IDR
Internet-Draft
Intended status: Standards Track
Expires: 4 June 2026

K. Wang
M. Styszynski
W. Lin
M. Subramaniam
HPE
T. Kampa
Audi
D. Singh
Oracle Cloud Infrastructure

1 December 2025

BGP Deterministic Path Forwarding (DPF)
draft-wang-idr-dpf-00

Abstract

Modern data center (DC) fabrics typically employ Clos topologies with External BGP (EBGP) for plain IPv4/IPv6 routing. While hop-by-hop EBGP routing is simple and scalable, it provides only a single best-effort forwarding service for all types of traffic. This single best-effort service might be insufficient for increasingly diverse traffic requirements in modern DC environments. For example, loss and latency sensitive AI/ML flows may demand stronger Service Level Agreements (SLA) than general purpose traffic. Duplication schemes which are standardized through protocols such as Parallel Redundancy Protocol (PRP) require disjoint forwarding paths to avoid single points of failure. Congestion avoidance may require more deterministic forwarding behavior.

This document introduces BGP Deterministic Path Forwarding (DPF), a mechanism that partitions the physical fabric into multiple logical fabrics. Flows can be mapped to different logical fabrics based on their specific requirements, enabling deterministic forwarding behavior within the data center.

Copyright Notice

Table of Contents

1. Introduction

Modern data center (DC) fabrics typically employ Clos topologies with External BGP (EBGP) [RFC7938] for plain IPv4/IPv6 routing. While hop-by-hop EBGP routing is simple and scalable, it provides only a single best-effort forwarding service for all types of traffic. This single best-effort service might be insufficient for increasingly diverse traffic requirements in modern DC environments. For example, loss and latency sensitive AI/ML flows may demand stronger Service Level Agreements (SLAs) than general purpose traffic. Duplication schemes which are standardized through protocols such as Parallel Redundancy Protocol (PRP) [IEC62439-3] require disjoint forwarding paths to avoid single points of failure. Congestion avoidance may require more deterministic forwarding behavior.

Traditionally, traffic engineering requirements like these can be served using technologies like RSVP-TE [RFC3209] or Segment Routing [RFC8402] in MPLS networks. However, according to the reasons stated in [RFC7938], modern data centers mostly use IP routing with EBGP as their sole routing protocol. BGP DPF is a lightweight traffic engineering alternative designed specifically for the IP Clos fabrics with EBGP as the routing protocol. It partitions the physical fabric into multiple logical fabrics by coloring the EBGP sessions running on the fabric links. Routes are also colored so that they are only advertised and received over the matching colored EBGP sessions. Together, they provide a certain level of deterministic forwarding behavior for the flows to satisfy the diverse traffic requirements of today's data centers.

1.1. 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.

2. BGP DPF

BGP DPF use BGP session coloring and route coloring to direct flows to different logical fabrics.

2.1. BGP Session Coloring

Figure 1 shows how a physical fabric is partitioned into two logical fabrics, the red fabric and the blue fabric. Leaf1 and Leaf2 can communicate using the red fabric via Spine1, or using the blue fabric via Spine2. Link Spine1-Leaf1 and Spine1-Leaf2 belong to the red fabric and link Spine2-Leaf1, Spine2-Leaf2 belong to the blue fabric. Instead of coloring the links directly, BGP DPF colors the EBGP sessions running on the corresponding links. The color of an EBGP session is configured on both ends separately, using the Color Extended Community as defined in Section 4.3 of [RFC9012].

There are two modes for session coloring, the strict mode and the loose mode. In the strict mode, the EBGP session MUST NOT come to Established state unless both ends are configured with the same color. In the loose mode, mismatched colors on both ends of an EBGP session SHALL NOT prevent the session from coming up.

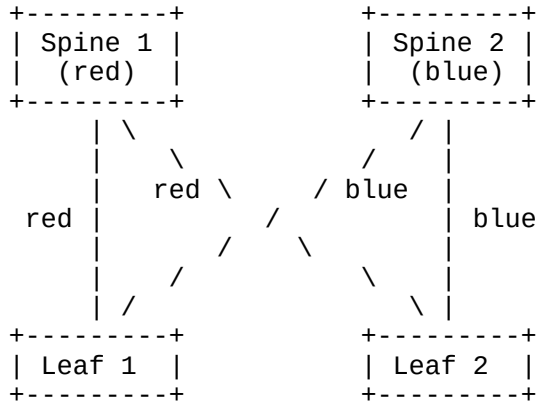

Figure 1: Divide one physical fabric into two logical fabrics

2.1.1. Strict Mode

When running in the strict session coloring mode, a BGP speaker uses the Capability Advertisement procedures from [RFC5492] to determine whether the color configured locally matches the color configured on the remote end. When a color is configured for an EBGP session locally, the BGP speaker sends the SESSION-COLOR capability in the OPEN message. The fields in the Capability Optional Parameter are set as follows. The Capability Code field is set as TBD. The Capability Length field is set as 4. The Capability Value field is set as the 4-octet Color Value of the Color Extended Community, as defined in Section 4.3 of [RFC9012]. Note, even though the BGP session is colored using a Color Extended Community, the only field

useful is the Color Value of the Color Extended Community. The Flags field is ignored. That is why only the 4-octect Color Value is included in the SESSION-COLOR Capability. The SESSION-COLOR capability format is shown in Figure 2:

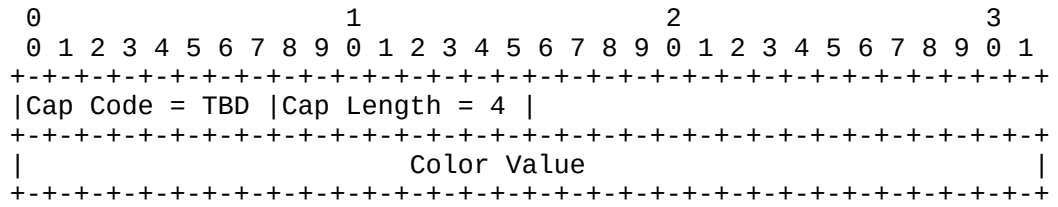

Figure 2: SESSION-COLOR Capability

When receiving the OPEN message for an EBGP session, the BGP speaker matches the SESSION-COLOR capability against its locally configured session color. Session color is considered as a match for one of the following conditions:

No color on both ends:
The receive OPEN message has no SESSION-COLOR capability and the EBGP session is not configured with a color.

Same color on both ends:
The received OPEN message has SESSION-COLOR capability and its color is the same as the session color configured locally for the EBGP session.

All other cases MUST be considered as session color mismatch. When a session color mismatch is detected, the BGP speaker MUST reject the session by sending a Color Mismatch Notification (code 2, subcode TBD) to the peer BGP speaker.

2.1.2. Loose Mode

The strict session coloring mode ensures that an Established EBGP session must have matching session colors on both ends. It helps to detects the color misconfigurations earlier. However, exchanging session colors through a Capability in BGP OPEN message requires BGP session flaps whenever session colors are changed. To address this session flap issue, the loose session coloring mode is introduced. When running the loose session coloring mode, session colors are not carried in the BGP OPEN message therefore change of the session color will not lead to the session flap. In this case, if the colors configured on both ends of the EBGP session mismatch, the routes received over the session will only match the color of the remote end but mismatch the color of the local end, as described in Section 2.2.

A route received with mismatched color MUST NOT be accepted.

[I-D.ietf-idr-dynamic-cap] allows Capabilities to be exchanged without flapping the session. That might allow us to gradually phase out the Loose Mode once dynamic capability is widely deployed.

2.2. Route Coloring

Once the EBGP sessions are colored accordingly, the physical fabric is partitioned into multiple logical fabrics. Routes can also be colored at the egress leaves to indicate which EBGP sessions (or which logical fabrics) they should be advertised over.

2.2.1. Route Coloring at the Egress Leaf

There are several ways to color a route at an egress leaf:

One color:
: When a route is configured with one color at the egress leaf, it is advertised over the same colored or uncolored EBGP sessions, with the corresponding Color Extended Community attached. This is the easiest way to make use of the logical fabrics.

One primary color and one backup color:
: When a route is configured with one primary color and one backup color at the egress leaf, it is advertised over the EBGP sessions of the primary color, with the primary Color Extended Community and an AIGP metric [RFC7311] of value zero. It is also advertised over the EBGP sessions of the backup color, with the backup Color Extended Community. In case there are uncolored sessions, the route is also advertised over the uncolored sessions, without Color Extended Community. The AIGP metric will help the receiving node to identify the primary colored paths. This allows traffic to fall back to the backup logical fabric when the primary logical fabric fails.

One primary color and all-colors as backup colors:
: When a route is configured with one primary color and all-colors as backup colors at the egress leaf, it is advertised over the EBGP sessions of the primary color, with the primary Color Extended Community and an AIGP metric of value 0. It is also advertised over the EBGP sessions of all other colors, with the Color Extended Community same as the corresponding session color. In case there are uncolored sessions, the route is also advertised over the uncolored sessions, without Color Extended Community. The AIGP metric will help the receiving nodes to identify the primary colored paths. By specifying all-colors as backup colors, traffic can be spread over all remaining logical fabrics when the

primary fabric fails. In the single backup color approach, traffic from the failed primary logical fabric might congest the backup fabric. By spreading the failed primary logical fabric traffic to all backup logical fabrics, the chance of congestion on the backup logical fabrics will be significantly reduced.

All-colors:
When a route is configured with all-colors at the egress leaf, it is advertised over the EBGP sessions with any color, with the Color Extended Community same as the corresponding session color. In case there are uncolored sessions, the route is also advertised over the uncolored sessions, without Color Extended Community. This allows the ingress router to map different flows of the route to different logical fabrics.

No color:
An uncolored route from the egress leaf can be advertised over EBGP sessions with any color or no color. It is advertised without Color Extended Community. Uncolored routes could be useful to carry routing protocol PDUs which do not use much bandwidth but needs to be sent over any links regardless of the logical fabrics.

Since AIGP metric is used in the primary/backup color cases, it is expected that all BGP speakers MUST support AIGP if we need DPF primary/backup protection.

2.2.2. Color Matching at the Spine and Super Spine

At the transit nodes (Spines or Super Spines), the Color Extended Community of the route is used to match against the EBGP session color to decide whether the route should be advertised over the session:

Advertising over an uncolored EBGP session: If the session is uncolored, the route is re-advertised following the existing route advertisement rules defined in [RFC4271].

Advertising over a colored BGP session: If the active route has no Color Extended Community or a Color Extended Community which is the same as the session color, then the active route is advertised over the session. If the active route has a Color Extended Community mismatching the session color, then check whether there is an inactive route with a Color Extended Community matching the session color. If yes, advertise the active route to the session, except that the AIGP attributed (if any) MUST be stripped and the Extended Color Community MUST be replaced with the session’s Color Extended Community. Otherwise, don’t advertise the route.

Matching the session color against the inactive routes is necessary because a backup route needs to be re-advertised to the backup fabric. So, when a packet arrives from the backup fabric, it is forwarded over the primary fabric to the destination, unless the primary fabric is down.

2.2.3. Flow Mapping at the Ingress Leaf

At the ingress leaf, flows can be mapped to different logical fabrics based on the route coloring approaches from the egress leaf:

One color: When a route is configured with one color at the egress leaf, the ingress leaf will receive the route from the EBGP session(s) with that color only. Flows towards this destination will be mapped to the logical fabric of this color only.

One primary color and one backup color: When a route is configured with one color as primary color and one color as backup color at the egress leaf, the ingress leaf will receive the route from EBGP sessions of both the primary color and the backup color. The routes received from the primary color sessions will be preferred due to AIGP. The routes received from the backup color sessions can be used as the backup paths. Flows towards this destination will be mapped to the primary logical fabric. In case the primary logical fabric fails, flows towards this destination will be mapped to the backup logical fabrics. Note that fallback to the backup logical fabric could happen at the ingress leaf as well as the spines and super spines.

One primary color and all-colors as backup color: When a route is configured with one color as primary color and all-colors as backup color at the egress leaf, the ingress leaf will receive the route from EBGP sessions of all colors. The routes received from the primary color sessions will be preferred due to AIGP. The routed received from all other colored sessions can be used as backup paths. Flows towards this destination will be mapped to the primary logical fabric. In case the primary logical fabric fails, flows towards this destination will be mapped to all backup logical fabrics. Note fallback to backup logical fabrics could happen at the ingress leaf as well as the spines and super spines.

All colors: When a route is configured with all-colors at the egress

leaf, the ingress leaf will receive the route from EBGP session of all colors. The routes from all sessions can be used to forward traffic. The ingress leaf can map flows towards this destination to routes with different Color Extended Communities, using mechanisms such as the Access Control List (ACL) filter. The details of mapping different flows to different routes of the same destination is out of the scope of this document.

Apart from mapping IP flows as described above, the ingress leaf could also map VPN flows, such as EVPN-VXLAN flows, to different logical fabrics. For example, the egress leaf can advertise multiple VXLAN tunnel endpoint routes, each with its own color. When a VXLAN tunnel endpoint is chosen for a MAC VRF at the ingress leaf, flows of that MAC VRF will be mapped to the logical fabric corresponding to the color of the tunnel endpoint route.

3. Use Cases

The most common use cases related to the BGP-DPF are:

* AI/ML backend training DC networks
* AI/ML frontend DC Inference networks
* IP Storage networks
* DCI - Data Center Interconnect
* Industrial hybrid DC/Campus networks

3.1. AI/ML backend training Data Center network

In the context of the AI/ML data centers (DC), especially where the training of LLM (Large Language Models) is the primary goal, there might be some challenges with the traditional IP ECMP packet spraying, such as sending the packets in an unordered manner due to the way load balancing is performed or maintaining consistency of performance between different phases of the job executions. AI/ML training in a data center refers to the process of utilizing large-scale computing infrastructure to train machine learning models on massive datasets. This process can take weeks or sometimes months for larger models. LLM training is taking place in DCs with GPU-enabled servers interconnected in the Rail Optimized Design within the IP Clos scale-out fabrics. In such architectures, every GPU of the server is linked to a 400G/800G NIC card, which connects to a different ToR (Top of Rack) leaf Ethernet switch node. The typical AI training server uses eight GPUs, so each server requires eight NIC cards, each connecting to a different ToR. A typical Rail is based

on eight 400G/800G/1.6Tbps switches, and rail-to-rail communication between strips is achieved through multiple spine nodes (typically 32 or more).

The transport used by the GPU servers between the rails or within the rail is either based on ROCEv2, or UEC transport (UET) in the future. The number of these flows per GPU/NIC is sometimes limited. A single ROCEv2 flow can utilize a massive bandwidth, and the characteristics of the flows may have very low entropy - the same source UDP and destination UDP are used by the ROCEv2 transport between the GPU servers during the given Job-ID. This may lead to short-term congestion at the spines, triggering the DCQCN reactive congestion control in the AI/DC fabric, with the PFC (Priority Flow Control) and ECN (Explicit Congestion Notification) mechanisms activated to prevent frame loss. Consequently, these mechanisms slow down the AI/ML session by temporarily reducing the rate at the source GPU server and extending the time needed to complete the given Job-ID. If congestion persists, frame loss may also occur, and the given Job-ID may need to be restarted to be synced across all GPUs participating in the collective communication. With packet spraying techniques or flow-based Dynamic Load Balancing, this is a less common situation in a well-designed Ethernet/IP fabric, but the GPU servers NIC cards must support the Out Of Order delivery. Additionally, it may still reduce performance or cause instability between Job-IDs or between tenants connected to the same AI/DC fabric.

This is where deterministic path pinning-based load balancing of flows can be applied, and where the BGP-DPF can be utilized to color the paths of a given tenant or a specific AI/ML workload, controlling how these paths are used. When the given ROCEv2 traffic is identified through the destination QPAIR in the BTH header at the ToR Ethernet switch, it can be allocated to a specific DPF color ID using ingress enforcement rules or TCAM flow awareness at the ASIC level. The AI/ML flows can be load-balanced across different DPF fabric color IDs and remain on the specified fabric color for the duration of the AI/ML Job. Thanks to that, not only does the given AI workload get a dedicated fabric color ID, but it also becomes isolated from the other AI workloads, which offers more predictable performance results (consistent tail latency and same Job Completion Time (JCT)) when compared to packet spraying based load balancing across all of the IP ECMP paths.

In this case, the probability of encountering congestion is also lower, as the given workload is assigned a dedicated path and is not competing with other AI workloads. When pinning the AI workload to a specific path, this means that there will be no packet reordering at the destination/target server, as the ROCEv2/UET packets will follow the same path from the beginning to the end of the given session.

The Rail Optimized Design shown in Figure 3 may also run two LLM training sessions simultaneously from two different tenants. This is also where IP path diversity of the DPF comes into play - by simply coloring the two workloads from the two LLMs, we can forward them across a different set of spine switches.

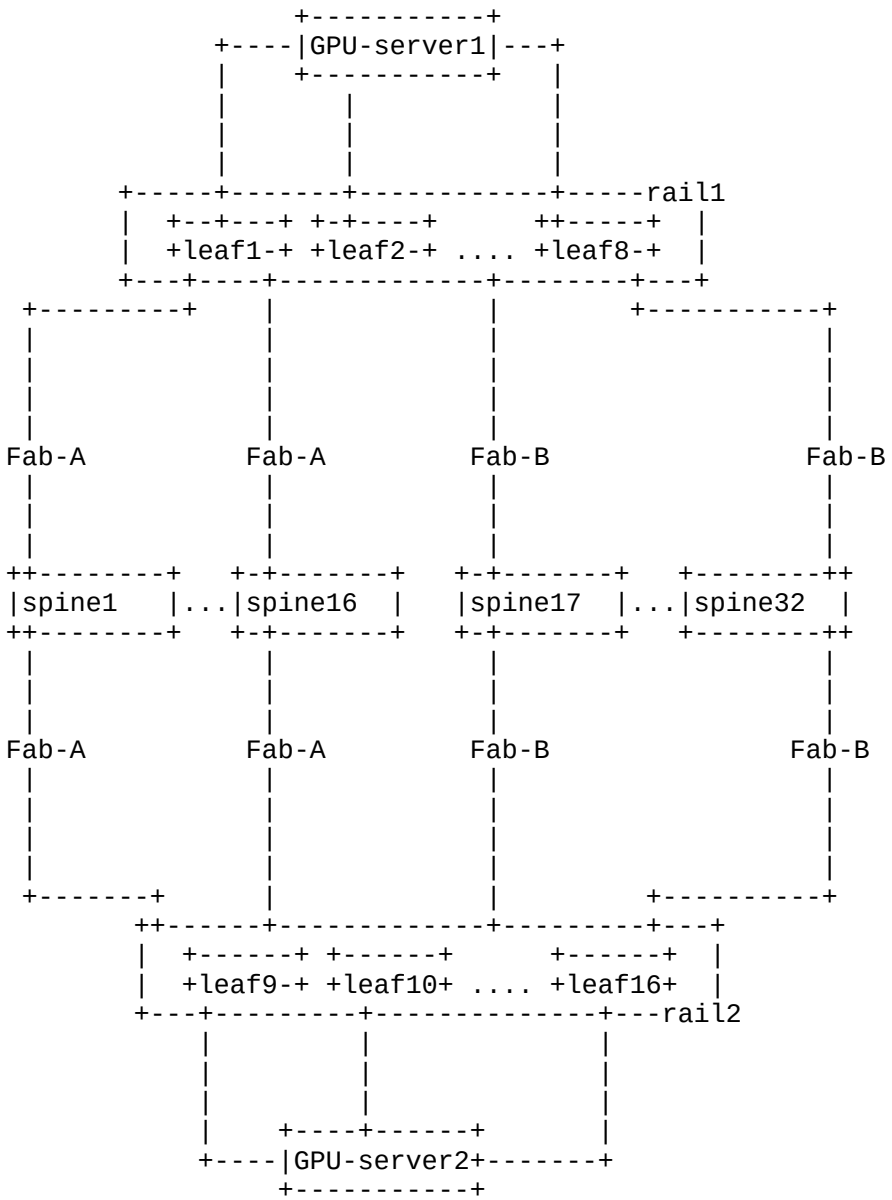

Figure 3: AI/ML backend training Data Center network

For example, 16 spines are allocated to the LLM-A training, and the other 16 spines are mapped to the LLM-B. Within each group of colored spines, IP ECMP with Dynamic Load Balancing can still operate on a per-flow or per-packet basis. Each tenant LLM with this approach receives half of the fabric's capacity of the fabric, and if required, this can be adjusted to be reduced or increased. The given fabric color fab-A and fab-B can be also allocated to the tenants enabled with EVPN-VXLAN overlays.

In summary, using BGP-DPF in backend DC network could achieve:

* Predictable and more efficient load balancing of the AI/ML workloads with the path pinning (for example, the ROCEv2 Op Code-based pinning or the destination ROCEv2 QPAIR-based path pinning in case of the ROCEv2 traffic)

* Isolation of the tenants inside the larger-scale AI/ML IP Clos fabric

* Consistency of the performances and faster AI workload ramp time

* Eliminated or highly reduced utilization of the PFC/ECN in the lossless fabric

3.2. AI/ML frontend DC and the Inference network

In the context of an AI/ML data center, an inference network refers to the computing infrastructure and networking components optimized for running already trained machine learning models (inference) at scale. Its primary purpose is to deliver low-latency, high-throughput predictions for both real-time and batch workloads. ChatGPT is a large-scale inference application deployed in a data center environment that utilizes real-time data. Still, it employs a generative AI model, such as GPT, which has been trained for several weeks in the training domain, as explained in Section 3.1 above.

The reason we mention it is that in many cases, cloud or service providers will run inferences in parallel for multiple customers simultaneously. Multi-tenancy is likely to be used at the network level - for example, utilizing EVPN-VXLAN-based tenant isolation in the leaf/spine/super-spine IP Clos fabric, or using MAC-VRFs or Pure RT5 IPVPN. In such cases, many inference applications can be enabled simultaneously within the same physical fabric. In some cases, the tenant/customer may request to be fully isolated from the other tenants, not only from a control plane perspective but also from a data plane perspective when forwarding traffic between the two ToR switches.

For example, the tenant-A and tenant-B may each be allocated to a different RT5 EVPN-VXLAN instance, and these instances are mapped to two different BGP-DPF color-ids. With this approach, the overlays of tenant A and tenant B will never overlap and will utilize different fabric spines. The outcomes here are that the latency, which is critical for inference applications, is also becoming more predictable if the fabric paths for the two tenants are different. The two overlays are more correlated with the underlay path. In some cases, with the explicit definition of the backup color ID at the BGP-DPF level, the fast convergence will become an additional outcome for the frontend EVPN-VXLAN fabrics.

3.3. IP Storage networks with Fab-A/Fab-B path diversity

In the context of the DC, storage networks are a key component of the infrastructure that manages and enables servers with scalable block or object storage systems. For block storage, such as NVMe-o-F (using NVMe-o-RDMA or NVMe-o-TCP), the Fab-A/Fab-B design is often used, where Fabric-A serves as the primary and Fabric-B as the backup path for performing read or write operations on the remote storage arrays. The given server inside the DC typically has dedicated storage NICs. For redundancy purposes, two NIC ports are generally used - one connected to Fab-A and another to Fab-B. As in the case of traditional storage, such as Fiber Channel(FC), the recommended approach is to make sure that the storage dedicated fabric supports complete path isolation. In case of failure, at least one of the two fabrics becomes available.

This is also where BGP DPF can help, by explicitly defining the IP Storage paths for Fab-A and Fab-B. Besides the storage redundancy aspect, the capacity planning is also essential here. After the failover from A to B, the same read and write capacity is offered to all IP Fabric-connected servers. Fab A/B offers 100% capacity in the event of failure, while all operations are managed at the logical level using the BGP DPF.

3.4. DCI - Data Center Interconnect

In case of critical applications, disaster recovery plans usually require a second availability zone for redundancy and resilience. Concepts foresee either the replication of persistent storage data, or to run the same application in parallel in a backup location, or to load balance across multiple DCs.

When replicating data or synchronizing the application state between two places, it is sometimes also necessary to isolate the paths across long-distance connectivity. If the connection between DC1 and DC2 use a mesh of links or partial mesh and the DCI connect solution uses EVPN-VXLAN or Pure IP connections, some workloads may require communication in a more deterministic way by correlating the underlay and overlay when both uses the BGP as IP routing protocol - one path may have better latency and jitter than the other when connecting between the two remote locations so the admin may decide to push one EVPN-VXLAN instance (MAC-VRF and/or RT5 IPVPN) through very well selected underlay path of the dark fiber connection. In this use case, we assume the DCI is using the underlay IP EBGP, and some links may be colored using the BGP-DPF. EVPN-VXLAN can use the EVPN-VXLAN to EVPN-VXLAN tunnel stitching [RFC7938], with the DCI underlay links colored by BGP-DPF as red and blue paths. Different MAC-VRFs and RT5 instances are assigned to various DPF colors to control the forwarding of the workloads between the two DC locations.

The outcome of this use case is that the DCI admin can anticipate failovers and allocate EVPN-VXLAN-connected workloads based on the capacity and performance (including latency and jitter) of the DCI links.

3.5. Industrial/factory hybrid DC/Campus networks

Industrial and factory automation is increasingly adopting distributed computing concepts to leverage the benefits of virtualization and containerization. This change often comes with a shift of application into a remote DC, which imposes stringent requirements on the networking infrastructure between DC and the respective process. These hybrid DC campus networks require a high level of resiliency against failures as certain applications tolerate zero loss of frames. Duplication schemes like PRP [IEC62439-3] are being leveraged in these scenarios to provide zero loss in face of failures but require disjoint paths to avoid any single point of failures.

When the Campus and DC fabrics utilize modern solutions, such as EVPN-VXLAN overlays, IP ECMP from leaf to spine is frequently employed. This might lead to PRP duplicates being forwarded across

the same spine and bring processes to a standstill in case of a spine maintenance or physical failure. That’s where the BGP-DPF based underlay network can guarantee that the EVPN-VXLAN overlays are always forwarded over their predefined nominal and backup paths, allowing for disjoint paths across the fabric. The primary and backup paths taken by PRP frames are well-defined, enabling fault-tolerant communication, i.e., between robots on the shop floor and control applications running on a distributed environment in the DC. With the PRP frames destined for LAN A and LAN B being sent through EVPN-VXLAN MAC-VRF-A and MAC-VRF-B, over diverse paths DPF color-A and DPF color-B, critical communication flows are being controlled in terms of forwarding and recovery for the deterministic behavior they require.

4. Operational Considerations

When routes are colored with both primary and backup colors at the egress leaf, we need to make sure the network is a strictly staged network to avoid potential routing and forwarding loops. A strictly staged network ensures that packet always goes to the next stage and never come back. In the Clos topology with EBGP, staged routing is guaranteed by configuring the same AS number on the spines and super spines in the same stage. Only leaves have unique AS numbers.

5. IANA Considerations

A new BGP Capability will be requested from the "Capability Codes" registry within the "IETF Review" range [RFC5492].

A new OPEN Message Error subcode named "Color mismatch" will be requested from the "OPEN Message Error subcodes" registry.

6. Security Considerations

Modifying Color Extended Community of a BGP UPDATE message by an attacker could potentially cause the routes to be advertised to the unintended logical fabrics. This could potentially lead to failed or suboptimal routing.

7. References

7.1. Normative References

[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>.

[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>.

[RFC5492] Scudder, J. and R. Chandra, "Capabilities Advertisement with BGP-4", RFC 5492, DOI 10.17487/RFC5492, February 2009, <https://www.rfc-editor.org/info/rfc5492>.

[RFC9012] Patel, K., Van de Velde, G., Sangli, S., and J. Scudder, "The BGP Tunnel Encapsulation Attribute", RFC 9012, DOI 10.17487/RFC9012, April 2021, <https://www.rfc-editor.org/info/rfc9012>.

[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>.

[RFC7311] Mohapatra, P., Fernando, R., Rosen, E., and J. Uttaro, "The Accumulated IGP Metric Attribute for BGP", RFC 7311, DOI 10.17487/RFC7311, August 2014, <https://www.rfc-editor.org/info/rfc7311>.

7.2. Informative References

[RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, DOI 10.17487/RFC3209, December 2001, <https://www.rfc-editor.org/info/rfc3209>.

[RFC8402] Filsfils, C., Ed., Previdi, S., Ed., Ginsberg, L., Decraene, B., Litkowski, S., and R. Shakir, "Segment Routing Architecture", RFC 8402, DOI 10.17487/RFC8402, July 2018, <https://www.rfc-editor.org/info/rfc8402>.

[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>.

[I-D.ietf-idr-dynamic-cap] Chen, E. and S. R. Sangli, "Dynamic Capability for BGP-4", Work in Progress, Internet-Draft, draft-ietf-idr-dynamic-cap-17, 6 July 2025, <https://datatracker.ietf.org/doc/html/draft-ietf-idr-dynamic-cap-17>.

[IEC62439-3]
International Electrotechnical Commission, "Industrial communication networks High availability automation networks Part 3: Parallel Redundancy Protocol (PRP) and High-availability Seamless Redundancy (HSR)", IEC 62439-3:2016, 2016.

Appendix A. Alternative Solutions

An alternative way to achieve part of the BGP DPF functionalities is to use BGP export and import policies. Instead of coloring the EBGP sessions and routes, one could choose to use export policies to specify which session(s) a route should be advertised. On the receiving side, one could also choose to use import policies to ensure a route is only received from certain EBGP sessions. The alternative approach is not chosen due to the following factors:

* The policy configurations have to be done on each nodes and might need to change when new routes are added.

* Policy configurations are less intuitive than session coloring and could be prone to configuration mistakes.

* Certain functionalities in DPF, like the primary and backup logical fabrics, might not be achievable using popular policies.

Acknowledgements

TBD.

Contributors

Jeffrey Haas
HPE
Email: jeffrey.haas@hpe.com

Authors' Addresses

Kevin Wang
HPE
Email: kevin.wang@hpe.com

Michal Styszynski
HPE
Email: mlstyszynski@juniper.net

Wen Lin
HPE
Email: wen.lin@hpe.com

Mahesh Subramaniam
HPE
Email: mahesh-kumar.subramaniam@hpe.com

Thomas Kampa
Audi
Email: thomas.kampa@audi.de

Diptanshu Singh
Oracle Cloud Infrastructure
Email: diptanshu.singh@oracle.com