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), and flowlet-based e-ECMP in fat-tree topologies. By isolating routing decisions from transport and queuing effects, the simulator quantifies how increased spatial and temporal path diversity improves load balance and reduces 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 [2]. Standard Equal-Cost Multi-Path (ECMP) routing distributes flows by hashing each flow to a single equal-cost path [4], which can lead to persistent load imbalance due to hash collisions. Recent large-scale AI training clusters extend ECMP by splitting flows across multiple Queue Pairs (e-ECMP) and further reshuffling routing decisions over time using flowlets, as demonstrated in modern RDMA over Ethernet deployments [3]. Understanding the isolated impact of these routing extensions motivates this study.

2 Research Question

How do ECMP, e-ECMP, 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 three routing schemes:

- **ECMP:** Each flow is deterministically hashed once to a single equal-cost path. This represents standard static multipath routing in data center networks.
- **e-ECMP:** Each flow is split into multiple Queue Pairs (QPs), where each QP is hashed independently. This increases spatial path diversity by distributing a single flow across multiple equal-cost paths.
- **Flowlet-based e-ECMP:** Each QP is periodically re-hashed across equal-cost paths at discrete flowlet intervals. This introduces temporal path diversity, allowing traffic to escape persistent hash-induced congestion 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. Since the simulator isolates routing behavior from queuing and transport effects, the metrics focus on how evenly traffic is distributed across network links.

- **Tail Link Load (p99):** This metric represents the 99th percentile of link loads across the network. Intuitively, it captures the severity of congestion hotspots: a high p99 indicates that a small fraction of links carry disproportionately large traffic, which in real systems would lead to queue buildup and increased latency. Reducing p99 is therefore a primary goal of effective load balancing.
- **Coefficient of Variation (CV = std/mean):** This metric measures the global fairness of load distribution. A lower CV indicates that traffic is spread more uniformly across links, while a higher CV reflects uneven utilization and potential inefficiencies. Unlike p99, which focuses on extreme hotspots, CV captures overall balance across the entire topology.

Together, p99 and CV provide complementary views: p99 highlights worst-case congestion behavior, while CV quantifies global load balance.

6 Experimental Setup

Three parameter sweeps are conducted using the experiment scripts provided in the repository.

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

Additional intermediate and large topology sizes ($k = 6, 32$) were tested to verify consistency of observed trends.

7 Results

7.1 Impact of Multi-QP Routing

The QPs sweep shows that standard ECMP exhibits high tail link loads due to static hash collisions. Increasing the number of QPs in e-ECMP significantly reduces p99 link load and improves global balance. Flowlet-based e-ECMP further decreases tail load by periodically reshuffling paths, preventing persistent congestion patterns.

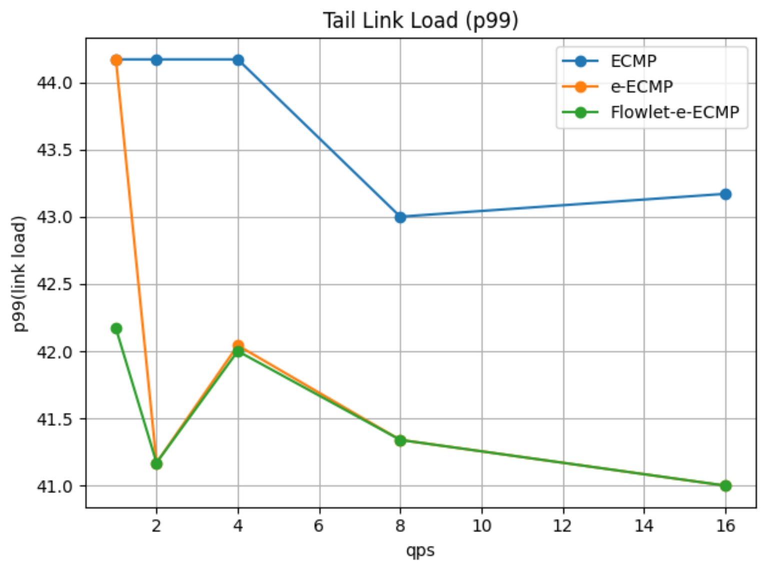

Figure 1: Tail link load (p99) vs. number of QPs per flow.

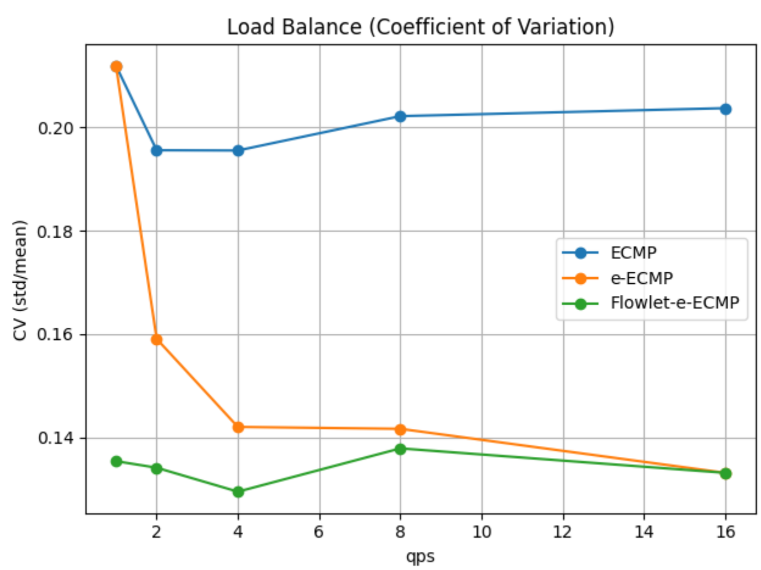

Figure 2: Load balance (CV) vs. number of QPs per flow.

7.2 Impact of Flowlet Reshuffling

The flowlet sweep demonstrates that temporal path reshuffling provides additional load smoothing beyond spatial multipath. Performance improves rapidly when increasing the number of flowlets from 1 to 10 epochs, after which diminishing returns are observed.

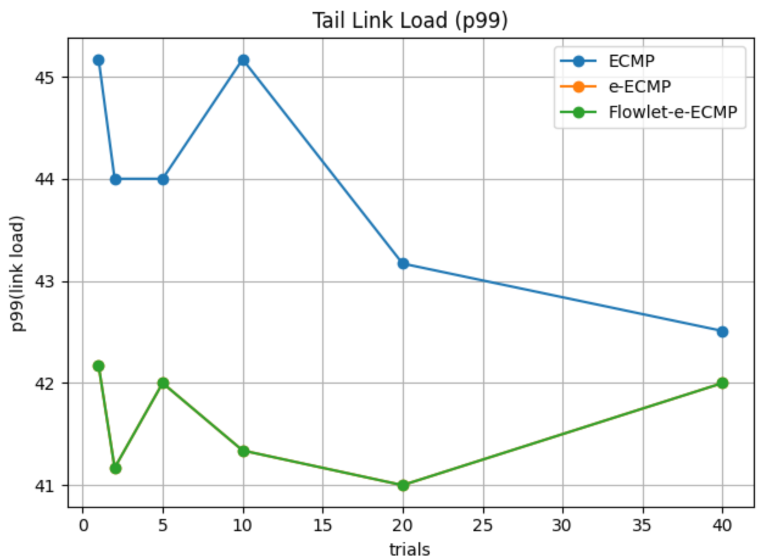

Figure 3: Tail link load (p99) vs. number of flowlet epochs.

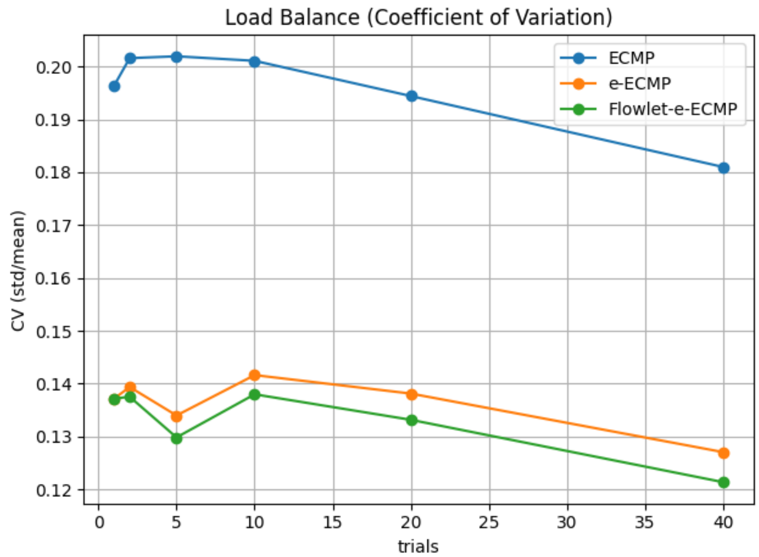

Figure 4: Load balance (CV) vs. number of flowlet epochs.

7.3 Impact of Topology Size

The topology sweep confirms that larger fat-tree networks naturally provide more path diversity, improving ECMP performance. Nevertheless, e-ECMP and flowlet-based e-ECMP consistently achieve superior load balance across all topology sizes.

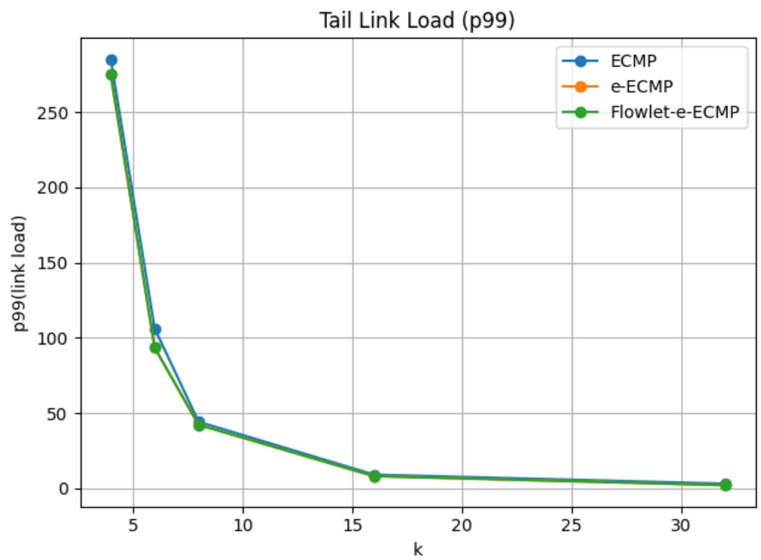

Figure 5: Tail link load (p99) vs. fat-tree size k.

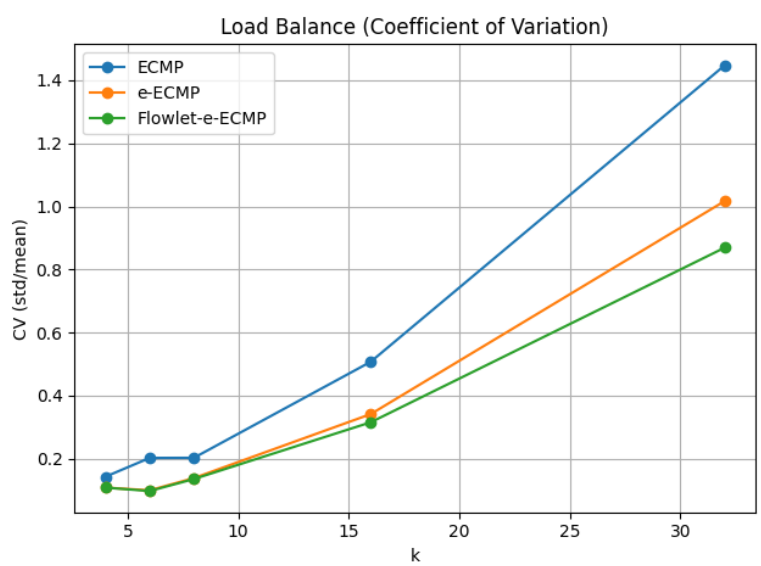

Figure 6: Load balance (CV) vs. fat-tree size k.

8 Discussion

Results indicate that spatial path diversity (via QPs) is the dominant factor in reducing congestion, while temporal reshuffling (flowlets) further mitigates persistent hash-induced hotspots.

These findings align with broader efforts toward adaptive routing in data center networks [1].

9 Limitations

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

10 Conclusion

This project presents a lightweight routing-level simulation framework for studying multipath routing in fat-tree data center networks. Comparative evaluation shows that e-ECMP substantially improves load balance over ECMP, while flowlet-based e-ECMP further reduces tail congestion through temporal path diversity. The framework provides a reproducible basis for future exploration of routing-transport interactions in large-scale data centers.

11 Reproducibility

The simulator, experiment scripts, and output plots are available in the project repository: https://github.com/KenziVisor/fat-tree-topology-sim-with-ecmp/tree/main

References

[1] Mohammad Al-Fares et al. Hedera: Dynamic flow scheduling for data center networks. In *Proceedings of USENIX NSDI*, 2010.

[2] Mohammad Al-Fares, Alexander Loukissas, and Amin Vahdat. A scalable, commodity data center network architecture. *Proceedings of ACM SIGCOMM*, 2008.

[3] Daniel Firestone et al. Rdma over ethernet for distributed ai training at meta scale. In *Proceedings of the ACM SIGCOMM Conference*, 2022.

[4] Christian Hopps. Equal-cost multi-path routing in ip networks. 2000.