

A Routing-Level Comparison of ECMP, e-ECMP and Flowlet-Based e-ECMP under Data Center Traffic Patterns

Kobi Kenzi

January 2026

Abstract

Modern data center networks rely on multipath routing to support high-bandwidth distributed workloads. This project presents a routing-level simulation framework for comparing Equal-Cost Multi-Path (ECMP), enhanced ECMP using multiple Queue Pairs (e-ECMP), Flowlet-only routing, and Flowlet-based e-ECMP in fat-tree topologies. By isolating routing decisions from transport and queuing effects, the simulator quantifies how spatial and temporal path diversity improve load balance and reduce congestion hotspots.

1 Background and Motivation

Fat-tree topologies are widely used in data center networks due to their high path diversity and scalability, as originally proposed in [1].

Standard Equal-Cost Multi-Path (ECMP) routing distributes flows by hashing each flow to a single equal-cost path [3], which can lead to persistent load imbalance due to hash collisions.

Recent large-scale AI training clusters based on RDMA over Ethernet (RoCE) have introduced practical extensions to ECMP in production environments. In particular, Meta's large-scale AI training fabric demonstrates the use of multiple Queue Pairs (e-ECMP) to increase spatial path diversity and mitigate congestion hotspots [2].

In parallel, flowlet-based routing was proposed to introduce temporal path diversity by periodically reshuffling routing decisions, preventing persistent congestion in multipath networks [4].

Understanding the isolated impact of these spatial and temporal routing extensions motivates this project.

2 Research Question

How do ECMP, e-ECMP, Flowlet-only routing, and Flowlet-based e-ECMP differ in their ability to distribute traffic load and mitigate congestion hotspots in fat-tree data center networks?

3 Methodology

A discrete-time Python simulator is implemented over a k -ary fat-tree topology. Uniform random host-to-host flows are generated, each carrying equal normalized load. For every source-destination pair, all equal-cost paths are precomputed.

The simulator operates at the flow and flowlet level without modeling queues or packet delays, thereby isolating routing behavior.

4 Routing Schemes

The simulator evaluates four routing schemes:

- **ECMP:** Each flow is deterministically hashed once to a single equal-cost path. This represents standard static multipath routing.
- **Flowlet-only:** Each flow is periodically re-hashed across equal-cost paths at discrete flowlet intervals, introducing temporal path diversity without spatial splitting.
- **e-ECMP:** Each flow is split into multiple Queue Pairs (QPs), where each QP is hashed independently. This introduces spatial path diversity.
- **Flowlet-based e-ECMP:** Each QP is periodically re-hashed across equal-cost paths at discrete flowlet intervals, combining spatial and temporal path diversity while preserving in-order delivery within each flowlet.

To ensure fair comparison, total offered load is normalized so that all schemes inject identical traffic volume.

5 Evaluation Metrics

Routing performance is evaluated using per-link simulated loads.

- **Tail Link Load (p90):** The 90th percentile of link loads across the network. This metric captures the severity of congestion hotspots at high-load links while being more stable than extreme tail metrics. A lower p90 indicates that fewer links experience disproportionately high load.

- **Coefficient of Variation (CV = std/mean):** This metric measures global fairness of load distribution. A lower CV indicates that traffic is spread more uniformly across links, while a higher CV reflects uneven utilization.
- **Link Utilization Fraction:** The fraction of links that carry nonzero load. Higher utilization indicates better exploitation of available path diversity and network capacity.
- **Empirical CDF of Link Loads:** The cumulative distribution of per-link loads provides a complete view of load distribution shape, highlighting heavy-tail behavior and uniformity.

6 Experimental Setup

Three parameter sweeps are conducted.

- **QPs Sweep:** $k = 8$, 2000 flows, QPs $\in \{1, 2, 4, 8, 16, 30\}$ flowlet epochs.
- **Flowlet Sweep:** $k = 8$, 2000 flows, QPs=8, flowlet epochs $\in \{1, 2, 5, 10, 20, 40\}$.
- **Topology Sweep:** $k \in \{4, 8, 16, 32, 32\}$, 2000 flows, QPs=8, 20 flowlet epochs.

All experiments use identical traffic realizations for fair comparison.

7 Results

Results are organized by sweep parameter in the following order: QPs, Flowlet epochs, Topology size, and finally the link-load CDF.

7.1 Impact of Spatial Multipath: QPs Sweep

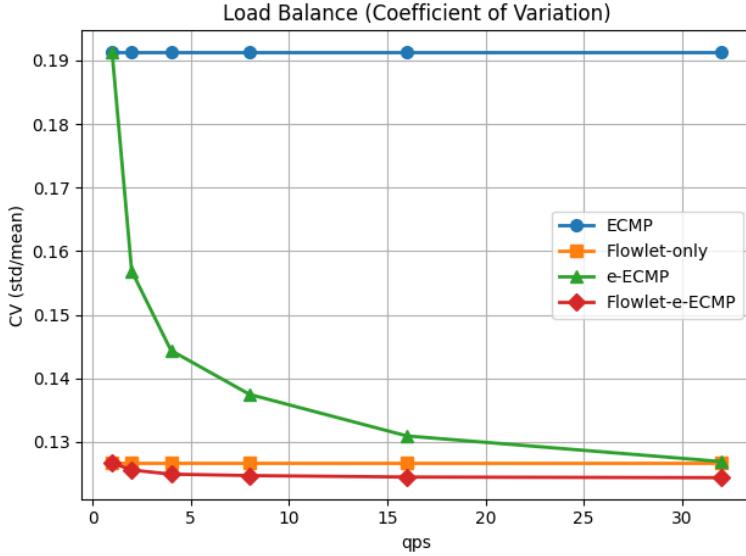


Figure 1: Load balance (CV) vs. number of QPs per flow.

Figure 1 shows that ECMP maintains constant CV across QP values, confirming that ECMP does not exploit QP-level splitting. e-ECMP exhibits a rapid reduction in CV when increasing QPs, demonstrating improved global load fairness through spatial multipath. Flowlet-based e-ECMP consistently achieves the lowest CV, indicating that temporal reshuffling further smooths residual imbalance.

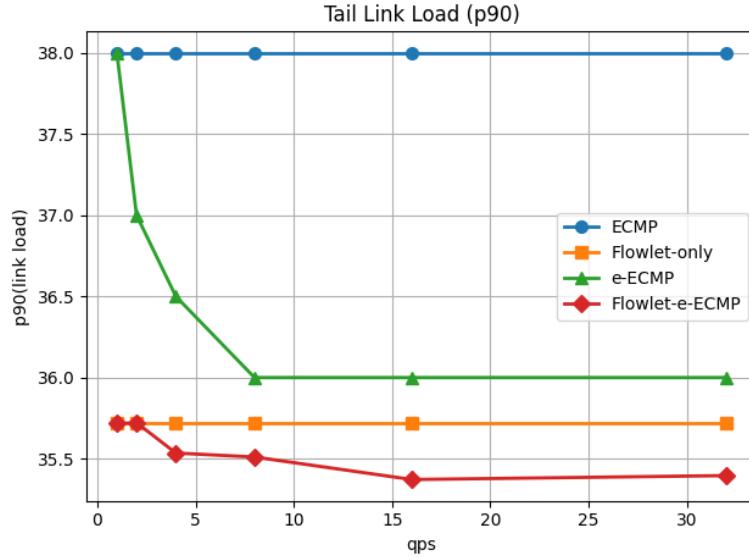


Figure 2: Tail link load (p90) vs. number of QPs per flow.

Figure 2 shows that ECMP's p90 remains constant across QP values, confirming persistent hash-induced congestion. e-ECMP significantly reduces p90 as QPs increase, showing effective mitigation of congestion hotspots. Flowlet-based e-ECMP achieves the lowest p90, confirming that temporal reshuffling further reduces persistent high-load links.

7.2 Impact of Temporal Multipath: Flowlet Epoch Sweep

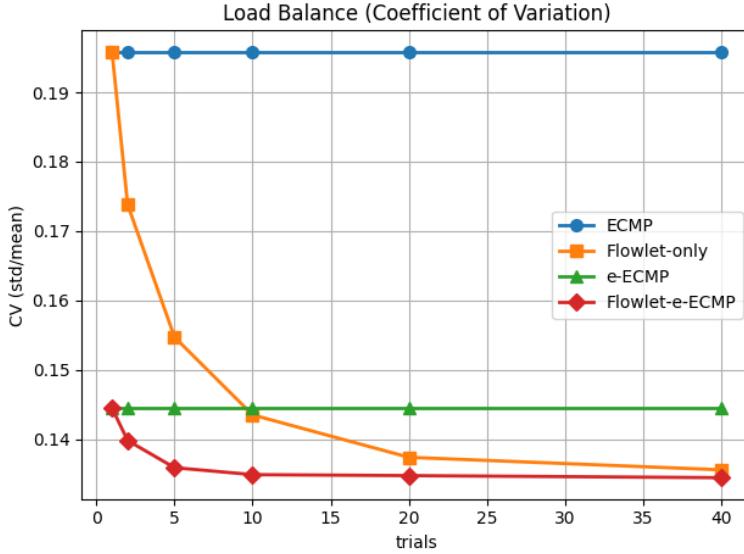


Figure 3: Load balance (CV) vs. number of flowlet epochs.

Figure 3 shows that Flowlet-only routing improves global fairness as reshuffling frequency increases. Flowlet-based e-ECMP achieves consistently lower CV, confirming that temporal and spatial multipath reinforce each other. Improvements saturate after approximately 10–20 epochs, indicating diminishing returns from excessive reshuffling.

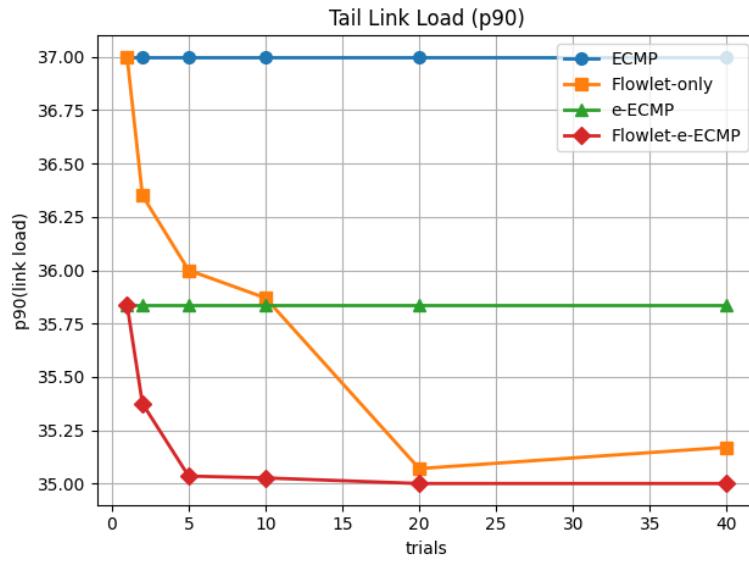


Figure 4: Tail link load (p90) vs. number of flowlet epochs.

Figure 4 confirms that temporal reshuffling reduces congestion hotspots. Flowlet-based e-ECMP consistently achieves the lowest p90, indicating superior suppression of persistent high-load links.

7.3 Impact of Topology Size: Fat-Tree Scaling

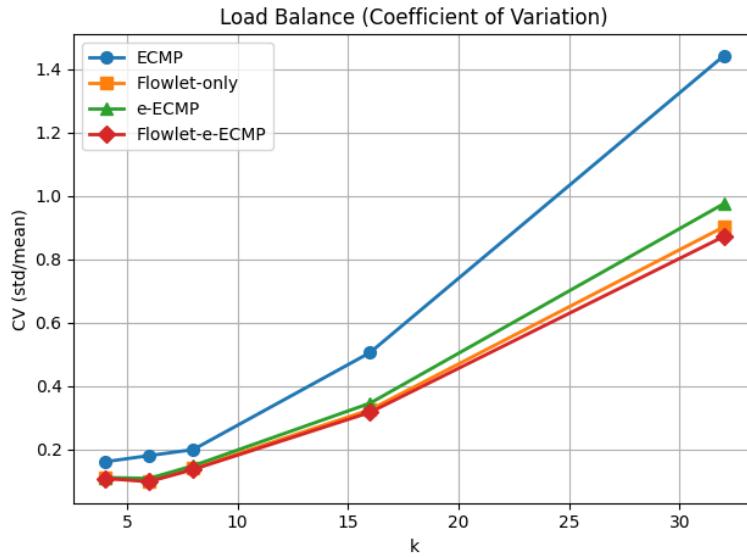


Figure 5: Load balance (CV) vs. fat-tree size k .

Figure 5 shows that ECMP suffers from increasing imbalance as topology scales. e-ECMP and Flowlet-based e-ECMP consistently improve fairness by exploiting available path diversity.

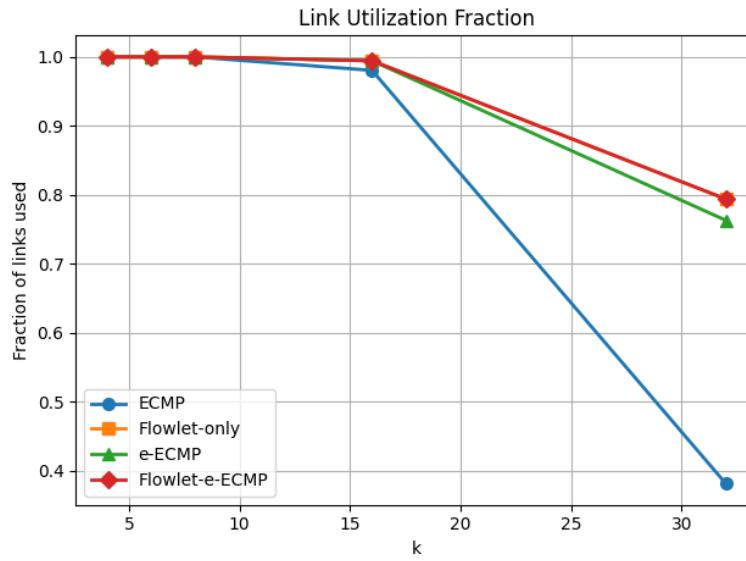


Figure 6: Fraction of utilized links vs. fat-tree size k .

Figure 6 shows that ECMP increasingly underutilizes available links as topology grows. Multipath extensions maintain high utilization, demonstrating superior exploitation of network capacity.

7.4 Load Distribution Shape: Link Load CDF

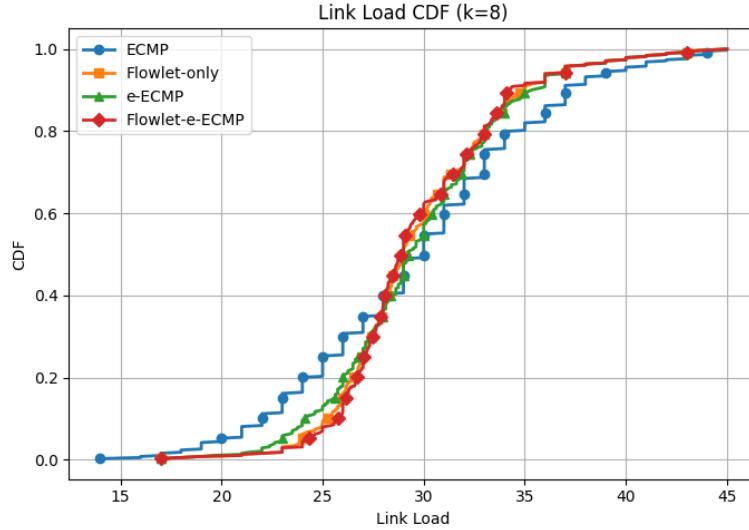


Figure 7: Empirical CDF of link loads for $k = 8$.

Figure 7 shows that ECMP produces a heavy-tailed load distribution due to static hash collisions. Flowlet-only routing shifts the distribution left via temporal reshuffling. e-ECMP further compresses the distribution through spatial multipath. Flowlet-based e-ECMP achieves the steepest CDF, indicating the most uniform load distribution.

8 Discussion

The experiments confirm that static ECMP suffers from persistent congestion hotspots caused by hash collisions. Spatial multipath through QP splitting significantly improves load distribution, while temporal multipath through flowlets further mitigates residual imbalance. Combining both mechanisms yields the most uniform utilization and lowest congestion across all evaluated scenarios.

9 Limitations

The simulator does not model queuing, transport protocols, or packet-level reordering. Latency and throughput are not evaluated. Future work may integrate queueing dynamics and collective communication patterns.

10 Conclusion

This work presents a routing-level simulation framework for studying multipath routing in fat-tree data center networks. Results show that e-ECMP significantly improves fairness over ECMP, and Flowlet-based e-ECMP further reduces congestion through temporal path diversity. These findings reflect the design principles of modern RDMA-based data center fabrics.

11 Reproducibility

Code and experiment scripts are available at:

<https://github.com/KenziVisor/fat-tree-topology-sim-with-ecmp>

References

- [1] Mohammad Al-Fares, Alexander Loukissas, and Amin Vahdat. A scalable, commodity data center network architecture. In *Proceedings of the ACM SIGCOMM Conference*, 2008.
- [2] Daniel Firestone et al. Rdma over ethernet for distributed ai training at meta scale. In *Proceedings of the ACM SIGCOMM Conference*, 2024.
- [3] Christian Hopps. Analysis of an equal-cost multi-path algorithm. RFC 2992, 2000.
- [4] Srikanth Kandula et al. Flowlet switching: Achieving load balance in multipath networks. In *Proceedings of ACM CoNEXT*, 2007.