RTGWG Working Group
Internet Draft
Intended status: Informational
Expires: September 3,2025

R. Wang
China Mobile
C. Lin
New H3C Technologies
W. Wang
China Mobile
W. Cheng
China Mobile

March 2, 2025

Routing mechanism in Dragonfly Networks Gap Analysis, Problem Statement, and Requirements
draft-wang-rtgwg-dragonfly-routing-problem-03

Abstract

This document provides the gap analysis of existing routing mechanism in dragonfly 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 September 3 2025.

Copyright Notice

Table of Contents

1. Introduction.....3
   1.1. Requirements Language.....3
   1.2. Terminology.....3
2. Existing Mechanisms.....4
   2.1. Basic Topology.....4
   2.2. Routing mechanisms in Dragonfly network.....5
3. Gap Analysis.....6
   3.1. Load In balance.....6
   3.2. Adaptive Routing Notifications.....6
4. Problem Statement.....8
5. Requirements for Dragonfly network Mechanisms.....8
6. Security Considerations.....9
7. IANA Considerations.....9
8. References.....10
   8.1. Normative References.....10
   8.2. Informative References.....10
Authors' Addresses.....11

1. Introduction

Dragonfly network is a type of high-performance computer interconnection network architecture that is commonly used in large-scale computing environments. It consists of a collection of interconnected groups, with each group containing several computing resources such as processors, storage devices, and nodes. The nodes within each group communicate with each other using a high-speed local network, while the groups themselves are connected through a global network. Dragonfly networks are designed to provide high bandwidth and low latency communication capabilities, making them ideal for applications that require large-scale data processing and intensive computing tasks. Overall, dragonfly networks offer a scalable, efficient, and flexible solution for connecting hundreds or even thousands of computing resources in a parallel computing environment.

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

Group: In a group, multiple nodes are organized into a physical topology structure and interconnected by a high-speed network.

Inter-group link: Link connecting different groups.

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

Topology: The physical and logical layout structure of the network. Dragonfly network is a type of topology.

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

Congestion control: When there is too much traffic in the network, adjusting the transmission rate and routing method, etc., to avoid network congestion.

MR : Minimal Routing

NMR Non-Minimal Routing

AR: Adaptive Routing

VLB: Valiant Load-Balanced Routing

2. Existing Mechanisms

2.1. Basic Topology

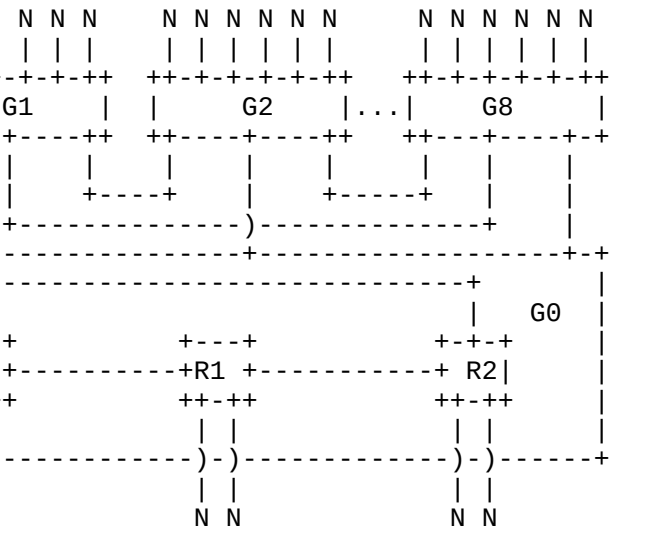

Figure 1: DragonFly network diagram

In the DragonFly network shown in Figure 1, there are a total of 9 groups, with each group consisting of 3 routers (G). Each router is connected to 2 nodes (N). The groups in the DragonFly network are connected through inter-group links. The routers within each group, as well as between routers and nodes, are connected through high-speed links within the group.

For data communication within a group, it is typically sufficient to forward traffic only through the links within the group. For data communication between groups, traffic needs to be forwarded through both the links within each group and the inter-group links. The specific path selection in the Dragonfly network is typically determined by the routing protocol used in the network. The routing protocol is responsible for dynamically determining the best path for data packets to travel from the source to the destination.

Various topologies can be used to form the intra-group connectivity.A typical intra-group topology is a fully connected

graphwhere all switches are directly connected to each other. An exampleof such an intra-group topology is shown in the G0 group in Figure 1. The intra-group connectivity in the Cascade architectureis a 2-dimensional all-to-all mesh.

2.2. Routing mechanisms in Dragonfly network

This section briefly introduces the existing routing mechanisms in dragonfly networks. Dragonfly networks use several routing mechanisms, each with its own advantages and disadvantages. Here are some brief overviews of several common routing mechanisms:

- Minimal Routing is the simplest and most commonly used routing mechanism in Dragonfly networks. It uses the path with the least number of channels to quickly deliver data to the destination node. The advantage of MR is that it is easy to implement and has low latency. However, since MR only focuses on the path with the least number of channels, the risk of load imbalance is relatively high.

- Non-Minimal Routing is a routing mechanism that avoids load imbalance by choosing a path other than the one with the least number of channels. The advantage of NMR is that the routing algorithm is intelligent and flexible, able to balance the load of network communication and reduce latency. However, NMR is more complex, requiring more computational resources and communication overhead.

- Adaptive Routing is a mechanism that can dynamically adjust the routing path by intelligently judging the network congestion status. AR’s strengths lie in its adaptability, which can control traffic in high-load situations and prevent congestion. The disadvantage is that its implementation is complex and requires more sophisticated algorithms and computational resources.

- Valiant Load-Balanced Routing uses the classic Valiant algorithm to select paths between global routing networks and then uses load-balancing routing algorithms between each group. The advantage of VLB is that it can achieve load balancing across the network range. The disadvantage is that it is complex, requiring more resources and computational costs.

Overall, the choice of routing mechanism in dragonfly networks requires a balance between performance, cost, and other factors and depends on specific application scenarios and requirements.

3. Gap Analysis

3.1. Load In balance

When the Dragonfly network routes through the minimum-route mechanism, the problem of load imbalance is easy to occur because the routing path is fixed and the communication volume between different groups in Dragonfly network may not be the same. When load imbalance occurs, the group with larger communication volume may be overly congested, affecting the overall performance of the network. We need a load balancing mechanism that can distribute the load between optimal and non-optimal links to avoid congestion on the main link.

There are several load balancing mechanisms that can achieve this goal. One common approach is to use a combination of Equal-Cost Multi-Path routing and Link Aggregation. ECMP distributes the traffic across multiple paths based on their cost, while Link Aggregation combines multiple physical links into a single logical link to increase bandwidth and provide redundancy.

However, non-minimum-route mechanism is required to calculate the distance of all possible paths, which requires more communication and computational resources, and cannot completely avoid the problem of load imbalance.

Adaptive routing mechanism can dynamically adjust the routing path according to the network congestion situation, making the network more adaptable to different traffic. However, the computational cost of this mechanism is high, and it occupies some of the bandwidth of the network, which may affect the performance of applications.

Valiant load-balancing routing algorithm can achieve load balancing across the entire network, but its design and implementation are complex and require more computing and communication resources. Although it can improve routing reliability and fault tolerance, it may not be necessary to adopt this mechanism in small-scale networks.

Due to the random network topology used in the Dragonfly network, the distance between each node internally is random, which may cause some unnecessary redundancies in routing and affect routing efficiency.

3.2. Adaptive Routing Notifications

The dynamic adjustment of flow paths based on the load situation is a traffic scheduling algorithm that can dynamically choose the

optimal flow path based on the load of nodes (or links) in the network to transmit data packets.

This algorithm usually uses two techniques: one is based on traffic measurement, and the other is based on protocol exchange between routers or switches. The traffic measurement-based technology measures the traffic in the network using network analysis tools or dedicated hardware embedded in routers or switches. Once some nodes or links with high loads are detected, the flow path can be automatically adjusted to alleviate the load. On the other hand, the protocol exchange-based technology relies on protocol communication between routers or switches to obtain the current network topology and node load data, and flow paths can be adjusted accordingly based on this information.

In this way, network administrators can ensure that there is always the best data flow path at any time, thereby maximizing network performance, reducing latency, and avoiding network congestion. At the same time, dynamic adjustment of flow paths can also provide robustness to the network, enabling it to automatically adapt to adverse events such as changes in network topology and node failures.

Whether based on traffic testing or protocol exchange between routers or switches, devices need to be able to communicate the current network performance in a quantitative manner.

Traffic testing technology requires the use of network analysis tools or dedicated hardware embedded in routers or switches to measure traffic in the network and obtain information about node load. These node load data needs to be translated into digital data and sent to the control plane through protocols or interfaces such as SNMP (Simple Network Management Protocol), Netflow, etc.

Protocol exchange technology uses protocol communication between routers or switches, such as OpenFlow, IS-IS, etc., to obtain the current network topology and node load information through the control plane. These information are often encoded into digital formats and transmitted to the operation plane through network transmission protocols.

Adaptive routing notifications are a communication protocol used to relay routing information and network load in a network. These notifications can be messages between nodes or between switches and routers.

In the Dragonfly network, adaptive routing notifications are utilized to implement adaptive routing mechanisms. When the network

load reaches a certain level, nodes and switches use notifications to dynamically choose routing paths. For example, during network congestion, switches and routers send notifications to prompt nodes to redirect traffic to different ports or nodes. These notifications can also include other information about network congestion and load balancing, such as bandwidth usage, device load and performance, and traffic rates.

The benefits of using adaptive routing notifications in the Dragonfly network are that they enable real-time adjustments of routing paths for nodes and switches, avoiding congestion and improving network performance. Additionally, adaptive routing notifications help network administrators identify and resolve network issues more easily, such as pinpointing congestion points and routing bottlenecks.

In summary, adaptive routing notifications play a significant role in the Dragonfly network and are a crucial component in implementing adaptive routing mechanisms.

Regardless of the approach, communication between devices needs to be standardized and routinized to achieve self-adaptation and interoperability across devices. Standardized and routinized communication between devices is critical to building adaptive networks.

4. Problem Statement

The current problem with the Dragonfly network is the lack of a concise and effective routing protocol for load balancing between optimal and non-optimal links.

Another problem is that for dynamic load balancing, it is necessary to standardize how network performance is quantified and communicated in a quantitative manner. This requires standardization.

5. Requirements for Dragonfly network Mechanisms

In the Dragonfly architecture, the routing protocol is a crucial component that guides packet transmission and route selection. Here are several aspects that the routing protocol in the Dragonfly architecture requires:

* Low latency: Low latency is essential in the Dragonfly architecture. Therefore, the routing protocol must be fast and

efficient to ensure that packets are transmitted to the destination node promptly.

* Load balancing: Load balancing is important in the Dragonfly architecture, and the routing protocol needs to support multiple available paths for load balancing. The routing protocol should dynamically select among multiple available paths to ensure fast packet transmission and distribute the load across network connections.

* Scalability: The Dragonfly architecture is typically deployed at large scale with a large number of nodes communicating with each other. Hence, the routing protocol needs to be scalable and capable of supporting route selection and packet transmission among a large number of nodes.

* Adaptability: The network topology in the Dragonfly architecture can change over time. The routing protocol needs to be adaptive and capable of re-computing optimal paths when the network topology changes, ensuring the selection of the best path for packet transmission.

* Reliability: The routing protocol in the Dragonfly architecture needs to ensure packet reliability. It should support link failure detection and recovery to ensure that packets can be correctly transmitted to the destination node in the event of link failures.

In summary, the routing protocol is a critical component in the Dragonfly architecture, requiring support for low latency, load balancing, scalability, adaptability, and reliability. Only with these requirements fulfilled can the routing protocol reliably operate in the Dragonfly architecture and provide efficient support for network communication.

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

Authors' Addresses

Ruixue Wang
China Mobile
China

Email: wangruixue@chinamobile.com

Changwang Lin
New H3C Technologies
China

Email: linchangwang.04414@h3c.com

Wenxuan Wang
China Mobile
China

Email: wangwenxuan@chinamobile.com

Weiqiang Cheng
China Mobile
China

Email: chengweiqiang@chinamobile.com