srv6ops
Internet-Draft
Intended status: Informational
Expires: 30 October 2026

S. Sangli
S. Hegde
M. Styszynski
HPE
M. Nanduri
Microsoft

28 April 2026

# BGP based SRv6 Routing Planes for DC network
draft-hss-srv6ops-srv6-routing-planes-01

## Abstract

This document introduces a BGP-based multi-planar routing architecture for modern data center networks, with a particular focus on environments running AI/ML workloads that demand traffic segregation. The proposed solution enables deterministic routing for workloads with characteristics such as collective communication and multi-tenancy. It allows the creation of multiple logical routing planes over a shared physical infrastructure by defining planes through three key elements: Constraints (e.g., fabric color inclusion/exclusion) Calculation types (e.g., shortest path) and Metric types (e.g., cost, delay, bandwidth).

Table of Contents



## 1. Introduction

Modern Data Center (DC) networks are typically built using Clos topologies, which provide an n-hop path (commonly 3, 5, or 7 hops) between ingress and egress with a minimal number of intermediate nodes. This design offers straightforward scalability as traffic demands increase. Several factors influence DC network buildout, including traffic characteristics, AI workload requirements, data generation rates, user distribution, and the placement of compute, storage, and application resources. DC networks generally operate as pure IP fabrics, using the BGP routing paradigm. Nodes (switches or routers) establish single-hop eBGP sessions with their neighbors [RFC7938].

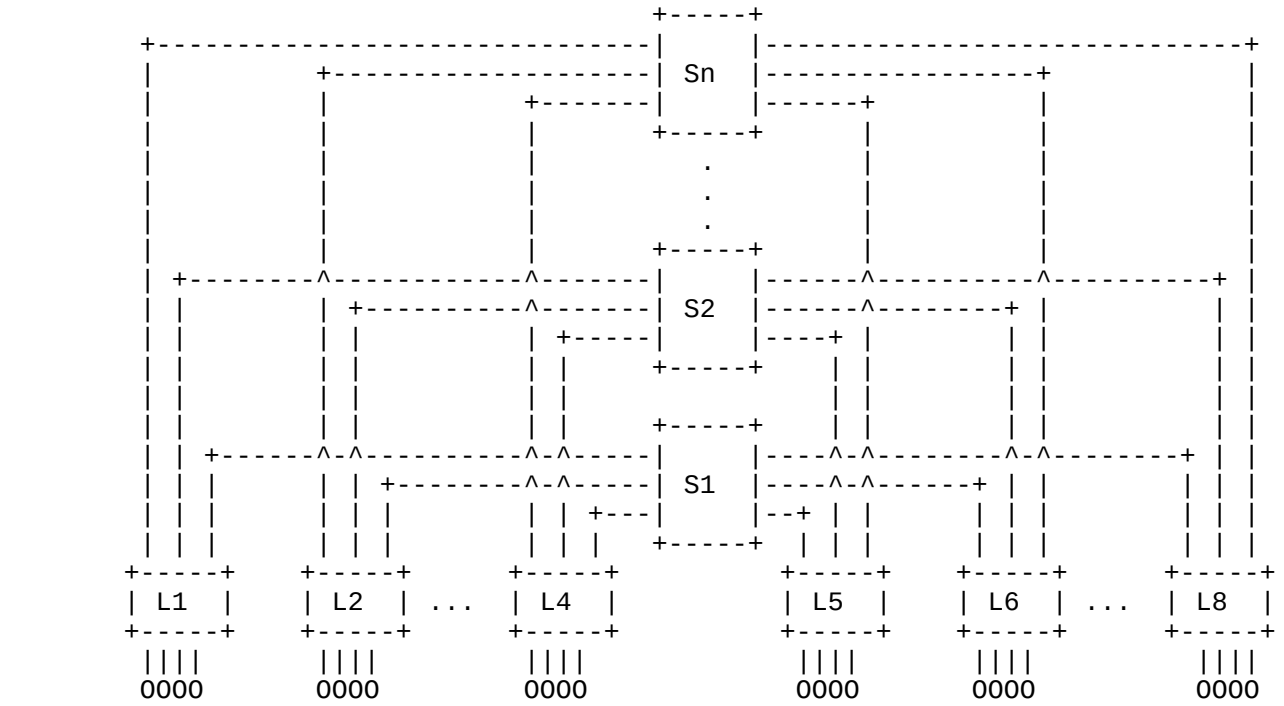
When hosting AI workloads, DC networks are optimized to maximize bandwidth usage and handle traffic with low entropy characteristics. AI models have grown dramatically, with parameter counts reaching billions or even trillions. This scale requires distributing workloads across multiple datacenters, where inter-DC networks must support mixed traffic types. Consequently, AI workloads share the same physical infrastructure with other applications such as storage, etc., each with distinct bandwidth and latency requirements.

Logical routing planes provide strict separation between traffic types while leveraging the same physical infrastructure. This ensures predictable performance across different types of applications. BGP is widely deployed in datacenters and often serves as the routing protocol for interconnecting regional datacenters located within close proximity (e.g., 100120 km). While mechanisms for logical routing planes have been defined for IGP protocols [RFC9350], a comparable capability is required for BGP.

## 2. Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

## 3. BGP based Routing Planes




Legend: S: Spine, L: Leaf, O: Compute Server NIC

Figure 1: Data Center Clos Network

This document proposes a BGP-based multi-planar architecture that enables the creation of multiple routing planes within a data center fabric.  The key characteristics are as follows:

* Routing Plane Definition:

Each routing plane is defined by a set of constraints, a calculation type, and a metric type.  These parameters, applied to the physical topology of nodes and links, form a logical routing plane.

* Fabric Colors, Metrics and Calculation-type:

Physical links can be tagged with Fabric Colors. Routing planes may include or exclude specific colors, and/or be differentiated by metric types such as cost, delay, or bandwidth. The calculation-type refers to the consistent way for best path selection that is applied within a routing plane. For example, all routers in a Routing Plane apply same criteria expressed via BGP import policy for best path computation.

* Expressing constraints:

The Routing Plane configuration in conjunction with BGP policy can combine multiple characteristics (e.g., exclude a fabric color while optimizing for delay) thereby providing flexibility.

* Pre-Built Configuration:

Routing planes are provisioned via configuration. The BGP routing protocol builds routes and next-hops according to defined constraints. Application traffic is mapped to one or the other routing planes based on application intent.

* Application Intent Expression:

Application intent is conveyed using BGP extended color communities, which are associated with prefix advertisements.

* Failure Handling:

In the event of link or node failures, a routing plane may become partitioned. Traffic can fallback to alternate planes.

* Policy-Based Control:

Routing plane definitions are applied as import/export policies in BGP advertisements. Importantly, this framework does not require new BGP protocol extensions.

Motivated by the deterministic path forwarding mechanism described in [I-D.wang-idr-dpf], the approach outlined here provides a generic and extensible framework for defining routing planes. The goal is to demonstrate how routing planes can be constructed in SRv6 networks by leveraging existing segment routing constructs.

## 4. BGP Routing Planes applied to SRv6 network

The following section describe the BGP Routing Plane solution applied to SRv6 networks.

Figure 1 diagram illustrates a multi-planar data center fabric in which nodes L1, L2, and spines S1, S2 belong to the Green routing plane, while nodes L5, L6 and spines S3, S4 belong to the Blue routing plane. Servers (e.g., Server1 and Server2) are dual-homed, with connections to both planes.

The requirement is to construct distinct Green and Blue routing planes across the fabric. Routing plane definitions can be consistently applied across the network, ensuring that each plane enforces its constraints and provides deterministic forwarding paths for application traffic.

Routing Plane Definition:

To achieve routing planes for the fabric described in Figure 1, the Routing plane definition is described below.

Green routing plane:

Calculation type: BGP Best path

Metric Type: standard metric

Set of constraints: Exclude Blue

Blue Routing Plane:

Calculation type: BGP Best path

Metric Type: standard metric

Set of constraints: Exclude Green

Each node in the fabric is provisioned with SRv6 locators along with the corresponding uN and uA SIDs derived from those locators. Nodes that belong to the Green routing plane are additionally configured with Green-specific locators, while nodes in the Blue routing plane are provisioned with Blue-specific locators.

SRv6 block for the fabric 2100:db8::/32

L1 instantiates the SID 2100:db8:0100::/48 associated with the uN instruction (End with NEXT-CSID, PSP & USD)
L2 instantiates the SID 2100:db8:0200::/48 associated with the uN instruction (End with NEXT-CSID, PSP & USD)
L5 instantiates the SID 2100:db8:0500::/48 associated with the uN instruction (End with NEXT-CSID, PSP & USD)
L6 instantiates the SID 2100:db8:0600::/48 associated with the uN

instruction (End with NEXT-CSID, PSP & USD)
S1 instantiates the SID 2100:db8:0900::/48 associated with the uN instruction (End with NEXT-CSID, PSP & USD)
S2 instantiates the SID 2100:db8:0a00::/48 associated with the uN instruction (End with NEXT-CSID, PSP & USD)
S3 instantiates the SID 2100:db8:0b00::/48 associated with the uN instruction (End with NEXT-CSID, PSP & USD)
S4 instantiates the SID 2100:db8:0c00::/48 associated with the uN instruction (End with NEXT-CSID, PSP & USD)

Green Routing Plane:

L1 instantiates the SID 2100:db8:1100::/48 associated with the uN instruction (End with NEXT-CSID, PSP & USD) corresponding to Green Routing Plane.
L2 instantiates the SID 2100:db8:1200::/48 associated with the uN instruction (End with NEXT-CSID, PSP & USD) corresponding to Green Routing Plane.
L5 instantiates the SID 2100:db8:1500::/48 associated with the uN instruction (End with NEXT-CSID, PSP & USD) corresponding to Green Routing Plane.
L6 instantiates the SID 2100:db8:1600::/48 associated with the uN instruction (End with NEXT-CSID, PSP & USD)corresponding to Green Routing Plane.
S1 instantiates the SID 2100:db8:1900::/48 associated with the uN instruction (End with NEXT-CSID, PSP & USD) corresponding to Green Routing Plane.
S2 instantiates the SID 2100:db8:1a00::/48 associated with the uN instruction (End with NEXT-CSID, PSP & USD) corresponding to Green Routing Plane.

Blue Routing Plane:
L1 instantiates the SID 2100:db8:2100::/48 associated with the uN instruction (End with NEXT-CSID, PSP & USD) corresponding to Blue Routing Plane.
L2 instantiates the SID 2100:db8:2200::/48 associated with the uN instruction (End with NEXT-CSID, PSP & USD) corresponding to Blue Routing Plane.

L5 instantiates the SID 2100:db8:2500::/48 associated with the uN instruction (End with NEXT-CSID, PSP & USD) corresponding to Blue Routing Plane.
L6 instantiates the SID 2100:db8:2600::/48 associated with the uN instruction (End with NEXT-CSID, PSP & USD)corresponding to Blue Routing Plane.
S3 instantiates the SID 2100:db8:2b00::/48 associated with the uN instruction (End with NEXT-CSID, PSP & USD)corresponding to Blue Routing Plane.

S4 instantiates the SID 2100:db8:2c00::/48 associated with the uN instruction (End with NEXT-CSID, PSP & USD)corresponding to Blue Routing Plane.

Figure 2: SRv6 SID

The BGP sessions in the Green routing plane are associated with Green admin-group [RFC5305] and the BGP sessions in the Blue routing plane are associated with Blue admin-group.

## 4.1. BGP Procedures for building SRv6 Based Routing plane

The network is provisioned with initial configurations as described in [SRv6-sids]. This configuration is performed once per routing plane and does not require modification based on changing traffic demands.

* Locator Advertisements:

Green locators are advertised as standard IPv6 prefixes (AFI-2, SAFI-1) and are tagged with the extended color community [RFC4360] corresponding to Green.

Blue locators are advertised similarly, with the extended color community corresponding to Blue.

* BGP Policy Mapping:

Each node is configured with BGP policies that map incoming extended color communities to the appropriate routing plane. When a policy maps to a routing plane definition, the routing plane' s characteristics are applied to the incoming advertisement to determine acceptance or rejection.

A locator advertisement tagged for the Green plane is accepted only if received on a BGP session associated with the Green admin-group.

Similarly, a locator advertisement tagged for the Blue plane is accepted only if received on a BGP session associated with the Blue admin-group.

Once the control plane has been established for multiple routing planes, collective communications can leverage the data plane mechanisms described in the Section 7 to forward traffic across the appropriate planes. BGP Routing Planes solution builds deterministic paths inside a fabric purely based on routing. It does not require any controller based or out-of-band path calculation, path provisioning etc.

```
       collective1 uses Blue routing plane :
              Srv6 encapsulated data packet loadbalanced across L5 & L6:
                  assuming destination prefix is associated with L5/L6
                   2100:db8:2500
                   2100:db8:2600
       collective2 uses Green routing plane
              Srv6 encapsulated data packet loadbalanced across L1 & L2:
                  assuming destination prefix is associated with L1/L2
                   2100:db8:1100
                   2100:db8:1200
```

Figure 3: BGP Routing based deterministic paths

## 5. Multi Tenancy

Cloud providers often face the requirement of supporting multiple customer AI/ML workloads simultaneously within the same data center. To ensure isolation, customer traffic must be carried on separate paths, preventing one workload from impacting another.

This separation can be achieved by constructing source-routed paths within the routing planes, using mechanisms described in [I-D.filsfils-srv6ops-srv6-ai-backend]. For example:

A source-routed path for Customer A, Collective Type 1 may be built using uA and uN SIDs defined for the Blue routing plane on node S3.

A source-routed path for Customer B, Collective Type 1 may be built using uA and uN SIDs defined for the Blue routing plane on node S4.

This approach ensures that each customer’ s workload traffic remains isolated within its designated routing plane, while still leveraging the shared physical infrastructure and this is possible only with source based routing.

Such source routing based solutions MUST require controller or any out-of-band mechanisms. With this, one can learn the fabric network topology, the details of the hosts network attachment. It is also very essential to collect the current operational state of the nodes and the links etc. for providing input to the soruce based path computation.

```
Source Routed path for customer A collective1 uses Blue routing plane S3:
            2100:db8:2b00:2500
Source Routed path for customer B collective1 uses Blue routing plane S4:
            2100:db8:2C00:2600
Source Routed path for customer A collective2 uses Green routing plane S1:
            2100:db8:1900:2100
Source Routed path for customer B collective2 uses Green routing plane S2:
            2100:db8:1a00:2200
```

Figure 4: Source routed paths

## 6. Minimum bandwidth as a constraint

In addition to fabric color constraints, routing planes can enforce a minimum-bandwidth requirement. The BGP link-bandwidth community [I-D.ietf-idr-link-bandwidth] can be used to convey end-to-end bandwidth of the path. When a BGP advertisement is received, the link-bandwidth value is extracted and compared with the minimum-bandwidth configured in the routing-plane definition, excluding paths that do not meet the threshold.

## 7. Scaling across multiple data-centers

AI/ML training models continue to grow in size and complexity, often requiring deployment across multiple datacenters. In such scenarios, the Data Center Interconnect (DCI) network must be designed to optimize for the lowest delay metric, ensuring efficient distribution of workloads.

Operators may deploy either IGP or BGP for DCI routing; in many cases, BGP is preferred due to its flexibility and widespread use. The mechanism for advertising delay metrics in BGP is defined in [I-D.ietf-idr-bgp-generic-metric]. Delay values may be configured statically or measured dynamically using protocols such as TWAMP [RFC5357].

To construct a routing plane based on delay:

* The metric-type in the routing plane definition Section 4 is set to delay.

When multiple BGP advertisements exist for the same prefix, best path selection is performed using the delay metric carried in the generic-metric attribute.

This framework is generic and extensible, allowing operators to define multi-planar networks using a variety of metric types (e.g., cost, bandwidth, delay) and constraints, depending on operational requirements.

## 8. Data Plane Considerations

Traffic in data center and interconnect networks typically consists of two patterns: bandwidth-intensive “elephant flows” and short-lived “mice flows.” These traffic patterns exhibit low entropy, and because AI computations are highly sensitive to latency, any congestion in the network can significantly degrade performance. Coping with congestion requires a combination of strategies: avoidance, detection, notification, and reaction.

* Congestion Avoidance:

Mechanisms such as strategic traffic segregation via routing planes and packet spraying across available links are employed to reduce the likelihood of congestion:

* Congestion Detection and Notification:

Techniques like Explicit Congestion Notification (ECN) and latency measurements can be scoped to individual routing planes. This allows congestion signals to be delivered to the sender with plane-specific granularity.

* Congestion Reaction:

Within a routing plane, BGP can select multiple paths to a destination, designating one or more as primary and others as backup. Backup paths can be pre-programmed, enabling traffic to switch at millisecond granularity when congestion occurs.

* Policy Enforcement:

Routing plane policies can reflect customer intent. For example, links experiencing quality degradation may be excluded, and traffic can be redirected to an alternate routing plane designated as backup.

The traffic can be classified based on DSCP marking to distinguish the collectives it belongs to.

## 9. IANA Considerations

TBD

## 10. Acknowledgements

The authors would like to thank Jeffrey Haas, Zhaohui(Jeffrey) Zhang, Kevin Wang and Ron Bonica for their valuable feedback.

## 11. References

### 11.1. Normative References

[RFC4271] Rekhter, Y., Ed., Li, T., Ed., and S. Hares, Ed., "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, DOI 10.17487/RFC4271, January 2006, <https://www.rfc-editor.org/info/rfc4271>.

[RFC9350] Psenak, P., Ed., Hegde, S., Filsfils, C., Talaulikar, K., and A. Gulko, "IGP Flexible Algorithm", RFC 9350, DOI 10.17487/RFC9350, February 2023, <https://www.rfc-editor.org/info/rfc9350>.

### 11.2. Informative References

[I-D.filsfils-srv6ops-srv6-ai-backend] Filsfils, C., Martin, C., Pillai, K., Camarillo, P., Abdelsalam, A., Tantsura, J., and K. Patel, "SRv6 for Deterministic Path Placement in AI Backends", Work in Progress, Internet-Draft, draft-filsfils-srv6ops-srv6-ai-backend-03, 2 March 2026, <https://datatracker.ietf.org/doc/html/draft-filsfils-srv6ops-srv6-ai-backend-03>.

[I-D.ietf-idr-bgp-generic-metric] Sangli, S. R., Hegde, S., Das, R., Decraene, B., Wen, B., Kozak, M., Dong, J., Jalil, L., and K. Talaulikar, "Accumulated Metric in NHC attribute", Work in Progress, Internet-Draft, draft-ietf-idr-bgp-generic-metric-02, 6 January 2026, <https://datatracker.ietf.org/doc/html/draft-ietf-idr-bgp-generic-metric-02>.

[I-D.ietf-idr-link-bandwidth] Mohapatra, P., Das, R., Satya, M. R., Krier, S., Szarecki, R. J., and A. Gattani, "BGP Link Bandwidth Extended Community", Work in Progress, Internet-Draft, draft-ietf-idr-link-bandwidth-24, 7 January 2026, <https://datatracker.ietf.org/doc/html/draft-ietf-idr-link-bandwidth-24>.

[I-D.wang-idr-dpf]
Wang, K., Styszynski, M., Lin, W., Subramaniam, M., Kampa, T., and D. Singh, "BGP Deterministic Path Forwarding (DPF)", Work in Progress, Internet-Draft, draft-wang-idr-dpf-00, 1 December 2025, <https://datatracker.ietf.org/doc/html/draft-wang-idr-dpf-00>.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/info/rfc2119>.

[RFC4360] Sangli, S., Tappan, D., and Y. Rekhter, "BGP Extended Communities Attribute", RFC 4360, DOI 10.17487/RFC4360, February 2006, <https://www.rfc-editor.org/info/rfc4360>.

[RFC5305] Li, T. and H. Smit, "IS-IS Extensions for Traffic Engineering", RFC 5305, DOI 10.17487/RFC5305, October 2008, <https://www.rfc-editor.org/info/rfc5305>.

[RFC5357] Hedayat, K., Krzanowski, R., Morton, A., Yum, K., and J. Babiarz, "A Two-Way Active Measurement Protocol (TWAMP)", RFC 5357, DOI 10.17487/RFC5357, October 2008, <https://www.rfc-editor.org/info/rfc5357>.

[RFC7938] Lapukhov, P., Premji, A., and J. Mitchell, Ed., "Use of BGP for Routing in Large-Scale Data Centers", RFC 7938, DOI 10.17487/RFC7938, August 2016, <https://www.rfc-editor.org/info/rfc7938>.

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <https://www.rfc-editor.org/info/rfc8174>.

## Authors' Addresses

Srihari Sangli
HPE
Mahadevapura
Bangalore, KA 560048
India
Email: srihari.sangli@hpe.com

Shraddha Hegde
HPE
Mahadevapura
Bangalore, KA 560048
India
Email: shraddha.hegde@hpe.com

Michal Styszynski
HPE
France
Email: mlstyszynski@juniper.net

Mohan Nanduri
Microsoft
USA
Email: mohannanduri@microsoft.com