui/
- SRv6 for deterministic path placements and SONiC extensions
  - https://www.segment-routing.net/conferences/2025-08-28-sonic-mini-summit-open-source-summit-eu/
- AMD – SRv6 with multipath in AMD P4 NIC
  - https://www.youtube.com/watch?v=AgrpNj0YEKQ
- Alibaba Ecore with SRv6
  - https://www.segment-routing.net/conferences/2025-10-16-ocp-summit25-alibaba-ecore/
  - https://sonicfoundation.dev/driving-innovation-alibaba-and-cisco-co-dev-srv6-sonic-router/
- SRv6 for AI Backend IETF Draft
  - https://datatracker.ietf.org/doc/html/draft-filsfils-srv6ops-srv6-ai-backend-02

# References-2

- Nebius – DC Frontend with SRv6
  - https://www.segment-routing.net/conferences/Paris25-Nebius-Alexey-Gorovoy/
  - https://www.segment-routing.net/conferences/20250213-cleur-nebius-srv6-usid-dc-frontend-to-peering/
- IETF – SRv6 in Frontend DC Design
  - https://datatracker.ietf.org/doc/html/draft-filsfils-srv6ops-srv6-e2e-dc-frontend-wan-00
- Swisscom SRv6 deployment
  - https://www.segment-routing.net/conferences/Paris25-Swisscom-Rolf-Schmid/
  - https://www.segment-routing.net/news/2024-10-09-swisscom-srv6-usid-ipm/
- Rakuten SRv6 Deployment
  - https://www.segment-routing.net/conferences/Paris25-Rakuten-Akash-Agrawal/
- RWS SRv6 deployment - SRv6 deployment in a public sector organization
  - https://www.segment-routing.net/conferences/20250213-cleur-rws-srv6-deployment/