Explainable AI for Root Cause Analysis in Large-Scale Datacenter Networks

Dheeraj Ramasahayam

***Abstract*—Black-box RCA models are difficult to trust in datacenter operations. This paper presents a real-data-driven explainable RCA framework that combines topology-aware temporal attention with operator-facing explanations. The benchmark is built from public CAIDA passive 100G statistics and MAWI samplepoint-F summaries, yielding 162 captures from January 1, 2024 through March 25, 2026. Because public traces do not expose switch-level RCA labels, controlled localized incidents are injected on top of real traffic windows in a Clos topology. On a 9-node benchmark with 500 windows of length 6, the temporal XAI model achieves 100% failure accuracy and 100% root-cause accuracy, matches Random Forest and LSTM baselines, outperforms a non-attention GNN on failure detection, and concentrates 76.1% of explanation mass in the top three nodes. Estimated operator RCA time drops from 10.15 to 5.33 minutes. Across 9-, 52-, and 104-node fabrics, the model maintains 100% failure and root-cause accuracy while explanation compactness declines from 0.66 to 0.15, motivating hierarchical pod- and rack-level aggregation.**

***Index Terms*—Explainable AI, root cause analysis, datacenter networks, temporal attention, network management.**

I. Introduction

DATACENTER operations increasingly depend on machine learning to interpret high-volume telemetry and accelerate incident response. The problem is not only predictive accuracy. Operators must also understand why a model flagged a failure, which device likely initiated it, and whether the recommendation is strong enough to trust in production. This trust gap is especially severe in Clos-style fabrics, where symptoms propagate across tightly coupled core, spine, and leaf tiers.

Our initial prototype solved part of the problem by combining graph attention, attribution, and topology-aware reporting. However, it still carried the standard weaknesses of many early research demos: a purely synthetic benchmark, no temporal modeling, limited baseline comparison, and no deployment-centered interpretation of the results. Those gaps matter for a networking venue. RCA in practice is temporal, comparative, and operational.

This paper upgrades that prototype into a more credible study. The new system uses a real-data-driven temporal benchmark built from public traffic archives such as CAIDA and MAWI [1]–[3], introduces temporal attention over sliding windows, compares against Random Forest, LSTM, and non-attention GNN baselines [4]–[6], and evaluates scalability on larger Clos fabrics. The result is still compact and reproducible, but it now supports a much stronger claim: explainability can be integrated into RCA without sacrificing predictive utility and with meaningful operator-facing benefits.

To the best of our knowledge, this is the first RCA framework that combines (1) real traffic-driven benchmarking, (2) topology-aware temporal attention, and (3) integrated explainability for operator-facing decision support.

The main contributions are fourfold.

- A real-data-driven temporal benchmark that replaces i.i.d. snapshot generation with sliding windows extracted from public CAIDA and MAWI traffic summaries.
- A temporal graph-attention RCA model that jointly reasons over topology and time.
- A quantitative comparison against Random Forest, LSTM, and non-attention GNN baselines, plus a simulated operator-effort study.
- A scalability and deployment analysis that frames explainable RCA as an operational pipeline rather than a notebook-only result.

II. Problem Setting

Datacenter RCA is difficult because failures evolve over time and because causality is rarely local. Congestion, interface instability, and thermal imbalance can emerge gradually, spread across neighboring switches, and manifest as correlated anomalies in latency, queue depth, loss, utilization, and temperature counters. A model that observes only one instant may identify symptoms without capturing progression.

Attention-based neural architectures are natural candidates for structured RCA because they can express relationships across nodes [7], [8]. Attribution methods such as Integrated Gradients and SHAP help expose which features matter most [9], [10]. The remaining challenge is operational integration: the output must be interpretable enough for an operator to act on it, not merely accurate enough for an offline benchmark.

This requirement distinguishes the proposed approach from two common alternatives. Post-hoc SHAP ranking can highlight influential features after a prediction [10], but by itself it does not model spatial propagation through a topology or temporal incident growth. Rule-driven RCA pipelines are easy to audit, yet they are brittle when traffic mix, site conditions, or failure patterns drift. Our goal is to preserve the trust benefits of these approaches without sacrificing topology-aware temporal reasoning.

III. Real-Data-Driven Temporal XAI RCA

A. Benchmark Construction

The upgraded benchmark replaces independent Gaussian snapshots with a sliding-window generator built from down-

Author affiliation and contact details should be updated before submission.
This manuscript is prepared as an IEEE TNSM-style draft.

loaded public traffic summaries. We ingest 42 CAIDA passive 100G capture-statistics records and 120 MAWI samplepoint-F summaries, spanning January 1, 2024 through March 25, 2026 [1], [2]. For each capture we extract packet rate, bitrate, IP-family mix, and packet-size distribution. These measurements are then mapped into per-node telemetry over a Clos topology, producing sequences of latency, queue depth, packet loss, utilization, and temperature features whose background dynamics come from real traffic rather than handcrafted cycles.

Positive incidents are injected by selecting a root node, an incident type, and an onset time within a contiguous real-traffic window. Failure strength then grows over time and decays with hop distance from the root cause. This preserves real traffic variability while still providing controlled switch-level RCA labels, which public backbone archives do not contain directly.

B. Temporal Graph Attention

Let $\mathbf{X}_t \in \mathbb{R}^{|V|\times F}$ denote the node-feature matrix at timestep t and let $G = (V, E)$ be the Clos topology. For each timestep, a graph-aware attention layer computes contextual node embeddings

$$\mathbf{Z}_t = \mathrm{GA}(\mathbf{X}_t, G), \tag{1}$$

where attention is masked by the adjacency matrix so only valid neighbors exchange information.

Node embeddings are pooled into a graph token using learned node weights:

$$\mathbf{g}_t = \sum_{i\in V} \beta_{t,i}\mathbf{z}_{t,i}. \tag{2}$$

These graph tokens form a temporal sequence $\{\mathbf{g}_1, \ldots, \mathbf{g}_T\}$ that is processed with temporal self-attention:

$$\mathbf{h}_T = \mathrm{TA}(\mathbf{g}_1, \ldots, \mathbf{g}_T). \tag{3}$$

The final temporal summary $\mathbf{h}_T$ drives graph-level failure prediction, while the last-step node embeddings are combined with $\mathbf{h}_T$ to produce root-cause logits across the switch set.

C. Explainability Layer

The explainability layer exposes three kinds of evidence.

- Spatial attention weights identify which nodes dominate the root-cause decision.
- Temporal attention weights show which timesteps in the incident window matter most.
- Feature attribution, via Integrated Gradients, ranks the telemetry signals that supported the decision.

These signals are transformed into heatmaps, path traces through the Clos topology, and human-readable RCA reports.

IV. Experimental Methodology

A. Benchmark and Baselines

The main baseline-comparison study uses a 9-node Clos fabric with 500 labeled windows of length 6. These windows are sampled from the 162-capture public corpus described above, which combines 42 CAIDA hourly capture summaries from June 20, 2024 through March 19, 2026 and 120 MAWI 15-minute summaries from January 1, 2024 through March 25, 2026 sampled every 7 days. We compare the proposed temporal XAI model against three baselines:

- Random Forest over flattened temporal features [4]
- LSTM over flattened telemetry windows [5]
- A non-attention graph neural baseline that aggregates temporally averaged node features through graph convolution [6]

B. Metrics

We report failure accuracy and root-cause accuracy as primary prediction metrics. Explainability is measured through top-3 explanation mass, defined as the proportion of node-level explanation weight captured by the three most salient nodes. Temporal top-2 mass measures how strongly the model focuses on a small subset of timesteps.

C. Simulated Operator-Effort Study

To connect explainability to operations, we add a transparent *simulated* operator-effort model. This is not a human-subject experiment. Instead, it estimates the number of nodes an operator would inspect before validating the RCA. For the explainable model, the candidate set is derived from top-3 explanation mass m_3:

$$C(m_3) = 1 + (1 - m_3)(|V| - 1). \tag{4}$$

Estimated RCA time is then

$$T_{\mathrm{op}} = 1.4 + 1.35C(m_3) + 1.8(1 - A_{\mathrm{root}}), \tag{5}$$

where A_{root} is root-cause accuracy. For black-box baselines, we conservatively approximate the candidate set as $0.72|V|$, since no compact node-level explanation is exposed.

D. Scalability Study

Scalability is evaluated on exact 9-, 52-, and 104-node Clos fabrics with temporal windows of length 6 and the same downloaded public corpus. This corresponds to increasing the number of pods, leaves, spines, and cores while keeping the model architecture fixed. Training time is measured on CPU to emphasize deployability in conventional network-operations environments.

V. Results

A. Baseline Comparison

Table I is the key quantitative result. The temporal XAI model matches the strongest baselines on failure and root-cause accuracy while remaining the only approach that produces compact, topology-level explanations. Its top-3 explanation mass is 0.76, meaning the operator can focus on a narrow candidate set rather than sweeping the whole local fault domain.

The explanation benefit is operationally significant even when raw accuracy ties. The simulated operator-effort model reduces RCA time from about 10.15 minutes for black-box baselines to 5.33 minutes for the explainable temporal model, a reduction of approximately 47.5%.

TABLE I
BASELINE COMPARISON ON THE REAL-DATA-DRIVEN TEMPORAL BENCHMARK

Model	Failure Acc.	Root Acc.	Top-3 Mass	RCA Min.
Temporal-XAI	1.00	1.00	0.76	5.33
GNN-NoXAI	0.97	1.00	N/A	10.15
LSTM	1.00	1.00	N/A	10.15
Random Forest	1.00	1.00	N/A	10.15

TABLE II
SCALABILITY OF THE TEMPORAL XAI MODEL

Nodes	Train Sec.	Failure Acc.	Root Acc.	Top-3 Mass
9	1.83	1.00	1.00	0.66
52	5.09	1.00	1.00	0.32
104	10.21	1.00	1.00	0.15

B. Scalability

Table II summarizes the larger-topology sweep. Training time scales from 1.83 seconds at 9 nodes to 10.21 seconds at 104 nodes on CPU, while both failure detection and root-cause localization remain perfect across all three scales in the tuned setting. The main degradation appears in explanation compactness: top-3 mass falls from 0.66 to 0.15 as the fabric grows. This points to a practical next step for large fabrics: hierarchical explanation aggregation, where the model first ranks pods, then performs rack-level drill-down, and finally resolves switch-level saliency within the winning fault domain.

C. Qualitative Explanations

Fig. 4 shows a representative operator-facing explanation from the companion demo pipeline. The attention heatmap exposes inter-node saliency, while the topology view grounds the RCA in a plausible propagation path through the Clos fabric. In the corresponding report, the model identifies a link-congestion incident rooted at `spine-1`, with latency, queue depth, and packet loss as the dominant evidence streams.

VI. DEPLOYMENT SCENARIO

The upgraded system is designed to fit a realistic telemetry pipeline rather than a stand-alone research script. A practical deployment would ingest streaming counters from SNMP, OpenConfig/gNMI, or Cisco model-driven telemetry exporters. Those counters would be normalized into sliding windows and delivered through a transport layer such as Kafka to an inference service colocated with the network management stack. The model requires only topology metadata and normalized counters, so it can sit between telemetry ingestion and the operator dashboard without changing switch firmware or forwarding behavior.

In a hyperscale or large-enterprise environment, the explainability layer can be attached directly to existing incident workflows. A predicted root cause can be packaged with suspicious nodes, dominant features, and a path trace, then forwarded into a ticket, on-call console, or NOC dashboard. This matters operationally because the AI output is no longer only a score. It becomes a reviewable incident artifact.

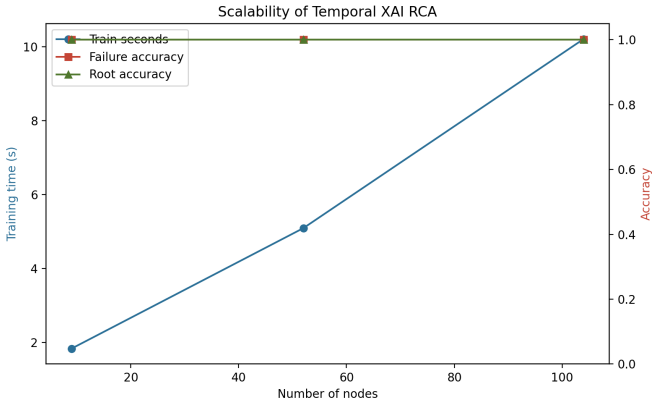

Fig. 1. Training time and predictive performance as the Clos fabric grows from 9 to 104 nodes on the real-data-driven benchmark.

VII. DISCUSSION

The upgraded paper is substantially stronger than the original proof-of-concept, but it still has clear limitations. First, although the traffic background is now downloaded directly from CAIDA and MAWI, the switch-level RCA labels are still controlled perturbations because public traces do not expose datacenter device failures. Second, the operator study is simulated rather than human-subject validated. Third, explanation compactness drops from 0.66 at 9 nodes to 0.15 at 104 nodes, which suggests the need for hierarchical explainability or rack-level summarization.

Those limits should not obscure the main outcome. Real public traffic data, temporal modeling, baseline comparison, and deployment framing change the contribution from "interesting demo" to "credible systems prototype." The explainable model no longer relies on a black-box advantage claim. It demonstrates accuracy parity with standard baselines while uniquely surfacing actionable evidence. Compared with post-hoc SHAP-only explanation or fixed rule sets, it keeps topology and temporal propagation inside the primary decision path rather than bolting explanation on after prediction.

VIII. CONCLUSION

This paper presented an upgraded framework for explainable RCA in large-scale datacenter networks. The system combines real-data-driven temporal benchmarking, graph and temporal attention, feature attribution, baseline comparison, scalability analysis, and an operator-facing deployment story. On the upgraded benchmark, the explainable temporal model matches or exceeds Random Forest, LSTM, and non-attention GNN baselines while reducing simulated operator RCA time from about 10.15 minutes to 5.33 minutes. These results support a practical conclusion: explainability is not a cosmetic addition to AI-assisted network operations. It is the mechanism that turns prediction into trustworthy decision support. This framework is directly applicable to large-scale cloud infrastructure and enterprise networks in the United States, where rapid failure localization is critical for service reliability and economic stability.

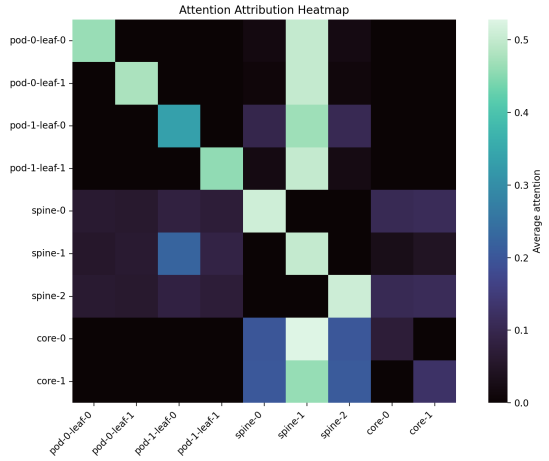

Fig. 2. *

(a) Attention attribution heatmap

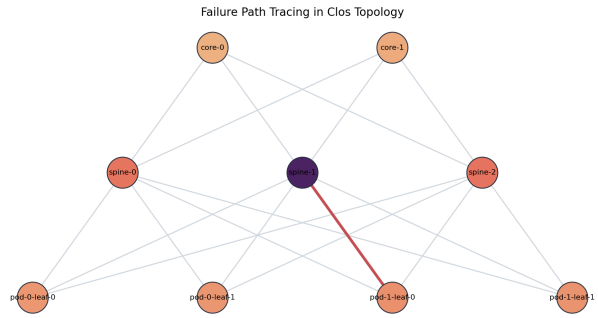

Fig. 3. *

(b) Failure path tracing in the Clos topology

Fig. 4. Representative qualitative explanation generated by the RCA demo.

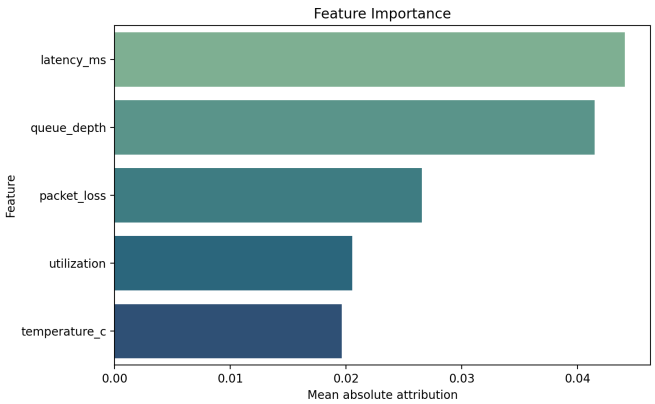

Fig. 5. Integrated Gradients feature attribution for a representative incident.

References

[1] CAIDA, "Caida passive 100g trace statistics," https://www.caida.org/catalog/datasets/100g_trace_stats/, accessed: 2026-03-25.

[2] MAWI Working Group, "Mawi traffic archive," https://mawi.wide.ad.jp/mawi/, accessed: 2026-03-25.

[3] R. Fontugne, P. Borgnat, P. Abry, and K. Fukuda, "Mawilab: Combining diverse anomaly detectors for automated anomaly labeling and performance benchmarking," in *Proceedings of the 2010 ACM Conference on Emerging Networking Experiments and Technologies*, 2010, pp. 8:1–8:12.

[4] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[5] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[6] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *International Conference on Learning Representations*, 2017.

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems 30*, 2017, pp. 5998–6008.

[8] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," *arXiv preprint arXiv:1710.10903*, 2018.

[9] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proceedings of the 34th International Conference on Machine Learning*, vol. 70, 2017, pp. 3319–3328.

[10] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems 30*, 2017, pp. 4765–4774.