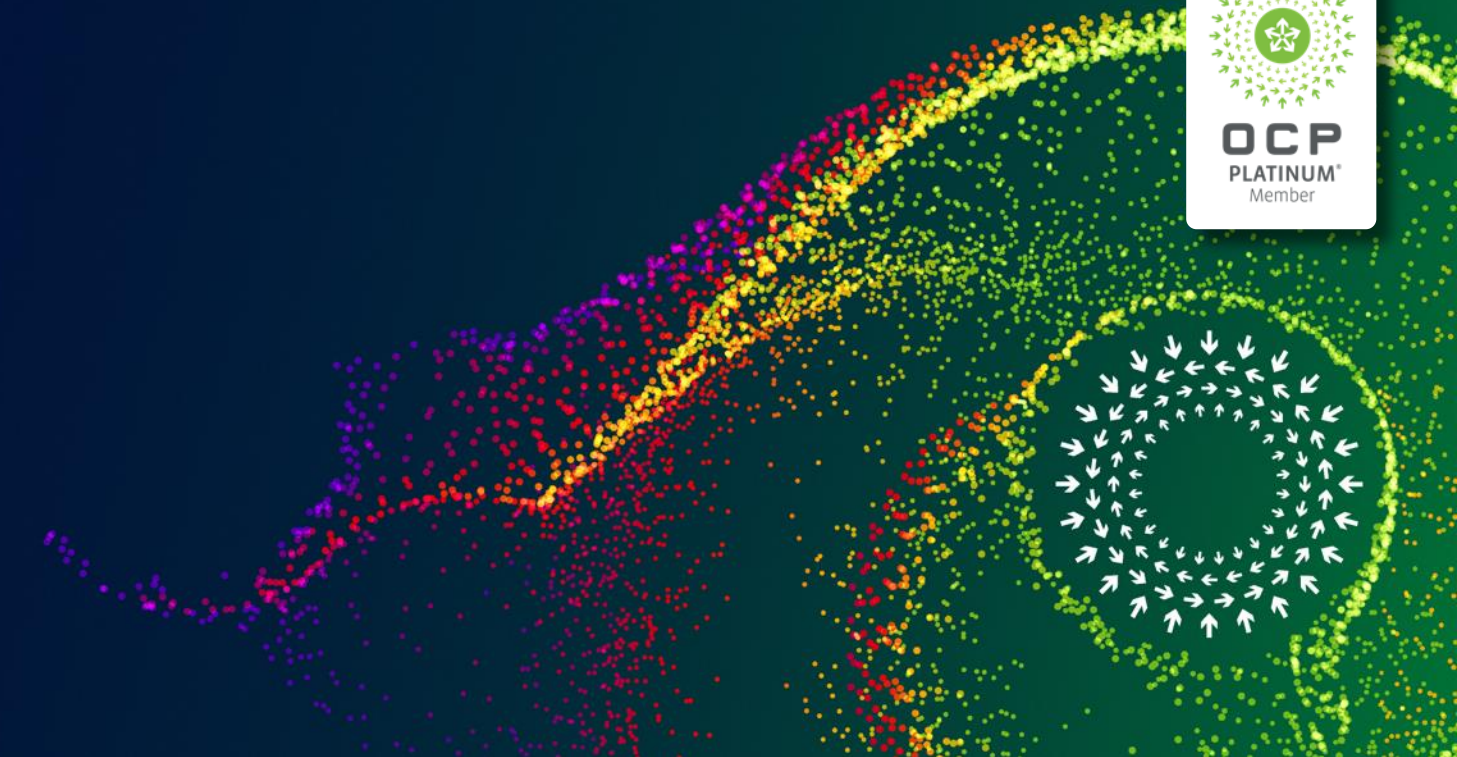
IT INFRASTRUCTURE:
NETWORK & STORAGE

Innovating Network Protocol Stacks for Scalable AI Infrastructure

Clarence Filsfils, Cisco Systems

Guohan Lu, Microsoft

OCP PLATINUM Member



OCP EMEA SUMMIT | 29-30 April, 2026
Barcelona, Spain

Legacy protocol stack

- Development of a myriad of per-domain shim layers to enrich IP
  - InfiniBand in AI/HPC
  - VxLAN in the DC
  - MPLS in the core
- Proprietary solutions & vendor lock-in
- Low scale
  - No summarization in MPLS
- Poor End-to-End Services and High Operation Cost
  - Neither TE nor stateless service chaining in VxLAN
  - Translation gateways between domain (VxLAN/MPLS)

Programmable IPv6

- 128-bit DA becomes:
  - Ordered list of SRv6 uSID instructions (uSID)
  - SRv6 uSID instruction bound to anything .. Underlay TE, overlay, service chaining
  - 6 SRv6 uSID instructions in DA
  - All deployments requires less than 6 uSID
- SRv6 uSID
  - Outperforms IPv4 shim layers: MPLS, VxLAN, GTP
  - Delivers Any Service, End-to-End, Across domains
  - Standard, Open, Interoperable

End-to-end control from the application

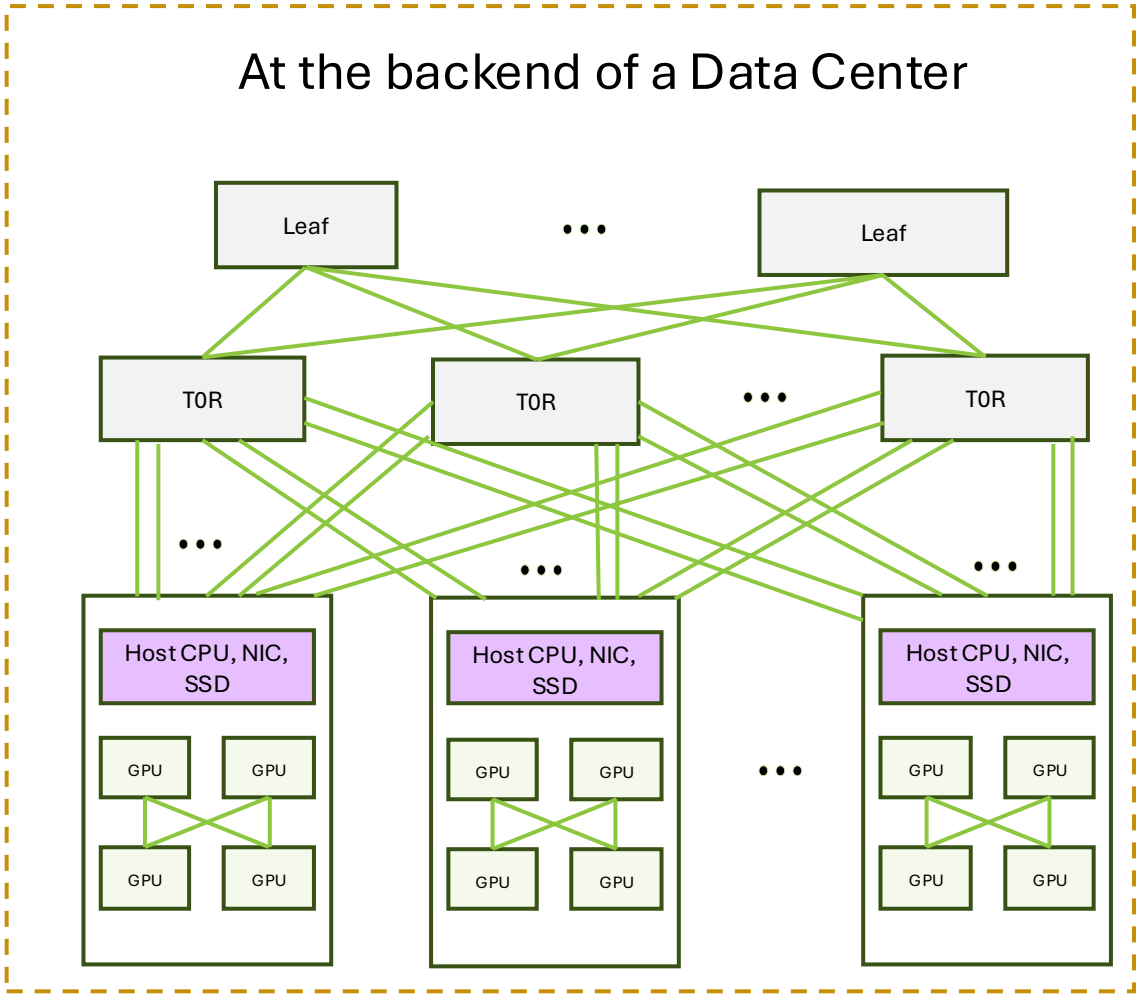
- SRv6 empowers applications to select their path
- Deterministic path for GPU-to-GPU
  - Microsoft AI backend [Public Reference]
- Stateless firewall service insertion
  - Nebius AI Frontend [Public Reference]

Scale out

SRv6 for Deterministic place placement to achieve Optimal Load-balancing

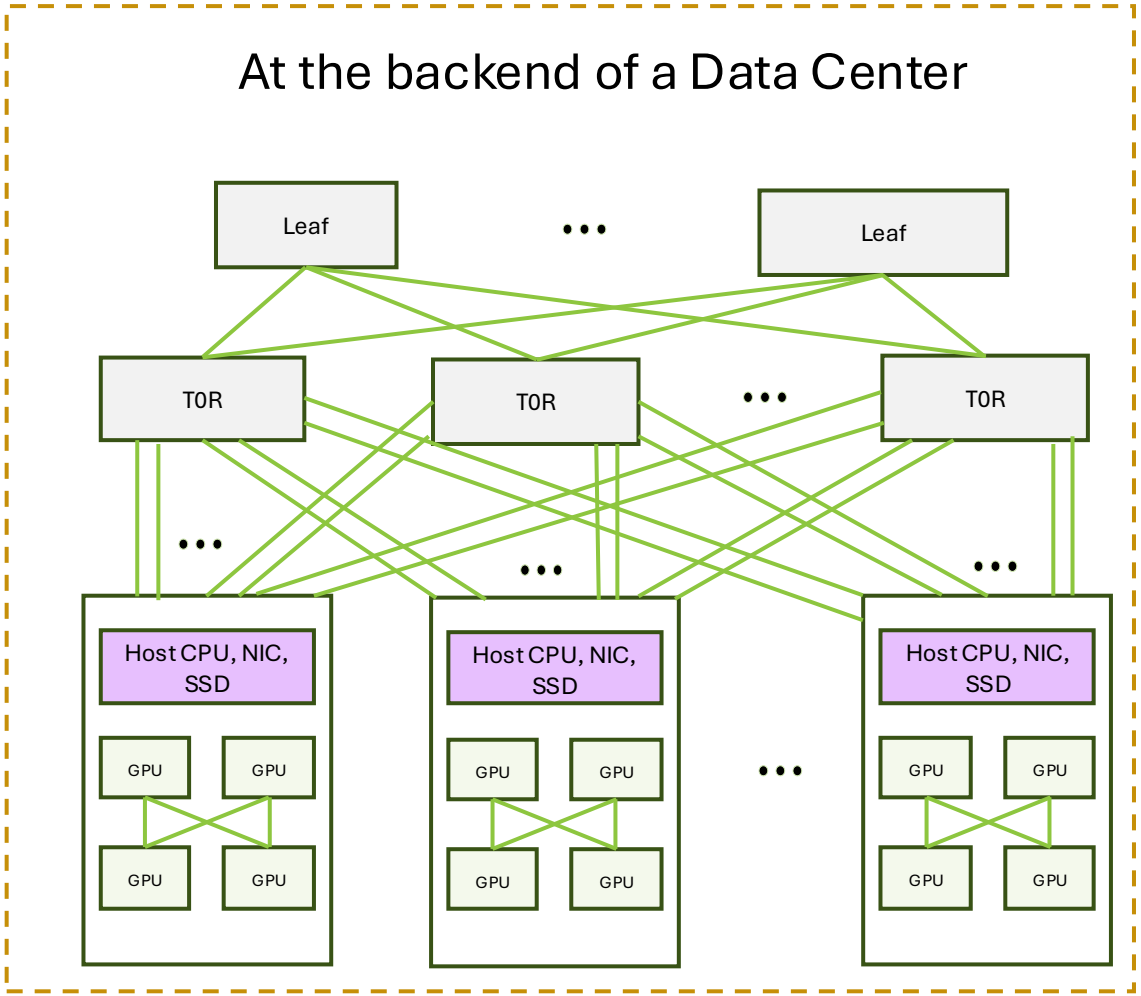
OCP EMEA SUMMIT
29-30 April, 2026
Barcelona, Spain

AI - Raising the Bar for Hyperscalers Networks



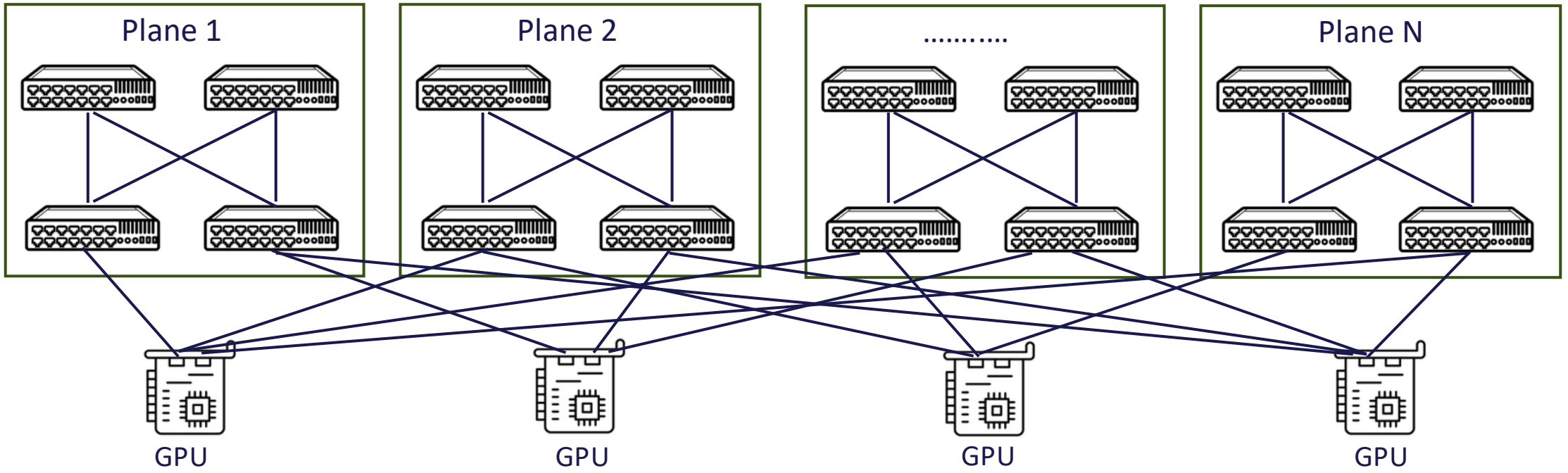
- New Traffic Pattern:
  - Small number of large flows
  - Periodic bursts of data sent synchronously

AI - Raising the Bar for Hyperscalers Networks



- Challenges:
  - Traditional passive ECMP-based load balancing mechanisms suffered from low entropy problem.
  - Failures of communications in LLM training are costly
  - An epoch of training is blocked until the synchronized collective communications of last epoch finishes
  - GPU hours are expensive resources because of tight supply
  - If an ongoing training job crashes, all progress since the last checkpoint may be lost.

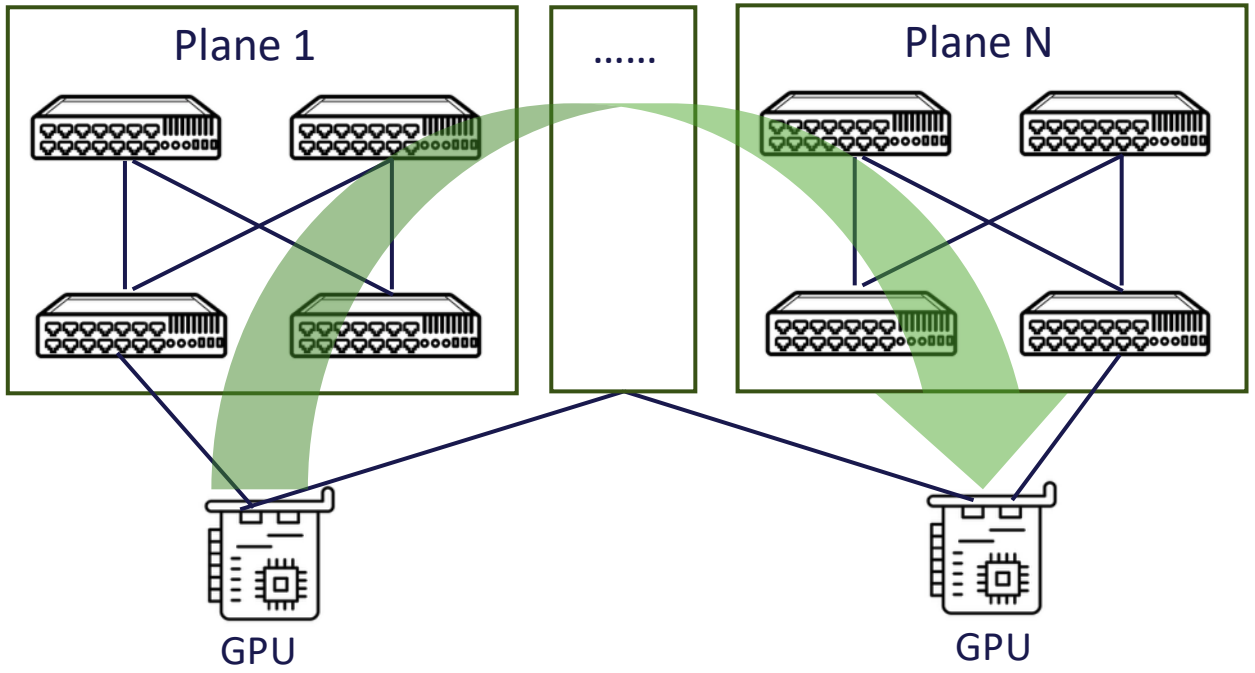
Microsoft Fairwater DC



Multi-Plane design allowing up to 524,288 GPUs!

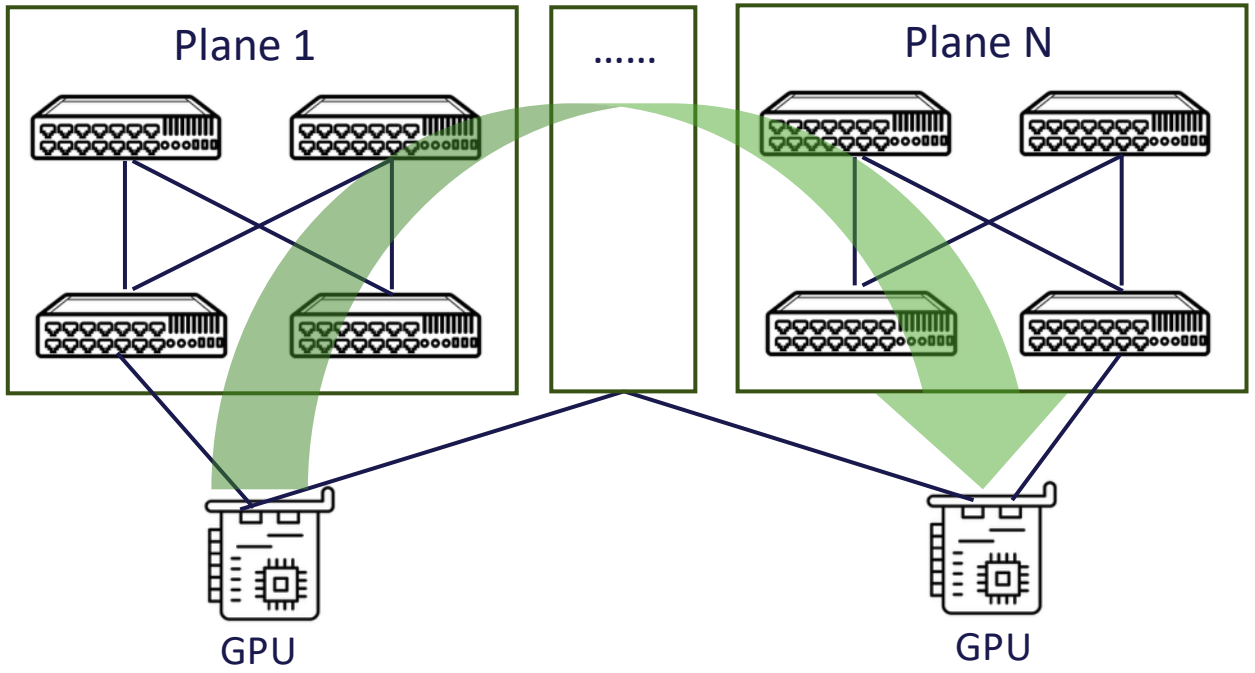
SRv6 in Microsoft Fairwater DC

- Fined-grained control based on Source Routing
- Path enumeration for traffic management
- Integration with AI workloads flow scheduling provides optimal network performance
- Source quickly reroute upon path failures or congestion
- Reliability: No dependency on Routing Protocol
- Ship-in-the-night with traditional traffic
- Standard, Open, Muti-vendor
- Ease of operation and troubleshooting



SRv6 in Microsoft Fairwater DC

- The uSID Stack is communicated to NIC
- Source NIC
  - Encapsulates the traffic with the uSID List
- Switch
  - Forward the packet based on the uSID List - Strict Source Routing
  - If drop, Packet Trimming
- Destination NIC
  - Decapsulates the uSID List and delivers the packet
- Both SRv6 uN and uA are supported.

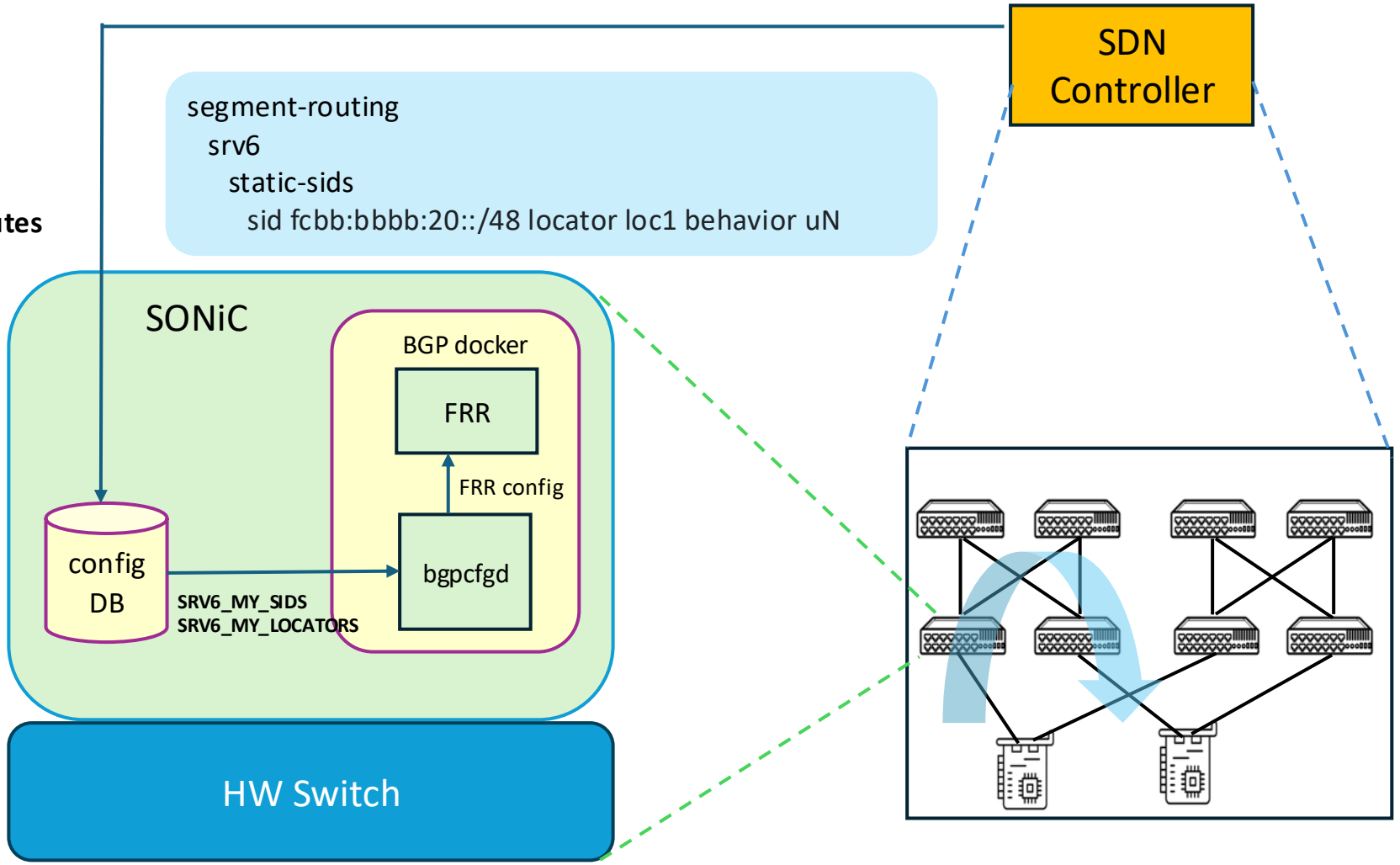


SRv6 in SONiC for AI Backend

SRv6 Configs

IPv6 Static Routes

```
segment-routing
  srv6
    static-sids
      sid fcbb:bbbb:20::/48 locator loc1 behavior uN
```



Config DB

SRV6_MY_LOCATORS Table
serves as a configuration source for SRv6 locators

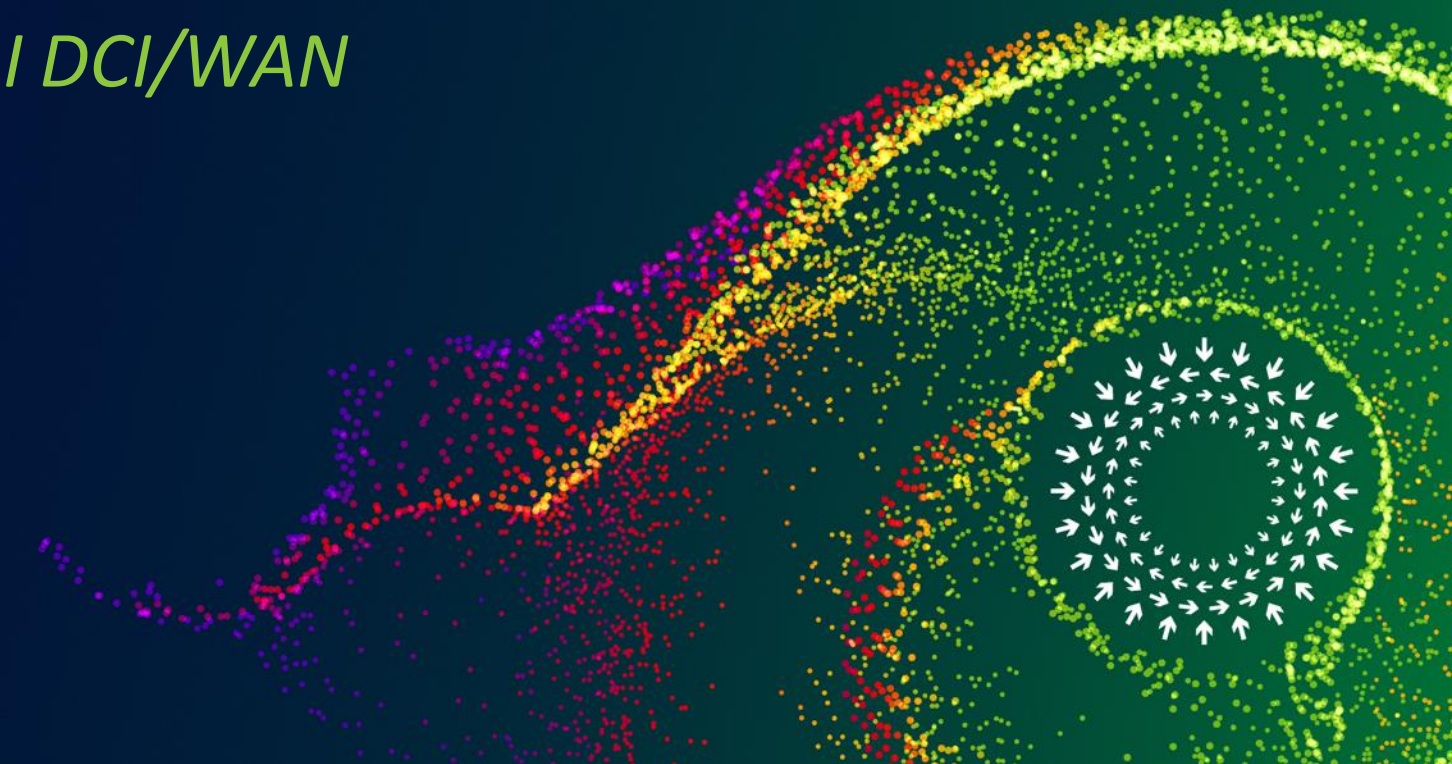
SRV6_MY_SIDS Table
configures segment identifiers

SRv6 in SONiC – Ecosystem Driven

- Mature:
  - AI Backend: SDN Provisioned SRv6 Fabric
  - SRv6 L3VPN, SRv6 GRT
  - SRv6 Underlay Traffic Engineering, SRv6 Steering
  - SRv6 SID Manager –Interoperability between SONiC and Other NOS
  - BGP-LS: IS-IS and eBGP fabrics
- Deployed:
  - Microsoft Fairwater DC [Public Reference]
  - Alibaba eCore – [Public Reference]
- Rich Ecosystem:
  - Contributors: Cisco, Microsoft, Alibaba, Broadcom, Nvidia
  - Mainline: 485 PR across SAI/SONiC/FRR.

Scale Across

SRv6 for Traffic Engineering in AI DCI/WAN

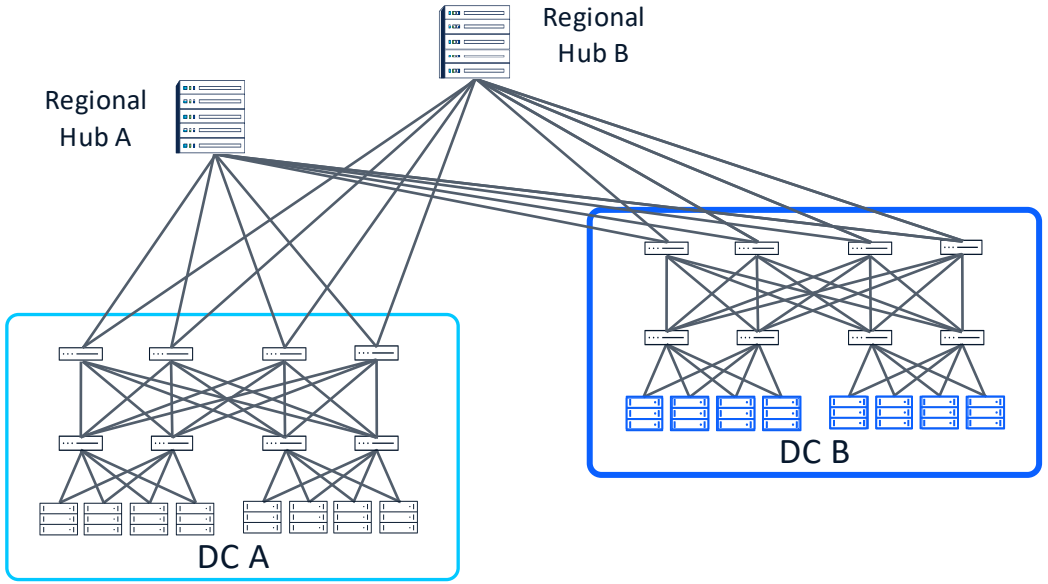


OCP EMEA SUMMIT

29-30 April, 2026
Barcelona, Spain

AI DCI/WAN is crucial

- Power limitations
- Workload spans clusters scattered over a regional network



Traffic Engineering (TE) is key for AI DCI

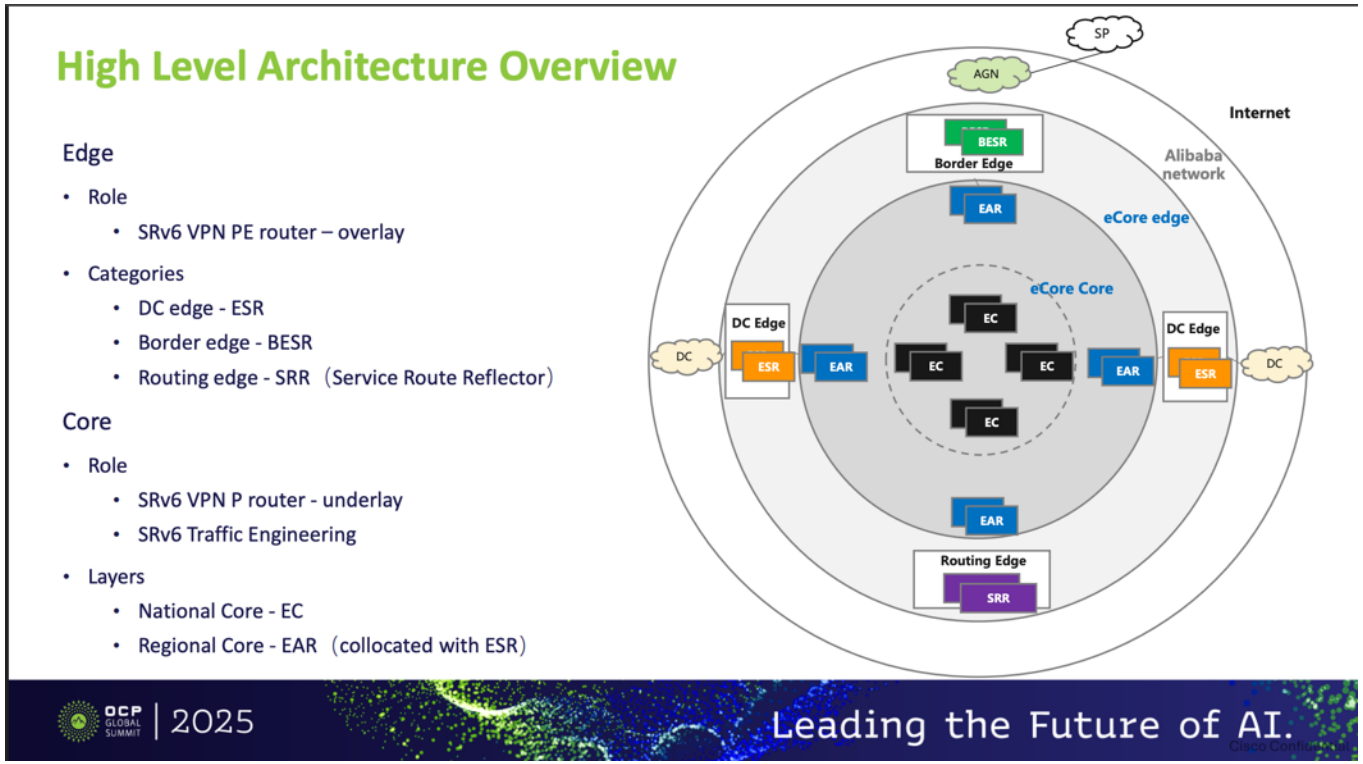
- The topology is less symmetric
- The capacity is more scarce
- Latency matters more
- Weirdness:
  - Shortcuts between clusters only for specific workloads
  - SRLG’s

SRv6 TE - Perfect fit for AI DCI

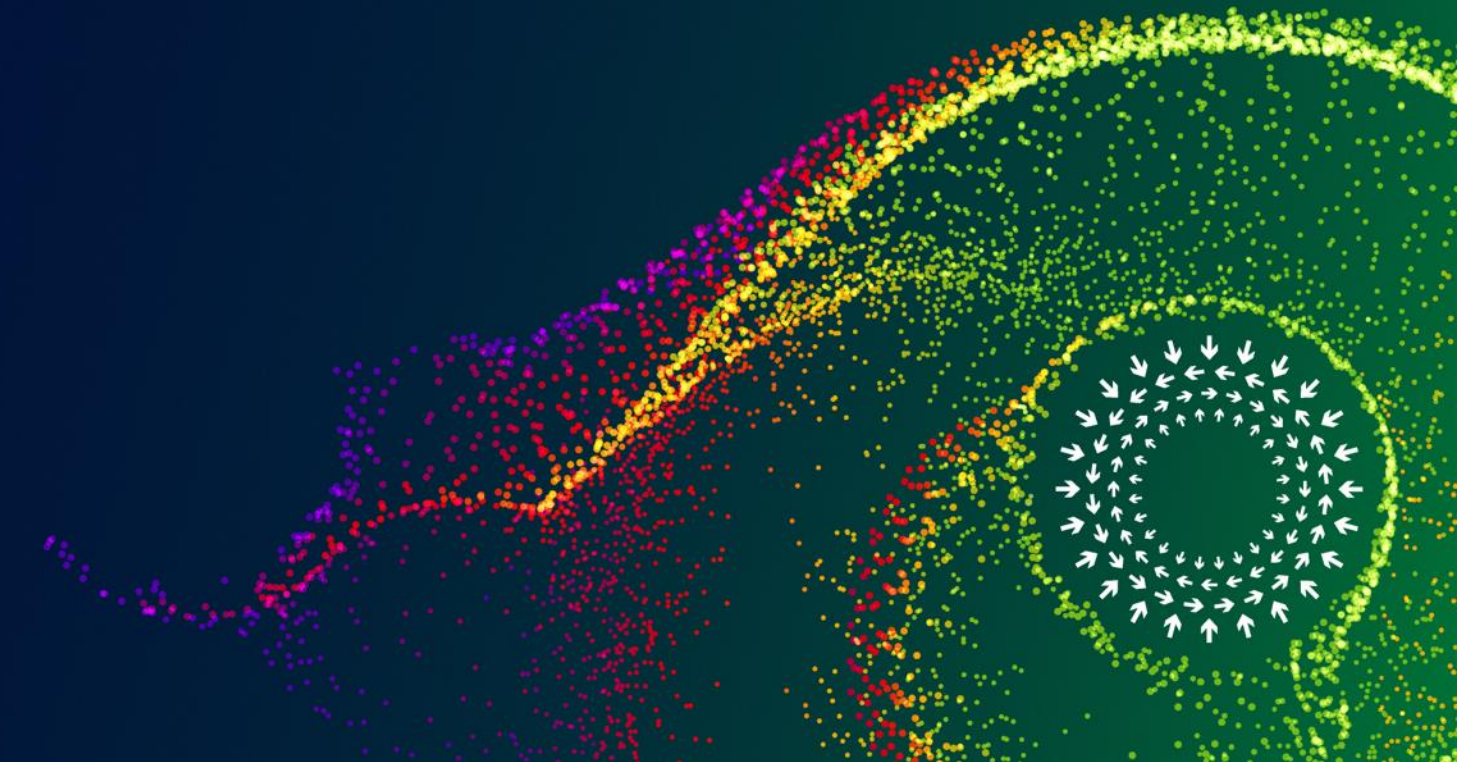
- The “IP nature” of legacy DCIs
  - Most pre-SRv6 uses IPinIP for source routing
- SONiC expansion in AI DCI
  - Leverage the SRv6 richness/maturity in SONiC from AI DC
- SRv6 is rich/mature across DC HW
- End-to-End model
  - SRv6 end-to-end from DC to DC through DCI

SRv6 in AI DCI/WAN at Alibaba

- Ultra Scale AI DCI/WAN deployment across in China
- SONiC/SRv6 in DCI
- Publicly presented at OCP



Scale Up



Scale Up Requirements

- Ultra-low Overhead
  - Remain highly efficient for 32, 64, 128-byte packet sizes (Flit transactions)
- Simplified packet processing
  - Single Header Architecture
  - Skip unnecesary building-blocks in HW (e.g., LB)
- Native strict source routing support
- Two-tier fabric support
  - Scale Up and combined Scale Up/Scale Out fabrics
- High-radix support (>576)

Call to Action

- For more information about SRv6 https://www.segment-routing.net/
- Inviting contributions to SRv6 SONiC
  - SONiC Community meeting: https://sonic-net.github.io/SONiC/Calendar.html
  - SONiC Routing Workgroup: sonic-wg-routing@lists.sonicfoundation.dev | Home

Thank You!

OCP EMEA SUMMIT | 2026