TOR VERGATA
UNIVERSITÀ DEGLI STUDI DI ROMA

cnit consorzio nazionale interuniversitario per le telecomunicazioni

SRV6 WORKSHOP @ NETDEV 0X1A

SRv6 for the AI Backend

Stefano Salsano[1,2]

[1]University of Rome Tor Vergata, Italy · [2]CNIT, Italy

Netdev 0x1A — The Technical Conference on Linux Networking

SRv6 Workshop — *13 July 2026, Rome, Italy*

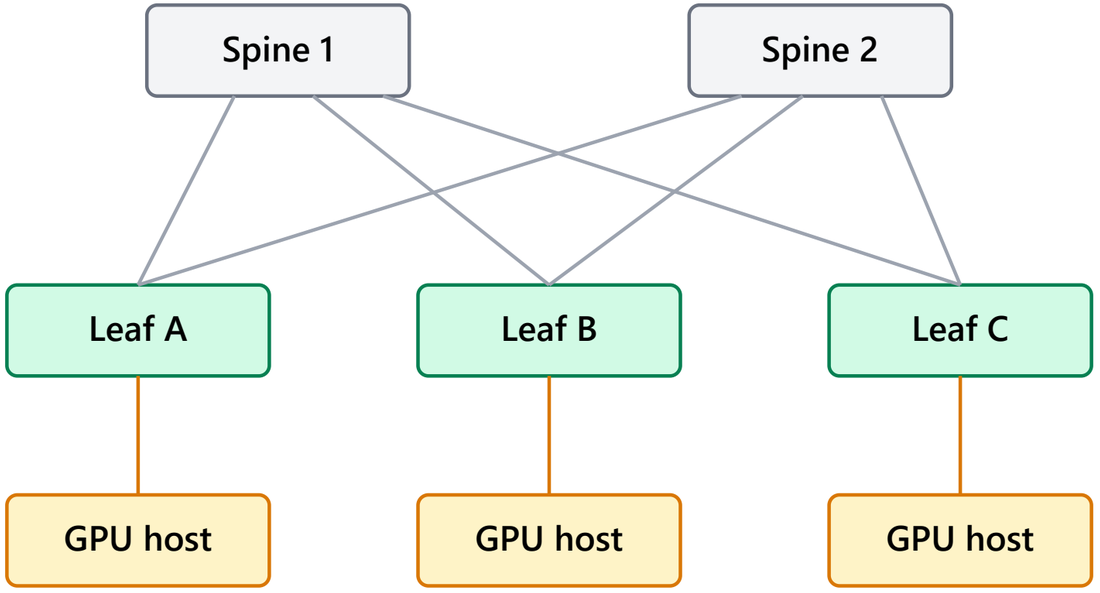
C. Filsfils, P. Camarillo, A. Abdelsalam, A. Quinci, A. Tulumello, A. Mayer, P. Loreti, L. Bracciale, S. Salsano,
"Toward Deterministic Path Placement in AI Backends: A Practical SRv6-Based Architecture", **IEEE CNSM 2025**

PART 1 — THE PROBLEM

An AI backend is not the Internet

- **One owner**, one fabric, known topology
- **Few, huge flows** — not millions of small ones
- Collectives are **synchronous**: all-reduce, all-to-all
- The **slowest flow** sets the job completion time
- GPUs idle while the network catches up

Consequence. Average throughput is irrelevant: the *tail* is what you pay for.



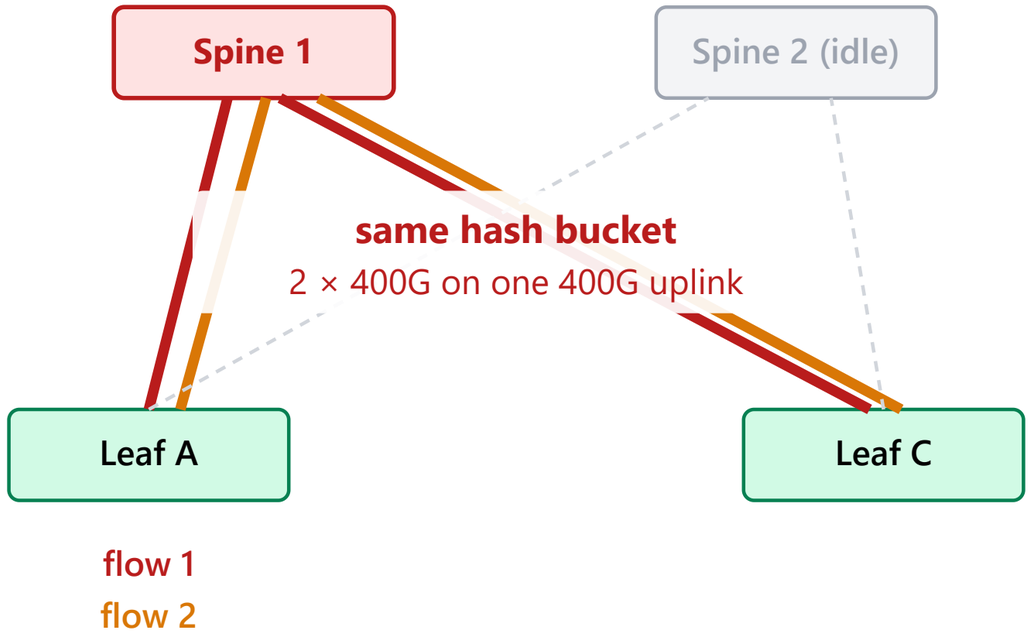
all-reduce: everybody waits for the last byte

PART 1 — THE PROBLEM

ECMP: spreading is not placement

- The switch hashes the 5-tuple — **it does not plan**
- **Hash collisions**: two elephants, one uplink
- No entity **owns** the end-to-end path
- On failure: wait for **control-plane convergence**
- Rehashing moves flows you did *not* want to move

The gap. We spread flows *statistically* on a fabric we control *exactly*.



half the fabric is idle, half the job is late

PART 2 — THE FIX

Source routing: the segment list *is* the path

1 **Decide at the source.**

The host (or the NIC) picks the exact fabric path, per packet or per flowlet.

2 **No hash, no collision.**

Placement is explicit: the fabric stops guessing and just executes the SID list.

3 **Failures are a local decision.**

Switch the SID list at the host — no waiting for control-plane convergence.

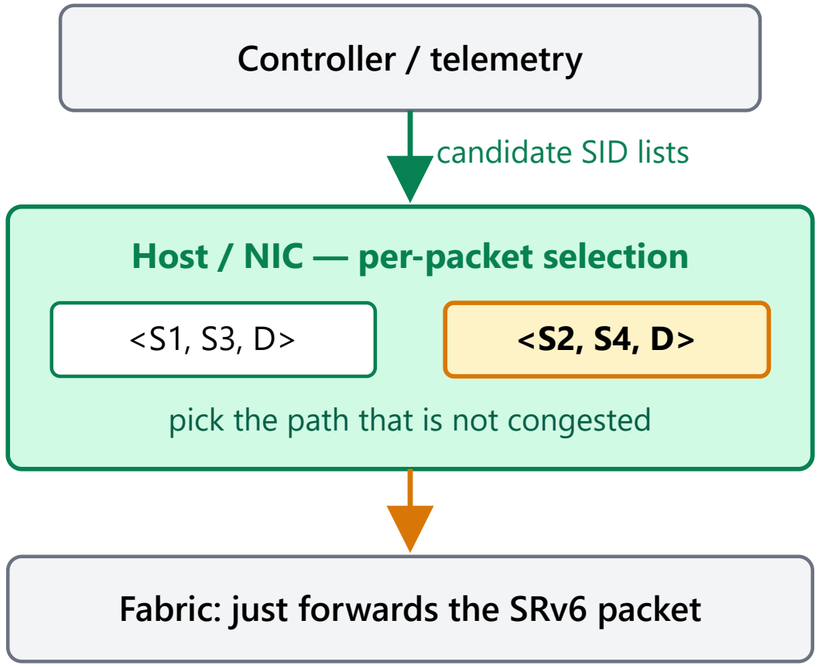
And it is plain IPv6. SRv6 needs no new dataplane in the fabric — the switches already forward it.

PART 2 — THE FIX

Deterministic, congestion-aware placement

- **Telemetry in**: link load, queue depth, failures
- **Path set out**: a few disjoint SID lists per destination
- The host **selects**, packet by packet
- Reaction time = **host RTT**, not convergence time

Division of labour. The controller *proposes* the paths, the host *disposes* of the packets.



no new dataplane, no hash lottery

PART 3 — LINUX

The Linux datapath

- `seg6` lwtunnel: encap, one static SID list per route
- `seg6local`: the behaviors, on the receive side
- The hard part: **choosing** the list per packet
- Not a routing lookup — a **policy** decision
- Needs feedback state next to the datapath

```
# one path, statically bound to a route
ip -6 route add fc00:d::/64 \
    encap seg6 mode encap \
    segs fc00:1::s1,fc00:3::s3 dev eth0

# what an AI backend actually needs
for each packet:
    p = select_path(dst, telemetry, state)
    push_srh(p)
```

The kernel gap. `seg6` can *apply* a path; nothing in it *picks* one.

Where the selection lives
eBPF at the socket / TC egress — or XDP, or the NIC itself.

PART 3 — LINUX

Doing it at line rate

	Where it runs	Per-packet choice	Line rate at 400G+
seg6 lwtunnel	kernel, route-bound	**no — static list**	CPU-bound encap
eBPF / XDP	kernel, programmable	**yes** — maps hold the path set	good, but CPU cycles per packet
NIC offload	hardware	**yes**, if the NIC exposes the SID list	**the only way to keep the GPUs fed**

The ask to this room. A clean split: *policy* in eBPF, *encap* in the NIC — with a kernel API that does not fight either.

PART 4 — THE ECOSYSTEM

It does not work alone

- **FRR.**
  SID allocation and distribution — the host needs to know which segments exist.
- **SONiC.**
  The switch side of the same story — see Ahmed's talk right after this one.
- **Telemetry.**
  Load and failure signals, fast enough to be worth reacting to.
- ★ **MRC.**
  This architecture underpins the industry MRC protocol — *main-track talk* at this conference.

Conclusions

What we have

An SRv6 architecture that turns an AI fabric from *statistically spread* into *deterministically placed* — on plain IPv6, no new dataplane.

What we need

Kernel and NIC support for *per-packet* segment-list selection at line rate, and the telemetry loop to drive it.

Take-home. In a fabric you own completely, guessing the path is a choice — and a bad one.

C. Filsfils, P. Camarillo, A. Abdelsalam, A. Quinci, A. Tulumello, A. Mayer, P. Loreti, L. Bracciale, S. Salsano,
“Toward Deterministic Path Placement in AI Backends: A Practical SRv6-Based Architecture”, **IEEE CNSM 2025**

stefano.salsano@uniroma2.it