

RTGWG Working Group
Internet Draft
Intended status: Informational
Expires: December 09, 2025

W. Cheng
China Mobile
C. Lin
New H3C Technologies
W. Wang
China Mobile
B. Xu
China Unicom
June 7, 2025

Reliability in AI Networks Gap Analysis, Problem
Statement, and Requirements
draft-cheng-rtgwg-ai-network-reliability-problem-03

Abstract

This document provides the gap analysis of existing reliability mechanism in AI networks, describes the fundamental problems, and defines the requirements for technical improvements.

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 December 09, 2025.

Copyright Notice

Copyright (c) 2025 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with

Table of Contents

1. Introduction.....	2
1.1. Requirements Language.....	3
1.2. Terminology.....	3
2. Existing Mechanisms.....	4
2.1. Routing Convergence in AI network.....	4
2.2. Spine-Leaf topology.....	5
2.3. Dragonfly topology.....	6
3. Gap Analysis.....	8
3.1. Fault detection Timing.....	8
3.2. Notifications Event Propagation Timing.....	9
3.3. Fault switchover Timing.....	9
4. Problem Statement.....	9
5. Requirements for AI network Mechanisms.....	10
6. Security Considerations.....	11
7. IANA Considerations.....	11
8. References.....	11
8.1. Normative References.....	11
8.2. Informative References.....	11
Authors' Addresses.....	12

1. Introduction

AI training places higher demands on network reliability for the following reasons:

Large-scale data transmission: AI training requires a significant amount of data for model training. These data often need to be obtained from distributed storage systems or cloud platforms and transmitted to the training servers. A highly reliable network ensures stable data transmission, preventing data loss or transmission errors.

Long training duration: AI model training typically takes hours or even days. During this process, the network connection should remain stable to ensure that the training process is not interrupted or terminated. Any network interruptions or failures can lead to training interruptions, requiring the process to be restarted and wasting time and resources.

High bandwidth requirements: AI training demands high network bandwidth. Operations such as large-scale data transmission, model parameter updates, and gradient calculations require fast and stable network connections to ensure efficient training. Network unreliability or low bandwidth can result in slower training speeds and impact training effectiveness and efficiency.

Distributed training: To accelerate training speed and improve model performance, AI training often employs distributed training methods that distribute computational tasks to multiple servers for parallel computing. This requires a highly reliable network to ensure data synchronization and communication in distributed training, ensuring model consistency and accuracy.

In summary, AI training places higher demands on network reliability, requiring stable data transmission, fast bandwidth, and stable connections to ensure smooth training processes and reliable results.

To ensure uninterrupted tasks during large-scale model training, it is crucial to address hardware failures. Take, for instance, a cluster that can accommodate 16,000 cards, with almost 100,000 optical modules. Considering the quality of actual hardware, let's assume that the Mean Time Between Failures (MTBF) of a single module is 10 million hours. MTBF denotes the average usage time of a hardware device prior to malfunction. However, with a large number of modules, even with a MTBF of 10 million hours, an average failure may display every four days approximately. In this situation, even low probability events become highly likely, considering the large number of modules involved. Therefore, AI networks concentrate on developing faster recovery capabilities from hardware failures.

This document provides the gap analysis of existing reliability mechanism in AI networks, describes the fundamental problems, and defines the requirements for technical improvements.

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.

1.2. Terminology

Routing: The path or strategy that data packets take to transmit through the network.

Topology: The physical and logical layout structure of the network.

Routing algorithm: The algorithm that determines the path or strategy for data packets to transmit through the network.

2. Existing Mechanisms

2.1. Routing Convergence in AI network

This section briefly introduces the existing routing convergence mechanisms in AI networks.

Traditional network failures rely on the control plane for detection and propagation of faults. The control plane then performs route convergence or uses Fast Reroute (FRR) mechanisms to quickly switch to backup paths. The convergence time for traditional network failures is typically around 50ms, and it is influenced by the working mechanism.

The following are several fast convergence methods,

The methods for link fault detection:

Bidirectional Forwarding Detection (BFD): BFD is used for fast fault detection. It provides a lightweight mechanism for quickly detecting faults and triggering a convergence process.

The methods for responding to local link faults and performing switchover.

Equal-Cost Multipath (ECMP): ECMP allows for fast fault switching by distributing traffic across multiple equal-cost paths. In the event of a failure on one path, traffic can be quickly redirected to an alternate path.

Fast Reroute (FRR): FRR is a mechanism that enables rapid switching to precomputed backup paths upon failure detection. It reduces the convergence time by bypassing the traditional control plane route convergence process.

The methods for responding to remote link faults and performing switchover.

BGP PIC (Prefix Independent Convergence): BGP PIC is a technique for fast iterative switching during network failures.

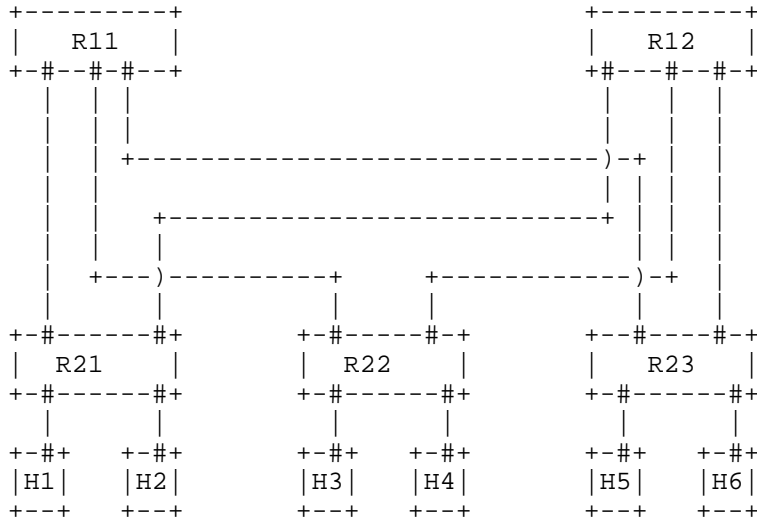


Figure 1: Spine-Leaf network diagram

In the commonly used Spine-Leaf topology for AI, there are two paths for communication between H1 and H5. The first path is R21->R11->R23, and the second path is R21->R12->R23. These two paths form ECMP (Equal Cost Multi-Path) paths, enabling load balancing of traffic.

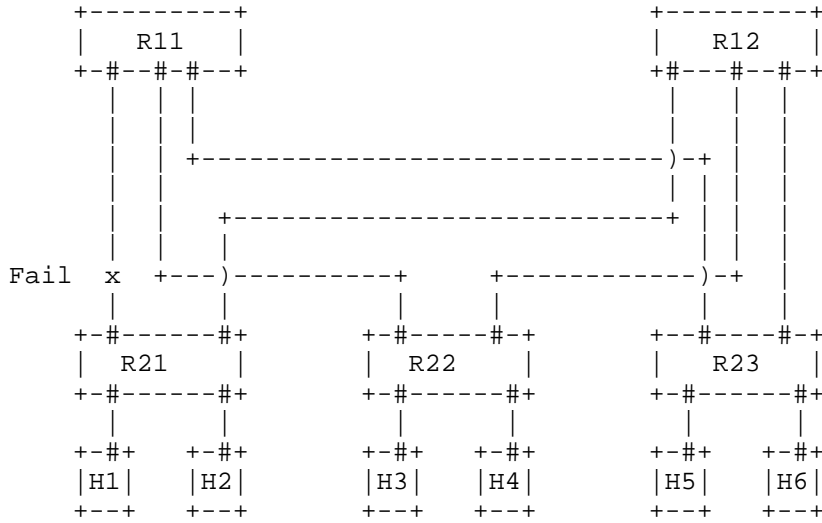


Figure 2: Local Link Failure

If a link failure occurs between R21 and R11, it is considered a local link failure for R21. Existing detection techniques such as

BFD can quickly identify this type of failure. When a local link failure (R21->R11->R23) is detected on one of the ECMP paths, the other equivalent path (R21->R12->R23) will be used for traffic forwarding. The duration of this process is mainly dependent on the time taken to detect the link failure.

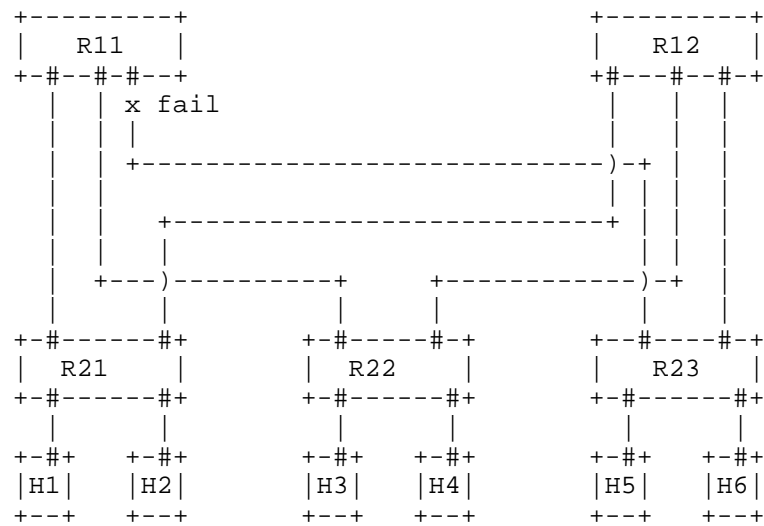


Figure 3: Remote Link Failure

If a link failure occurs between R11 and R23, this failure is considered a remote link failure for R21.

R11 propagates the link failure to R21 through IGP link state updates or BGP route withdrawal.

In the case of a remote link failure switchover, the process is mainly delayed by the propagation of fault information and the response switching of the remote link failure.

2.3. Dragonfly topology

Dragonfly is another widely used topology for AI training.

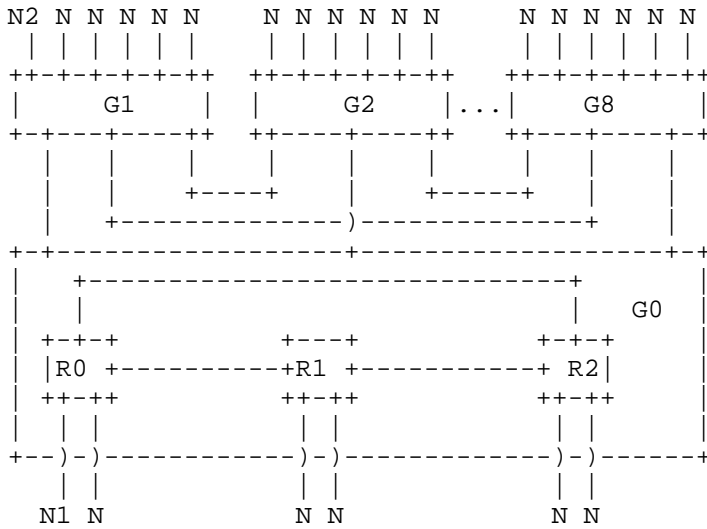


Figure 4: DragonFly network diagram

As shown in the diagram, N1 is connected to R0 in Group 0, and N2 is connected to the router in Group 1. The Inter-Group Link between Group 0 and Group 1 is assumed to be connected through R2. The traffic from N1 to N2 first goes through the Intra-Group Link from R0 to R2, then it is sent through the Inter-Group Link to Group1, and finally, it is forwarded to N2 via the Inter-Group Link in Group 1.

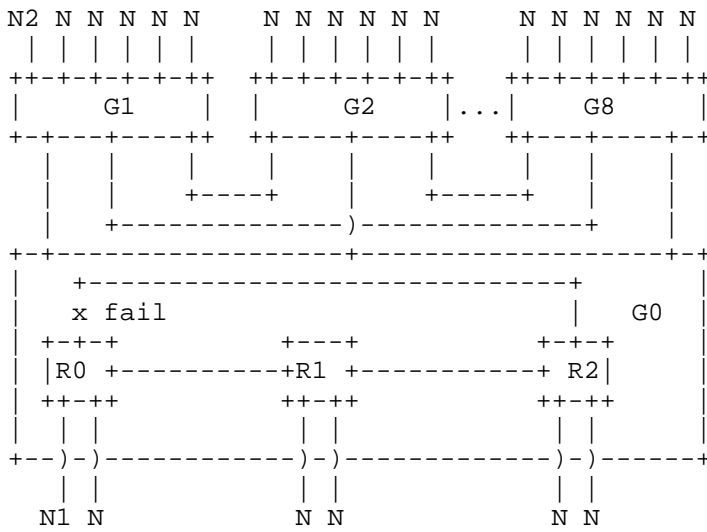


Figure 5: Intra-Group Link Failure

If a link failure occurs in Intra-Group link, The failure can be detected through BFD quickly by R0. Intra-Group link failure is a type of local link failure.

Once the failure is detected, R0 in the group switches the traffic to the backup path R0->R1->R2 for forwarding, Then the traffic is forwarded through the Inter-Group Link.

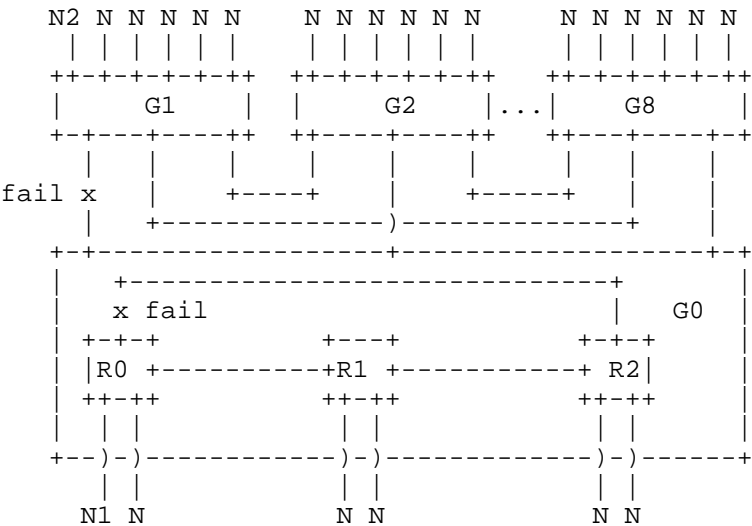


Figure 6: Inter Link Failure

If a link failure occurs in Inter-Group link, R0 cannot directly detect link failures and needs to be informed by a remote device detecting the link failure. R0 responds to the remote link failure by selecting a new path for forwarding. Inter-Group link failure is a type of remote link failure.

For Intra-Group Link failures, the main time taken for switching lies in the detection of the link failure.

For Inter-Group Link failures, it is necessary to detect the link failure, then transmit it to R0, and finally respond to the remote link failure by switching to a new path for forwarding.

3. Gap Analysis

3.1. Fault detection Timing

Ethernet links may support failure signaling or detection standards such as Connectivity Fault Management (CFM) as described in [IEEE8021Q]; this may make failure detection more robust.

Alternatively, some platforms may support Bidirectional Forwarding Detection (BFD) [RFC5880] to allow for sub-second failure detection and fault signaling to the BGP process. However, the use of either of these presents additional requirements to vendor software and possibly hardware. Since links in modern data centers are predominantly point-to-point fiber connections, a physical interface failure is often detected in milliseconds and subsequently triggers a BGP reconvergence.

3.2. Notifications Event Propagation Timing

After detecting a link failure, devices typically notify other devices through a link-state protocol or BGP route withdrawal, which typically takes milliseconds to complete.

3.3. Fault switchover Timing

Local link failure:

The existing mechanism allows for local detection of link failures, which can be directly handled by the hardware to switch between ECMP links. In the scenario depicted in Figure 1, when R11 detects a link failure to R23, the hardware switches directly to the second ecmp link. In this case, the switchover time is mainly determined by the link failure detection time.

Remote link failure:

Currently, there is no mechanism available to support this method of fast switchover for remote link failures. It can only rely on the routing protocol to perform a new routing calculation, including IGP SPF (Shortest Path First) or BGP route calculation, which typically takes seconds or even more.

4. Problem Statement

The number of parameters required for AI learning and training can vary greatly depending on the specific model and task at hand. For large AI models, the number of parameters for AI training can reach the millions.

And for large models, the training time for AI can take even several months or longer.

When a link failure occurs, the impact on AI training is as follows:

Performance impact: This includes issues such as training being stopped or RDMA not having a timeout processing mechanism.

Breakpoint reboot: The training process is paused and the system needs to be rebooted at a breakpoint. This can take anywhere from 30 minutes to several hours. The training task cannot proceed until the fault is resolved.

During AI training, the switch time for link failures should be as short as possible to minimize the impact on the training process. Typically, for most enterprises, the switch time for network link failures should be controlled within the millisecond or even microsecond range in order to minimize disruptions to the stability and performance of AI training. Otherwise, if there is a prolonged link failure, AI training would need to be restarted.

However, the current situation is that the failure rate of switches and optical modules is high, and the switch time is far from reaching the microsecond level, and even fails to achieve the millisecond level in most cases.

5. Requirements for AI network Mechanisms

In summary, For AI training networks, it is required to switch to an available link within microseconds after a link failure occurs. new requirements for the existing network for AI training include:

- 1) a new fault detection mechanism that can quickly detect the status of local and remote link failures; It is required to achieve link fault detection time in the microsecond range, while the current leading BFD (Bidirectional Forwarding Detection) for link detection requires at least several tens of milliseconds.
- 2) New techniques are needed to proactively eliminate link congestion that may be caused by link switchover. In the scenario of large workloads in AI training networks, once link congestion occurs, it will result in more severe network failures.
- 3) a new cross-device fault notification mechanism that enables other devices concerned with the fault to receive notifications quickly; It is required to achieve link fault detection time in the microsecond range, while the current leading BFD (Bidirectional Forwarding Detection) for link detection requires at least several tens of milliseconds.
- 4) a new fast table switching mechanism that can swiftly switch to backup links in response to remote link failures; For local link failure switchover, the current mechanisms like FRR can achieve millisecond-level performance, but further optimization is required for AI networks. On the other hand, for remote link failure switchover, there is currently no fast switchover mechanism available. It relies on re-routing calculation convergence through routing protocols. Even with optimizations

Internet-Draft AI network reliability problem June 2025
like BGP PIC, it only reduces the rate of table distribution from
the control plane to the forwarding plane.

- 5) expansion of the control plane to maintain this rapid remote link switching mechanism. If a suitable fast switchover solution at the forwarding plane is implemented for remote link failure, it would still require expanding the control plane protocols to maintain fast switchover entries and distribute them to the hardware.

6. Security Considerations

TBD.

7. IANA Considerations

This document does not request any IANA allocations.

8. References

8.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [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>>.

8.2. Informative References

TBD

Weiqiang Cheng
China Mobile
China

Email: chengweiqiang@chinamobile.com

Changwang Lin
New H3C Technologies
China

Email: linchangwang.04414@h3c.com

Wenxuan Wang
China Mobile
China

Email: wangwenxuan@chinamobile.com

Bohua Xu
China Unicom
China

Email: xubh15@chinaunicom.cn

