Network Working Group
Internet-Draft
Intended status: Informational
Expires: 3 September 2026

F. Clad, Ed.
C. Filsfils
Cisco Systems, Inc.
R. Jiang
D. Cai
Alibaba

2 March 2026

IP Fast Reroute for AI/ML Fabrics
draft-clad-rtgwg-ipfrr-aiml-00

Abstract

This document describes the requirements and mechanisms for achieving sub-100 microsecond convergence in Artificial Intelligence (AI) Data Center (DC) fabrics and Data Center Interconnect (DCI) environments. It explores the limitations of current IP Fast Reroute (RFC 5714) capabilities, such as ECMP, LFA, and TI-LFA, particularly in the context of large-scale, multi-tier Clos topologies and BGP-only fabrics. The draft highlights the requirements for hardware-accelerated network notification mechanisms and congestion-aware remote protection strategies to address the stringent performance demands of AI workloads.

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at https://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on 3 September 2026.

Copyright Notice

Table of Contents

1. Introduction 2
2. AI/ML Networks 3
2.1. Scale-Out Networks 3
2.1.1. Topology Architecture 3
2.1.2. ECMP Paths 4
2.1.3. Resiliency Mechanisms 6
2.2. Scale-Across Networks 6
2.2.1. Deployment Models and Topology Characteristics . . . 7
2.2.2. Resiliency Mechanisms 7
3. Limitations of Existing Resiliency Mechanisms 8
3.1. CPU-Based Activation Latency 8
3.2. Limited Scope of ECMP Protection 9
3.3. Binary Failure Model 9
3.4. Inefficient Local Repair Paths 10
4. Requirements for Enhanced Protection in AI/ML Fabrics 11
4.1. Hardware-Accelerated Protection Activation 11
4.2. Complete Topology Visibility 11
4.3. Hardware-Accelerated Network Notifications 12
4.4. Quality-Aware Remote Protection 12
5. Security Considerations 13
6. IANA Considerations 13
7. Informative References 13
Acknowledgements 15
Authors' Addresses 15

1. Introduction

AI training and inference workloads are characterized by high-bandwidth, synchronized "all-to-all" communication patterns ([Gangidi2024], [Qian2024]). These workloads are extremely sensitive to packet loss and jitter. In modern AI DC fabrics, the target convergence time for network failures is increasingly moving toward the sub-100 microsecond threshold.

Traditional network recovery mechanisms operate at two distinct timescales. Control-plane convergence, where routing protocols (BGP, IS-IS) recompute paths and update the Forwarding Information Base

(FIB), typically operates in the sub-second range. IP Fast Reroute mechanisms such as ECMP, LFA, and TI-LFA ([RFC5714]) can reduce this to the tens of milliseconds by pre-computing backup paths in the forwarding plane. However, even these mechanisms rely on CPU-based activation: when a failure is detected, an interrupt is processed by the line-card CPU to trigger FIB updates and activate the backup path. This CPU-mediated path introduces activation delays in the range of 10-50 milliseconds, which is still two orders of magnitude slower than the sub-100 microsecond target required for AI workloads.

This document discusses the transition toward hardware-accelerated notification and protection mechanisms that operate entirely within the forwarding plane to meet these stringent requirements.

While this document focuses on AI/ML backend networks given their particularly stringent convergence requirements, the mechanisms and benefits described herein are equally applicable to AI/ML frontend networks (serving inference requests), general data center networks, and data center interconnects for other workloads. Any environment with high-bandwidth, loss-sensitive traffic patterns can benefit from hardware-accelerated protection and sub-millisecond convergence, though AI training workloads represent the most demanding use case driving these architectural changes.

2. AI/ML Networks

2.1. Scale-Out Networks

Scale-out networks for AI and Machine Learning backends represent a specialized class of data center fabric designed to support the extreme communication demands of distributed training and large-scale inference workloads. These networks are characterized by all-to-all connectivity patterns where every compute node must communicate with every other node, often with synchronized, barrier-based communication patterns.

Scale-out AI networks typically employ BGP as the only routing protocol for disseminating reachability information and computing forwarding paths across the fabric ([RFC7938]).

2.1.1. Topology Architecture

AI backend networks typically employ a multi-tier folded Clos topology ([Clos53], [AlFares2008], [Greenberg2009]), most commonly implemented as a 2-tier or 3-tier architecture:

* *Leaf Layer*: Direct connection to compute nodes (GPUs). Each leaf switch connects to multiple compute nodes and uplinks to spine switches.

* *Spine Layer(s)*: Aggregation points that provide connectivity between leaf switches. In a 2-tier design, spines directly interconnect leaves. In a 3-tier design, an additional "super-spine" or "core" layer provides further aggregation.

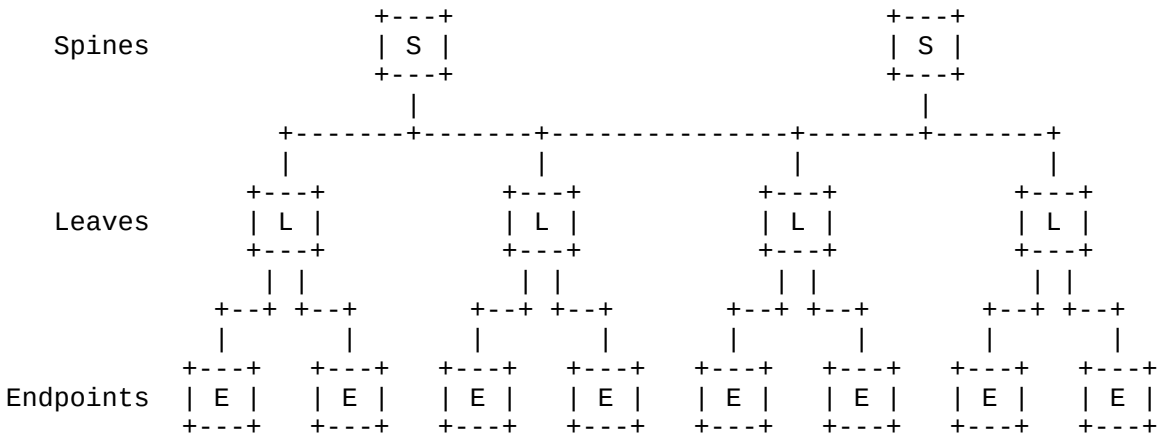

Figure 1: Example 2-tier folded Clos topology

For higher-scale deployments, multi-plane and multi-rail extensions are employed to increase bisection bandwidth and scaling capacity. In multi-plane architectures, each compute node connects to multiple independent fabric planes that operate in parallel, effectively multiplying the network capacity by the number of planes. In multi-rail designs, each compute node in a rack connects to an independent fabric rail, multiplying the endpoint scale by the number of rails at the cost of losing direct connectivity between nodes on different rails.

2.1.2. ECMP Paths

The set of ECMP paths between any two compute nodes in a k-ary 2- or 3-tier folded Clos fabric are determined by the location of the source and destination nodes within the topology.

For a source and destination node connected to the same leaf switch, there is a single path through that common leaf.

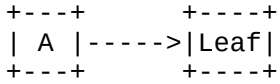

Figure 2: Path between nodes on the same leaf switch

For a source and destination node connected to different leaf switches in the same pod, the source leaf switch may select any of the k spine switches in that pod, resulting in k ECMP paths:

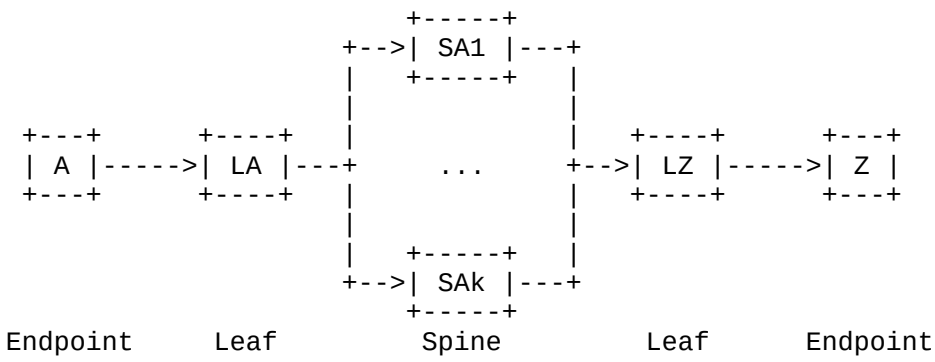

Figure 3: Paths between nodes on different leaf switches in the same pod

For a source and destination node connected to different leaf switches in different pods, the source leaf switch may select any of the k spine switches in its pod, and each of those spine switches may again select any of the k super-spine switches in the core layer, resulting in k^2 ECMP paths.

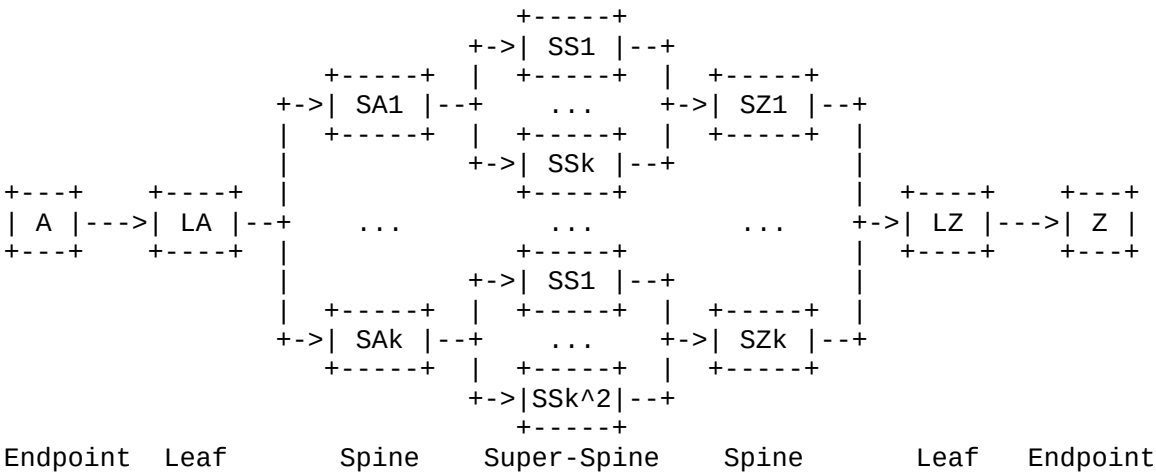

Figure 4: Paths between nodes in different pods

Notably, the ECMP path diversity is exclusively in the upward direction, from leaf to spine and from spine to super-spine. Once traffic reaches the highest tier switch on the path, there is only a single downward path to the destination leaf and compute node. This asymmetry in path diversity is a fundamental characteristic of folded Clos fabrics and has significant implications for failure modes and fast reroute strategies.

2.1.3. Resiliency Mechanisms

In scale-out fabrics, Equal-Cost Multi-Path (ECMP) is the primary mechanism for resiliency. When a local link fails, the Forwarding Information Base (FIB) is updated to remove the failed next-hop.

Traffic that was using the failed path can be handled in one of two ways: it can be shifted entirely onto a single backup ECMP member path, or it can be redistributed and load-balanced across all surviving ECMP member paths. The latter approach spreads the impact of the failure across multiple paths, reducing the likelihood of overloading any single backup path, but is also more complex to implement.

In addition, ECMP-based resiliency is fundamentally limited to failures occurring in the upward direction of the path—the first half of the route through the fabric where multiple paths exist. As described in Section 2.1.2, once traffic reaches the highest tier switch (spine or super-spine) on its path, there is only a single downward path to the destination leaf. A failure occurring in this second half of the path cannot be locally protected by ECMP because the node detecting the failure has no alternate ECMP paths available.

2.2. Scale-Across Networks

Scale-across networks interconnect multiple AI scale-out fabrics, either within the same data center campus, across geographically distributed data centers, or both. These networks enable distributed training and inference at mega-scale, allowing AI workloads to span hundreds of thousands of GPUs across multiple independent clusters. Scale-across networks operate at the boundaries between scale-out fabrics and represent a hybrid environment combining elements of both traditional DC networking and WAN characteristics.

Scale-across networks employ either link-state IGP (e.g., IS-IS) or BGP ([RFC7938]) as the routing protocol, depending on the deployment scenario and operational preferences.

2.2.1. Deployment Models and Topology Characteristics

Scale-across networks exist in several deployment scenarios, each with characteristic topologies and path properties:

* *Campus-scale*: Multiple clusters within the same data center campus, typically connected via dedicated high-bandwidth trunks with latency in the single-digit microseconds to tens of microseconds range. These connections may span multiple buildings or co-located facilities with diverse physical routing. Campus-scale deployments often employ full-mesh or partial-mesh topologies where every cluster gateway connects directly to every other cluster gateway (full-mesh) or connects through a central hub (partial-mesh with hub). These topologies provide maximum path diversity and lowest latency, with forwarding typically static or based on simple multi-path load balancing rather than dynamic routing protocols.

* *Metro/Regional*: Clusters distributed across metropolitan areas or regional data centers, with latencies typically in the hundreds of microseconds to low milliseconds. These deployments may leverage Dark Fiber or dedicated wavelengths to minimize latency. Topologies are often constrained by fiber routing and may form ring or chain topologies representing physical fiber routes, or may employ partial-mesh configurations with hub aggregation points at central locations.

* *Wide-area*: Clusters in geographically distant data centers, spanning multiple regions or countries. Latencies are typically in the milliseconds range. Wide-area topologies are inherently irregular and arbitrary, shaped by geographic constraints, business relationships, fiber availability, and historical infrastructure decisions. Connectivity between clusters varies significantly, with no guaranteed symmetry: some cluster pairs may have direct links, while others may be connected through multiple intermediate aggregation points. The number of available paths, path lengths, and path costs are highly uneven across different cluster pairs.

2.2.2. Resiliency Mechanisms

Scale-across networks employ a combination of resiliency mechanisms depending on the underlying routing protocol and topology characteristics.

ECMP: When multiple equal-cost paths exist to the destination node, ECMP provides the same basic protection as in scale-out fabrics. Upon detecting a local link failure, traffic is shifted to a backup

ECMP member or redistributed across all surviving paths. However, unlike the regular Clos topology of scale-out fabrics, the irregular nature of scale-across topologies means that equal-cost multipath opportunities are less predictable and may not exist for all source-destination pairs.

LFA (Loop-Free Alternates): In deployments using link-state IGPs, LFA ([RFC5286]) can provide pre-computed backup paths for link or node failures. LFA identifies alternate next-hops that guarantee loop-free forwarding without requiring the IGP to reconverge. However, LFA coverage is topology-dependent and may not provide 100% protection, particularly in irregular topologies with limited connectivity between nodes. Where LFA coverage gaps exist, traffic may be dropped until the control plane converges.

TI-LFA (Topology-Independent LFA): TI-LFA ([RFC9855]) utilizes Segment Routing to provide 100% protection coverage regardless of topology. TI-LFA can steer traffic around failures even when no naturally loop-free alternate exists, by encoding a repair path as a segment list. While this ensures full protection, TI-LFA may occasionally result in suboptimal "hairpin" routing where traffic has to traverse an upstream node on its way to a safe release node in the Q-Space of the destination (see Section 3.4).

3. Limitations of Existing Resiliency Mechanisms

The resiliency mechanisms described in Section 2.1.3 and Section 2.2.2 face several fundamental limitations when applied to AI/ML workloads in scale-out and scale-across networks.

3.1. CPU-Based Activation Latency

Traditional fast reroute implementations rely on the line-card CPU to detect interface failures and trigger FIB updates. When a physical link fails and the failure is detected at the hardware level (e.g., loss of optical signal), this event generates an interrupt that is processed by the line-card CPU. The CPU then retrieves the appropriate backup forwarding state and programs the forwarding hardware accordingly. This CPU-mediated path introduces an activation delay typically in the range of 10-50 milliseconds, depending on the CPU load, software architecture, and implementation efficiency.

For AI training workloads with synchronized all-reduce operations and barrier synchronization, even a few milliseconds of packet loss can cause significant performance degradation. The target convergence time for modern AI fabrics is increasingly moving toward the sub-100 microsecond range, two orders of magnitude faster than CPU-based protection activation can provide.

3.2. Limited Scope of ECMP Protection

As described in Section 2.1.3, ECMP-based protection is fundamentally limited to local failures. In the context of a multi-tier folded Clos fabric, this means ECMP can only protect failures in the upward direction, the first half of the path where multiple equal-cost paths exist through different spine or super-spine switches.

Once traffic has traversed upward to the highest tier switch on its path (spine or super-spine), there is typically only a single downward link to the destination leaf switch. A failure on this downward segment cannot be protected by ECMP at the upstream node because no alternate paths are available at that point in the topology. The traffic is effectively dropped until the control plane converges and updates the FIB at nodes multiple hops upstream that still have path diversity.

Recent proposals such as [I-D.ietf-idr-next-next-hop-nodes] and [I-D.zzhang-rtgwg-router-info] attempt to provide visibility into two-hop failures by advertising information about next-next-hops and the accompanying network notifications. While these mechanisms can extend protection to failures one hop beyond the local node, they do not address the general case of failures occurring multiple hops away. In a 3-tier Clos fabric, traffic from endpoint A to endpoint Z may traverse LA -> SA1 -> SS1 -> SZ1 -> LZ. As is apparent in Figure 4, if the spine-to-leaf link SZ1-LZ fails, the adjacent node (SZ1) has no alternate downward path to LZ and cannot reroute locally; the node one hop away (SS1) also has no alternate path to reach LZ without using SZ1. The closest node with path diversity that can avoid the failed link is LA, which is three hops away from the failure. Similarly, in scale-across networks with arbitrary topologies, failures may occur at any depth in the network, well beyond the visibility provided by one extensions to BGP visibility.

3.3. Binary Failure Model

Existing fast reroute mechanisms — ECMP, LFA, and TI-LFA — operate on a binary failure model where a link or node is either "up" or "down." Upon detecting a complete failure, these mechanisms activate a pre-computed backup path or redistribute traffic across surviving ECMP members.

However, in AI fabrics, failures may have more nuanced impacts. A single member failure within a link bundle between two switches does not sever connectivity; it reduces available capacity by 1/k (where k is the number of bundle members). Traffic can still traverse the bundle, but the aggregate bandwidth on that hop is reduced by one member while the ECMP set remains unchanged.

While control-plane mechanisms exist to adjust load-balancing weights based on bandwidth changes, such as by leveraging the BGP link bandwidth extended community ([I-D.ietf-bess-ebgp-dmz]), these rely on control-plane signaling and operate on sub-second to second timescales. For AI workloads requiring sub-100 microsecond convergence, this control-plane latency is insufficient.

3.4. Inefficient Local Repair Paths

TI-LFA provides 100% protection coverage by steering traffic around failures using Segment Routing-encoded repair paths. However, the repair paths computed by TI-LFA are not always optimal from the perspective of the source node and may result in "hairpin" routing where traffic must traverse an upstream node before reaching a safe release point in the Q-Space (the set of nodes that can reach the destination while avoiding the failed link).

Consider a network where traffic from node A to node D normally flows through node B (A -> B -> D), and the link B-D fails. A TI-LFA repair path at B might send traffic back to A and from there to a node C in the Q-Space of D. This forces the traffic to go back and forth between A and B, creating a hairpin (one time loop) that consumes additional bandwidth on the A-B link and introduces additional latency compared to the more direct A -> C -> D path.

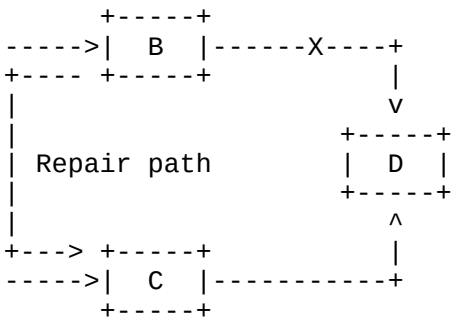

Figure 5: Hairpin for A -> D traffic on repair path from B to C

In the context of AI workloads, where traffic patterns are extremely bandwidth-sensitive and latency-sensitive, hairpin routing introduces multiple problems. First, the inefficient paths consume more bandwidth on potentially congested upstream links, potentially creating cascading congestion. Second, the additional latency introduced by hairpinning can disrupt the tight synchronization required by distributed AI training algorithms.

4. Requirements for Enhanced Protection in AI/ML Fabrics

To overcome the limitations described in Section 3, enhanced protection mechanisms for AI/ML fabrics should consider the following requirements.

4.1. Hardware-Accelerated Protection Activation

Traditional CPU-based protection with 10-50 millisecond activation delay is insufficient for AI workloads. Protection activation must occur entirely within the forwarding plane, using NPU-embedded logic.

Upon detecting a local link or interface failure, the hardware must immediately activate a pre-computed backup forwarding state without CPU intervention. This allows protection activation to occur in microseconds, with a target of sub-100 microseconds to meet the most stringent AI workload requirements.

The same hardware-accelerated activation principle applies to receiving network notifications (see Section 4.3). When a node receives a notification of a remote failure or capacity drop that impacts its forwarding paths, it should be able to adjust its forwarding state in hardware without waiting for CPU processing.

4.2. Complete Topology Visibility

Complete topology visibility is required to compute LFA and TI-LFA protection paths. The same is also required for any form of remote protection (ECMP, LFA, TI-LFA) at an arbitrary distance from the failure.

In networks running a link-state IGP, complete topology visibility is inherent to the protocol, as all nodes receive the full link-state database and can compute global shortest paths. However, in BGP-only networks, nodes typically do not have any visibility beyond their directly connected neighbors. Yet, complete topology visibility can easily be achieved in BGP-only networks ([RFC7938]) by incrementally enabling BGP Link-State (BGP-LS) without affecting normal BGP routing, as described in [I-D.ietf-idr-bgp-ls-bgp-only-fabric].

4.3. Hardware-Accelerated Network Notifications

A dedicated network notification mechanism is required to communicate failures, capacity changes, congestions, and other link performance degradation through the data plane in near real-time, allowing nodes at an arbitrary distance from the event to trigger appropriate remediation. When a node detecting a degradation generates and injects notification packets into the data plane, these packets must be able to traverse and be processed by any node in the network that requires corrective action, not limited to one or two hops away.

Beyond simple up/down signals, these notifications should convey quantitative information about the event impact— such as capacity reduction, congestion, or link quality degradation. This allows remote nodes to determine the scope of the event's impact on their own forwarding paths and make intelligent local protection decisions, rather than treating all events as failures.

The problem statement for those network notifications is discussed in [I-D.ietf-rtgwg-net-notif-ps].

4.4. Quality-Aware Remote Protection

Upon receiving a network notification, forwarding nodes must be able to adjust their forwarding state in real time based on the failure or quantitative information conveyed in the network notification. Rather than waiting for the control plane to converge, quality-aware protection allows a node to immediately trigger efficient remote protection actions in hardware within the sub-100 microsecond timeframe.

Quality-aware remote protection can consist of:

* *ECMP Weight Adjustment*: Adjust the load-balancing weights in the ECMP hash table to distribute traffic proportionally to remaining available capacity while staying within the existing ECMP set. This preserves the optimal path structure of the original topology.

* *Repair Path Outside ECMP Set*: When adjustment of weights alone cannot provide an appropriate remediation, activate one or more weighted repair paths that steer traffic toward the destination via alternate routes. These repair paths should be optimized to avoid any routing hairpin that would increase latency and bandwidth consumption.

5. Security Considerations

To be done.

6. IANA Considerations

This document does not require any IANA actions.

7. Informative References

[AlFares2008]
Al-Fares, M., Loukissas, A., and A. Vahdat, "A scalable, commodity data center network architecture", DOI 10.1145/1402946.1402967, August 2008, <https://doi.org/10.1145/1402946.1402967>.

[Clos53] Clos, C., "A study of non-blocking switching networks", DOI 10.1002/j.1538-7305.1953.tb01433.x, March 1953, <https://doi.org/10.1002/j.1538-7305.1953.tb01433.x>.

[Gangidi2024]
Gangidi, A., Miao, R., Zheng, S., Bondu, S. J., Goes, G., Morsy, H., Puri, R., Riftadi, M., Shetty, A. J., Yang, J., Zhang, S., Fernandez, M. J., Gandham, S., and H. Zeng, "RDMA over Ethernet for Distributed Training at Meta Scale", DOI 10.1145/3651890.3672233, August 2024, <https://doi.org/10.1145/3651890.3672233>.

[Greenberg2009]
Greenberg, A., Hamilton, J., Jain, N., Kandula, S., Kim, C., Lahiri, P., Maltz, D., Patel, P., and S. Sengupta, "VL2: a scalable and flexible data center network", DOI 10.1145/1592568.1592576, August 2009, <https://doi.org/10.1145/1592568.1592576>.

[I-D.ietf-bess-ebgp-dmz]
Litkowski, S., Satya, M. R., Vayner, A., Gattani, A., Kini, A., Tantsura, J., and R. Das, "BGP link bandwidth extended community use cases", Work in Progress, Internet-Draft, draft-ietf-bess-ebgp-dmz-09, 24 February 2026, <https://datatracker.ietf.org/doc/html/draft-ietf-bess-ebgp-dmz-09>.

[I-D.ietf-idr-bgp-ls-bgp-only-fabric]
Talaulikar, K., MahendraBabu, A. B., Filsfils, C., Ananthamurthy, K., Zandi, S., Dawra, G., and M. Durrani, "BGP Link-State Extensions for BGP-only Networks", Work in Progress, Internet-Draft, draft-ietf-idr-bgp-ls-bgp-only-

fabric-04, 20 October 2025,
<https://datatracker.ietf.org/doc/html/draft-ietf-idr-bgp-ls-bgp-only-fabric-04>.

[I-D.ietf-idr-next-next-hop-nodes]
Wang, K., Haas, J., Lin, C., and J. Tantsura, "BGP Next-next Hop Nodes", Work in Progress, Internet-Draft, draft-ietf-idr-next-next-hop-nodes-00, 20 October 2025,
<https://datatracker.ietf.org/doc/html/draft-ietf-idr-next-next-hop-nodes-00>.

[I-D.ietf-rtgwg-net-notif-ps]
Dong, J., McBride, M., Clad, F., Zhang, Z. J., Zhu, Y., Xu, X., Zhuang, R., Pang, R., Lu, H., Liu, Y., Contreras, L. M., Mehmet, D., and R. Rahman, "Fast Network Notifications Problem Statement", Work in Progress, Internet-Draft, draft-ietf-rtgwg-net-notif-ps-00, 11 February 2026, <https://datatracker.ietf.org/doc/html/draft-ietf-rtgwg-net-notif-ps-00>.

[I-D.zzhang-rtgwg-router-info]
Zhang, Z. J., Wang, K., Lin, C., Vaidya, N., Tantsura, J., and Y. Liu, "Advertising Router Information", Work in Progress, Internet-Draft, draft-zzhang-rtgwg-router-info-06, 23 February 2026,
<https://datatracker.ietf.org/doc/html/draft-zzhang-rtgwg-router-info-06>.

[Qian2024] Qian, K., Xi, Y., Cao, J., Gao, J., Xu, Y., Guan, Y., Fu, B., Shi, X., Zhu, F., Miao, R., Wang, C., Wang, P., Zhang, P., Zeng, X., Ruan, E., Yao, Z., Zhai, E., and D. Cai, "Alibaba HPN: A Data Center Network for Large Language Model Training", DOI 10.1145/3651890.3672265, August 2024,
<https://doi.org/10.1145/3651890.3672265>.

[RFC5286] Atlas, A., Ed. and A. Zinin, Ed., "Basic Specification for IP Fast Reroute: Loop-Free Alternates", RFC 5286, DOI 10.17487/RFC5286, September 2008,
<https://www.rfc-editor.org/rfc/rfc5286>.

[RFC5714] Shand, M. and S. Bryant, "IP Fast Reroute Framework", RFC 5714, DOI 10.17487/RFC5714, January 2010,
<https://www.rfc-editor.org/rfc/rfc5714>.

[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/rfc/rfc7938>.

[RFC9855] Bashandy, A., Litkowski, S., Filsfils, C., Francois, P., Decraene, B., and D. Voyer, "Topology Independent Fast Reroute Using Segment Routing", RFC 9855, DOI 10.17487/RFC9855, October 2025, <https://www.rfc-editor.org/rfc/rfc9855>.

Acknowledgements

Authors' Addresses

Francois Clad (editor)
Cisco Systems, Inc.
France
Email: fclad.ietf@gmail.com

Clarence Filsfils
Cisco Systems, Inc.
Belgium
Email: cf@cisco.com

Roy Jiang
Alibaba
China
Email: royjiang@aliyun-inc.com

Dennis Cai
Alibaba
China
Email: d.cai@alibaba-inc.com