POSTER: Vedrfolnir: RDMA Network Performance Anomalies Diagnosis in Collective Communications

Yuxuan Chen, Menghao Zhang, Xiheng Li, Fangzheng Jiao, Chunming Hu
Beihang University

ABSTRACT

Collective communication becomes increasingly crucial as large language models rapidly evolve, but the RDMA it uses inevitably faces network performance anomalies (NPAs). Vedrfolnir is an accurate and efficient diagnosis system for RDMA NPAs in collective communication, which (1) constructs waiting graphs through algorithm decomposition, (2) adaptively detects anomalies while efficiently collecting diagnostic data, and (3) precisely analyzes performance bottlenecks and root causes. Evaluation shows that Vedrfolnir can achieve accurate diagnosis results with low overhead.

CCS CONCEPTS

• **Networks** → **Network monitoring**; *Programmable networks*;

KEYWORDS

Remote Direct Memory Access; Network Performance Anomalies Diagnosis; Collective Communication;

ACM Reference Format:
Yuxuan Chen, Menghao Zhang, Xiheng Li, Fangzheng Jiao, Chunming Hu. 2025. POSTER: Vedrfolnir: RDMA Network Performance Anomalies Diagnosis in Collective Communications. In *ACM SIGCOMM 2025 Posters and Demos (SIGCOMM Posters and Demos '25), September 8–11, 2025, Coimbra, Portugal*. ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3744969.3748396

1 INTRODUCTION

With the rapid scaling of large language model size and its training clusters, collective communication has become increasingly crucial to the efficiency of data exchange among GPUs. Collective communication often utilizes RDMA networks for inter-node connectivity to achieve high throughput and low latency. However, due to mechanisms such as line-rate start and PFC flow control [12], RDMA networks inevitably face congestion, leading to network performance anomalies (NPAs). As collective communication usually utilizes multiple flows simultaneously, this puts significant challenges to diagnose the root cause for collective communication performance anomalies. For example, in Fig. 3, flow contention occurs between collective communication flows (F1 and F2) and a background flow (BF2), resulting in performance anomalies of collective communication. It is difficult to attribute the performance degradation of the collective communication to BF2.

Existing works fall short in accurately and efficiently diagnosing RDMA NPAs in collective communication. Firstly, existing diagnosis methods primarily focus on single-flow level analysis (e.g., Hawkeye [5, 10], SpiderMon [11], Sonata [2], and *Flow [7]), ignoring emerging paradigms such as collective communication, which consequently leads to insufficient capability for co-flow patterns. Unlike the static process at the single-flow level, in collective communication, a flow originating from a node may change over time and may have dependencies on other flows as collective communication proceeds. For example, in the Halving and Doubling [8] algorithm, the destination of a flow can change multiple times, and this change depends on the data transmitted from another flow. Current methods cannot accurately capture this dynamic feature.

Secondly, due to the involvement of numerous nodes and flows in collective communication, achieving accurate diagnosis with low overhead poses another significant challenge. Some approaches (e.g., NetSight [3], PINT [1]) collect telemetry data across all switches. While ensuring accuracy, they introduce substantial communication or analytical overhead. Other methods (e.g., SpiderMon [11], PrintQueue [4]) analyze queue contention only in partial switches while neglecting the chain propagation characteristics of PFC in RDMA, resulting in insufficient diagnostic precision. Hawkeye [5, 10] addresses these gaps via a PFC provenance-based methodology, but still lacks design for collective communication scenarios. For example, when deployed for anomaly diagnosis in such environments, Hawkeye's manually configured rigid trigger mechanism can repeatedly activate anomaly detection in multiple nodes in short intervals. This leads to redundant data collection and introduces substantial overhead.

To address the problems above, we propose Vedrfolnir, an efficient anomaly diagnosis system for RDMA network performance in collective communication. First, to characterize the dynamic behavior of collective communication, we decompose collective communication algorithms into steps and construct a waiting graph to describe the dependencies among the individual flows. Second, to ensure the accuracy and efficiency of diagnostic data collection, Vedrfolnir collects collective communication execution data on the host and captures anomaly-related network telemetry data across the switches, and reduces overhead by utilizing a step-aware adaptive detection mechanism. Finally, Vedrfolnir performs a comprehensive root cause analysis through correlated multi-source data fusion. We implement a prototype in NS3 to preliminarily validate the effectiveness of Vedrfolnir. Evaluation shows that

This work is supported in part by the National Natural Science Foundation of China (No. 62402025), the Huawei-BUAA Joint Lab, and the Fundamental Research Funds for the Central Universities. Menghao Zhang is the corresponding author.

SIGCOMM Posters and Demos '25, September 8–11, 2025, Coimbra, Portugal

ACM ISBN 979-8-4007-2026-0/25/09.
https://doi.org/10.1145/3744969.3748396

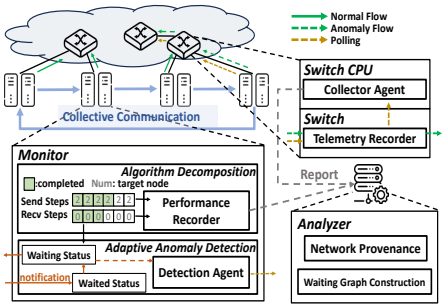

Figure 1: Vedrfolnir Framework.

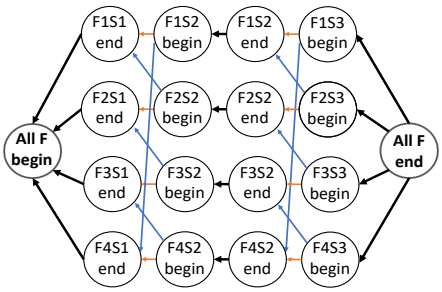

Figure 2: Waiting Graph.

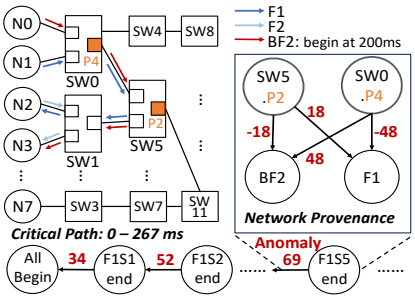

Figure 3: Topology and Diagnosis.

Vedrfolnir can accurately localize the root cause of NPAs and reduce telemetry collection overhead by 98% compared to Hawkeye.

2 VEDRFOLNIR DESIGN

Figure 1 illustrates the overall architecture of Vedrfolnir. During collective communication operations, monitors deployed on the hosts continuously record performance information of corresponding flows in real-time, which is reported to the analyzer. When performance anomalies occur, Vedrfolnir adaptively selects appropriate hosts to detect anomalies according to the steps decomposed by the algorithm. During the detection phase, a polling packet is sent by the host. This packet triggers the collection of telemetry information on switches related to the anomaly and the collected diagnostic information is subsequently reported to the analyzer. Ultimately, the analyzer gives the performance bottleneck of the entire collection communications and the root cause flows.

Algorithm Decomposition and Graphical Description. Firstly Vedrfolnir decomposes collective communication algorithms into steps. For a flow originating from a specific node, the transmitted data chunk or destination address changes between consecutive steps. For example, for Ring [6] algorithm, the flow transmits a different chunk of data for the source node with each step. For Halving and Doubling [8] algorithm, the destination node of the flow changes with each step. Based on these two changes, Vedrfolnir performs step divisions in parallel at each node and record in advance the sent and received targets for each step.

After the step splitting of the flows, Vedrfolnir constructs a waiting graph to describe the collective communication process. Taking a four-node ring-based reduce-scatter in Figure 2 as an example, we denote the flow where the sender is node i and the j-th step as F_iS_j, the begin or end of which is defined as a vertex. The directed edges are defined to represent the waiting relationship between vertexes, weighted by their corresponding waiting durations. Specifically, besides "F2S1 end", "F2S2 begin" also waits for "F1S1 end", because F1S1 needs to complete the transmission of the data that F2S2 depends on. Such waiting relationships are represented by light-colored edges with zero weight. Between the beginning and the end of F2S2 is represented by a dark-colored edge weighted by F2S2's execution duration.

Data Collection Workflow. (1) Host side. Vedrfolnir collects host-side collective communication execution information for waiting graph construction. When the system is running, the start time, execution time, and dependency relations of each flow step are recorded and reported to the analyzer. The analyzer builds the graph in reverse directed graph order. (2) Network side. When a flow performance anomaly is detected, the host sends polling packets to trigger telemetry information collection on switches. Vedrfolnir uses Hawkeye as the telemetry information collection mechanism to trace the path of the victimized flow. Vedrfolnir determines anomaly detection triggers through a step-aware adaptive mechanism. Specifically, when a step of a flow completes, the source host sends a notification packet to dependent flow source (if exists), marking it as "waited" states. If a "waited" host cannot deliver the marking to subsequent nodes while monitored flow RTT exceeds predefined thresholds, anomaly detection begins. Meanwhile, Vedrfolnir restricts the detection to only once per step of the flow. In this way, Vedrfolnir can theoretically reduce the redundancy overhead by half at least.

Comprehensive Root Cause Analysis. The analyzer of Vedrfolnir performs a comprehensive analysis on collected data from the host and the network. For host-side data, Vedrfolnir constructs a visualized waiting graph followed by pruning and analytical processing. As shown in Figure 2, during practical operation, nodes with two light-colored outgoing edges retain only one edge based on execution dependencies. For example, when F4S1 completes after F1S1, F1S2 actually waits only for F4S1, so the only outgoing edge of F1S2 is the blue one. Subsequently, Vedrfolnir recursively removes all nodes with zero in-degree. After the pruning is complete, Vedrfolnir computes the critical path of the graph, showing the flows that have the main impact on the execution time of the collective communication. For network-side data, Vedrfolnir adopts Hawkeye's analytical methodology to construct provenance graphs. By integrating both analytical perspectives, the system effectively pinpoints critical flows along with their root causes in network.

3 EVALUATION AND FUTURE WORK

We implemented an open-source prototype of Vedrfolnir [9] and used 8 nodes for ring `AllReduce` with two background flows in a k=4 fat-tree topology. The two background flows (BF1 and BF2) have flow contention with collective communication (F1) at two periods (40-200ms and 200-400ms) in sequence. Figure 3 illustrates the topology and diagnosis for BF2. Vedrfolnir accurately gives the performance bottleneck of collective communication by computing the critical path of the waiting graph and shows the results of network provenance. Moreover, Vedrfolnir collects 98% fewer telemetry bytes than Hawkeye. In future, we will give a clearer definition of the types of anomalies in collective communication, and evaluate Vedrfolnir in real testbeds.

REFERENCES

[1] Ran Ben Basat, Sivaramakrishnan Ramanathan, Yuliang Li, Gianni Antichi, Minian Yu, and Michael Mitzenmacher. 2020. PINT: Probabilistic In-band Network Telemetry. In *ACM SIGCOMM 2020*.

[2] Arpit Gupta, Rob Harrison, Marco Canini, Nick Feamster, Jennifer Rexford, and Walter Willinger. 2018. Sonata: query-driven streaming network telemetry. In *ACM SIGCOMM 2018*.

[3] Nikhil Handigol, Brandon Heller, Vimalkumar Jeyakumar, David Mazières, and Nick McKeown. 2014. I know what your packet did last hop: using packet histories to troubleshoot networks. In *USENIX NSDI 2014*.

[4] Yiran Lei, Liangcheng Yu, Vincent Liu, and Mingwei Xu. 2022. PrintQueue: performance diagnosis via queue measurement in the data plane. In *ACM SIGCOMM 2022*.

[5] Xiao Li, Shicheng Wang, Menghao Zhang, Zhiliang Wang, Mingwei Xu, and Jiahai Yang. 2024. POSTER: RDMA Network Performance Anomalies Diagnosis with Hawkeye. In *ACM SIGCOMM 2024*.

[6] Pitch Patarasuk and Xin Yuan. 2009. Bandwidth optimal all-reduce algorithms for clusters of workstations. *J. Parallel Distrib. Comput.* 69, 2 (Feb. 2009), 117–124.

[7] John Sonchack, Oliver Michel, Adam J. Aviv, Eric Keller, and Jonathan M. Smith. 2018. Scaling hardware accelerated network monitoring to concurrent and dynamic queries with *flow. In *USENIX ATC 2018*.

[8] Rajeev Thakur, Rolf Rabenseifner, and William Gropp. 2005. Optimization of Collective Communication Operations in MPICH. *Int. J. High Perform. Comput. Appl.* 19, 1 (Feb. 2005), 49–66.

[9] VEDRFOLNIR. 2025. Collective Communication Diagnosis. https://github.com/Networked-System-and-Security-Group/CollectiveCommunicationDiagnosis. (2025).

[10] Shicheng Wang, Menghao Zhang, Xiao Li, Qiyang Peng, Haoyuan Yu, Zhiliang Wang, Xiaohe Hu, Jiahai Yang, and Xingang Shi. 2025. Hawkeye: Diagnosing RDMA Network Performance Anomalies with PFC Provenance. In *ACM SIGCOMM 2025*.

[11] Weitao Wang, Xinyu Crystal Wu, Praveen Tammana, Ang Chen, and T. S. Eugene Ng. 2022. Closed-loop Network Performance Monitoring and Diagnosis with SpiderMon. In *USENIX NSDI 2022*.

[12] Yibo Zhu, Haggai Eran, Daniel Firestone, Chuanxiong Guo, Marina Lipshteyn, Yehonatan Liron, Jitendra Padhye, Shachar Raindel, Mohamad Haj Yahia, and Ming Zhang. 2015. Congestion Control for Large-Scale RDMA Deployments. In *ACM SIGCOMM 2015*.