MACIEJ BESTA

18.06.2025

Graphs & LLMs: Synergy

EvalNet: A Practical Toolchain for Generation and Analysis of Extreme-Scale Interconnects

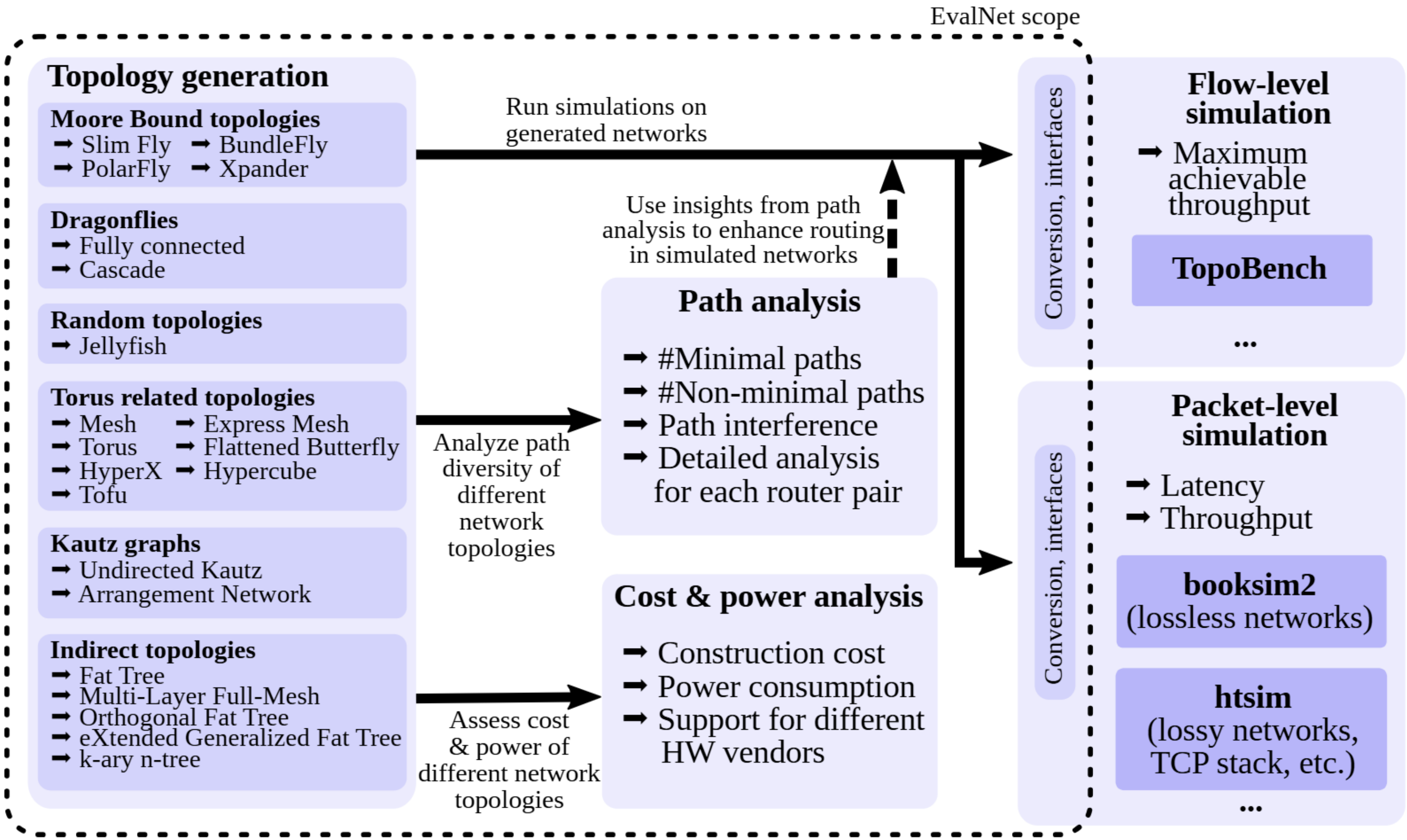
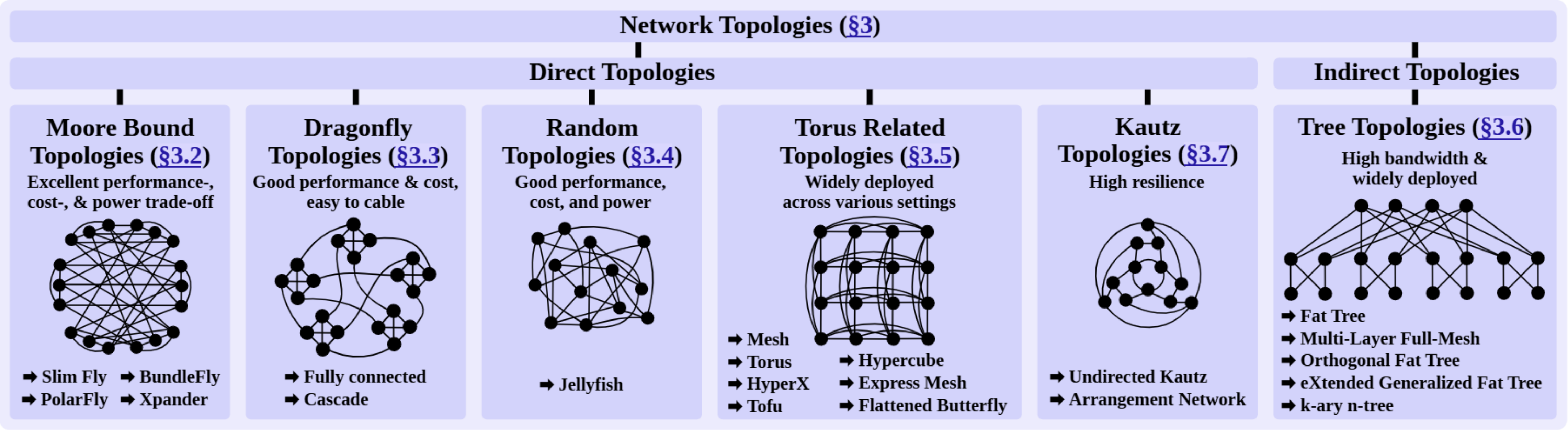


Figure 1: An overview of EvalNet.

EvalNet: A Practical Toolchain for Generation and Analysis of Extreme-Scale Interconnects

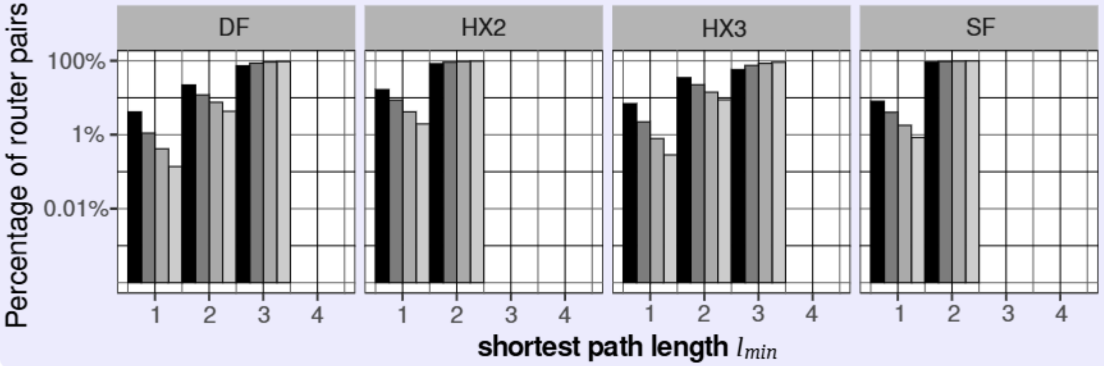


EvalNet: A Practical Toolchain for Generation and Analysis of Extreme-Scale Interconnects

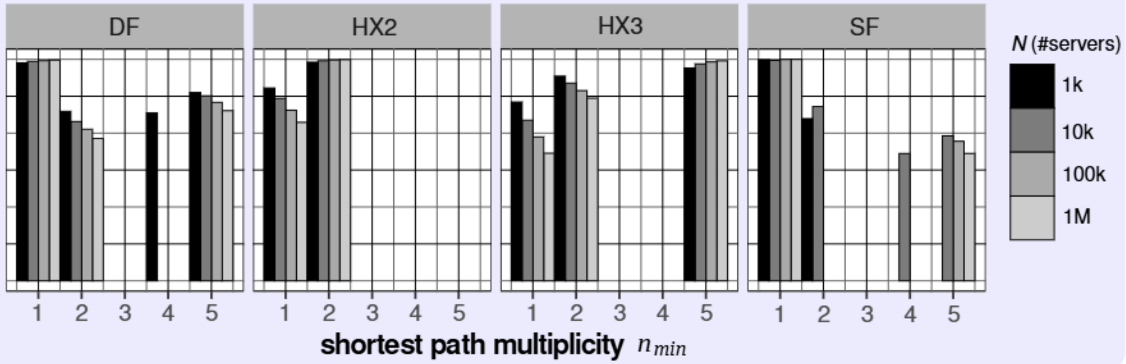
Evaluation metrics (§4)	Path diversity measures (§4.1)	Count of minimal paths	Shortest path lengths	Facilitates reasoning about the expected latency of a topology with shortest-path minimum routing, for different router pairs.
			Shortest path multiplicity	Enables taking informed decisions on whether to use multipathing over shortest paths (e.g., with ECMP), or whether one has to use non-minimal multipath routing.
		Count of non-minimal paths	Non-shortest path lengths	It facilitates reasoning about the expected latency of a topology with non-shortest path routing, for different router pairs, and offers hints on routing design.
			Non-shortest path multiplicity	Enables taking informed decisions on how to use non-minimal multipath routing.
		Path interference		Facilitates reasoning about the performance of both minimal and non-minimal routing, and of scheduling potential processes to different servers.
		Total network load (TNL)		Acts as an upper bound on the number of flows that a network can maintain without congestion.
	Cost and power (§4.3)	Construction cost		Enables comparisons between different baselines and insights into construction cost, for different cost models & equipment.
		Power consumption		Enables rudimentary comparisons between different baselines and preliminary insights into power consumption.
	Throughput and latency (§4.2)	Throughput (saturation for a given workload)		Measures how much load a network can accommodate, before its latency grows uncontrollably.
		Maximum achievable throughput (MAT)		Measures how much load a network can accommodate for a very adversarial traffic pattern, before its latency grows uncontrollably.
		Network latency		Measures the end-to-end communication delay without congestion effects in the network.
		Average path length		Provides insights into the impact of a routing design decision on the number of links that must be traversed.

EvalNet: A Practical Toolchain for Generation and Analysis of Extreme-Scale Interconnects

Histograms of **lengths of shortest paths** (between all pairs of routers)
Each color of bars (in a single plot) contitutes a separate histogram. Each such histogram illustrates the percentages of router pairs connected with shortest paths of different lengths.

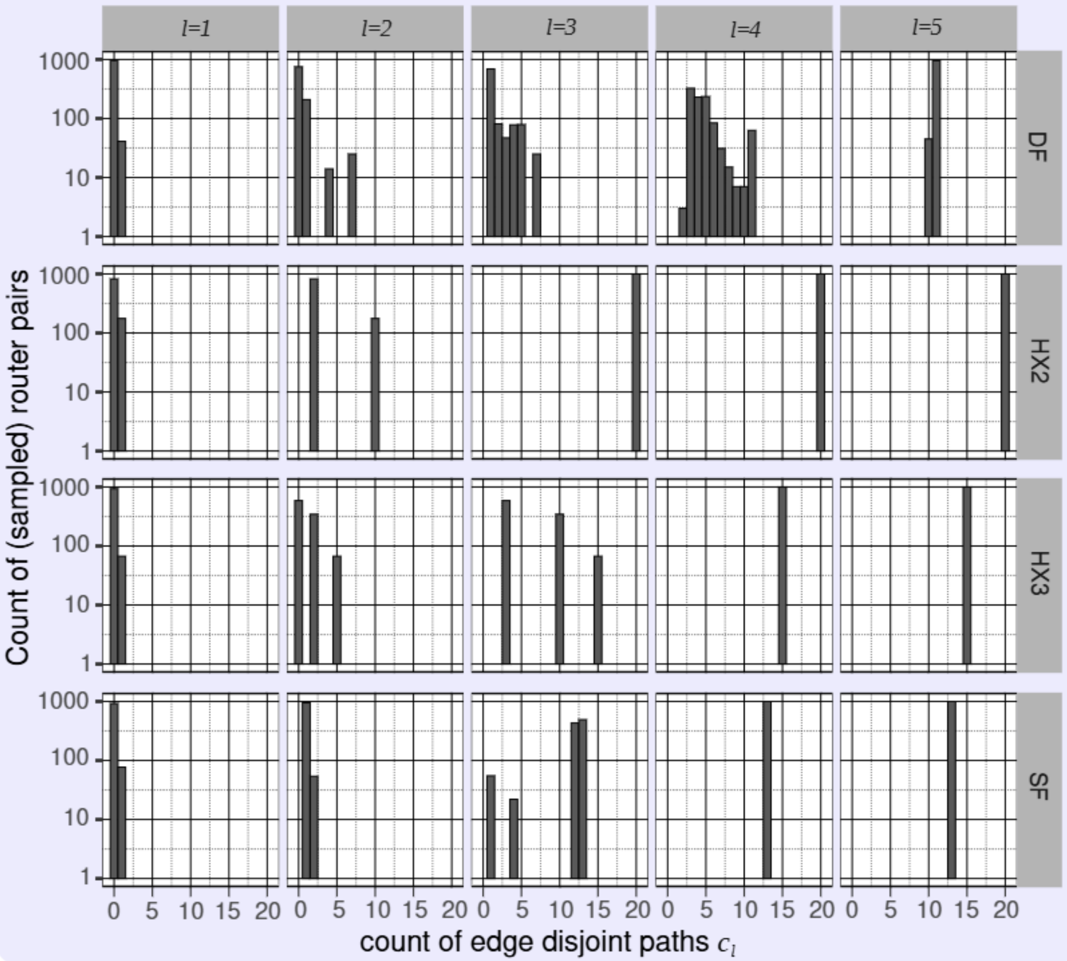


Histograms of **multiplicities of shortest paths** (between all pairs of routers)
Each color of bars (in a single plot) contitutes a separate histogram. Each such histogram illustrates the percentages of router pairs connected with a given number of shortest paths.

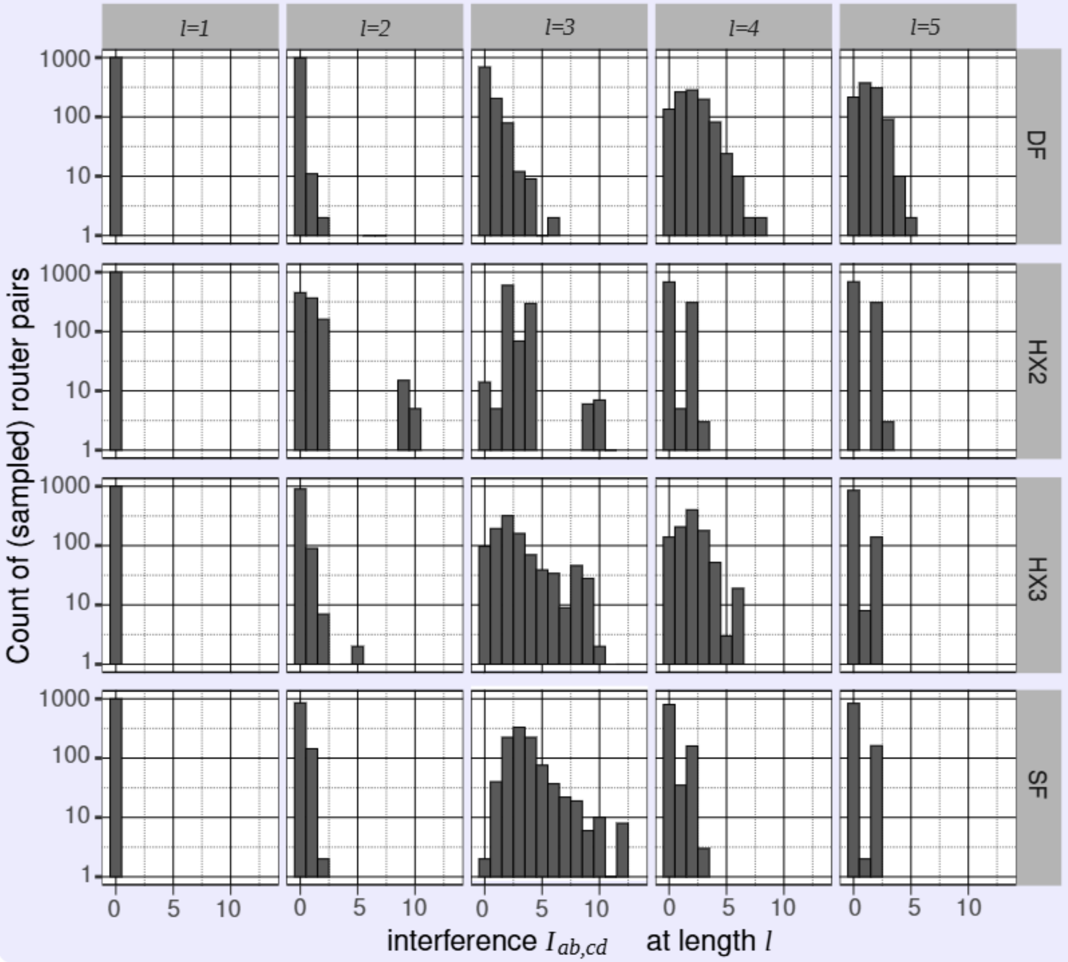


EvalNet: A Practical Toolchain for Generation and Analysis of Extreme-Scale Interconnects

Histograms of **multiplicities of non-shortest paths**
Each such histogram illustrates the counts of router pairs connected with a given number of edge-disjoint non-shortest paths of a given length (cf. columns) in a given network (cf. row).

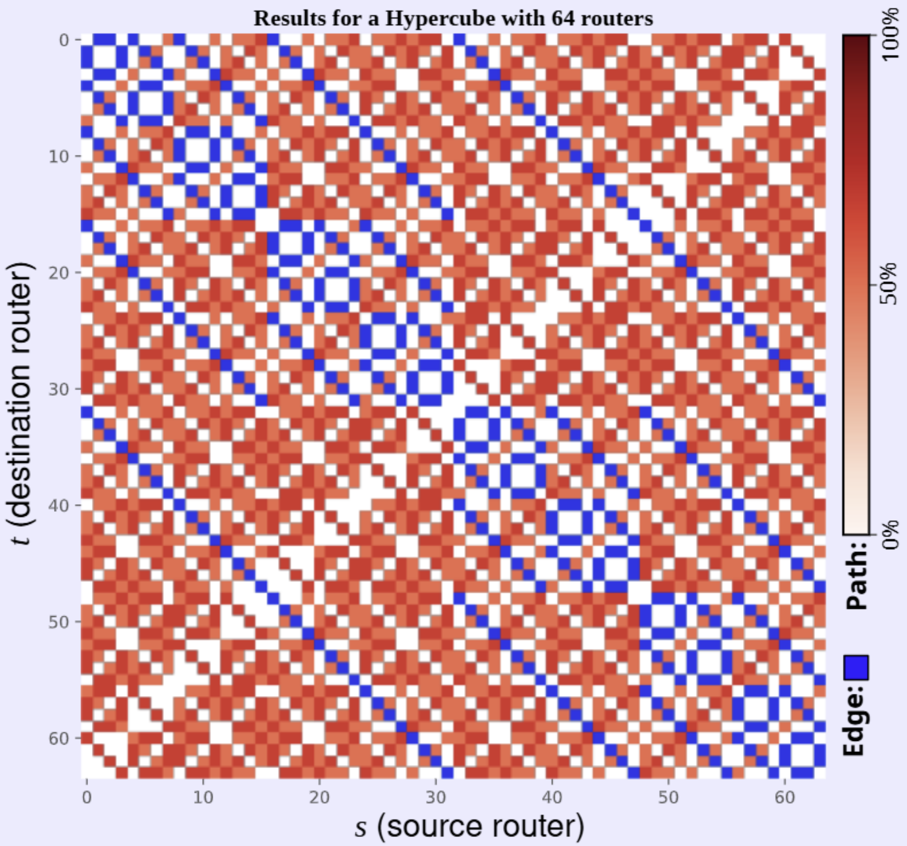


Histograms of **path interferences for shortest & non-shortest paths**
Each such histogram illustrates the counts of router pairs experiencing a given number of interfering edge-disjoint paths of a given length (cf. columns) in a given network (cf. row).

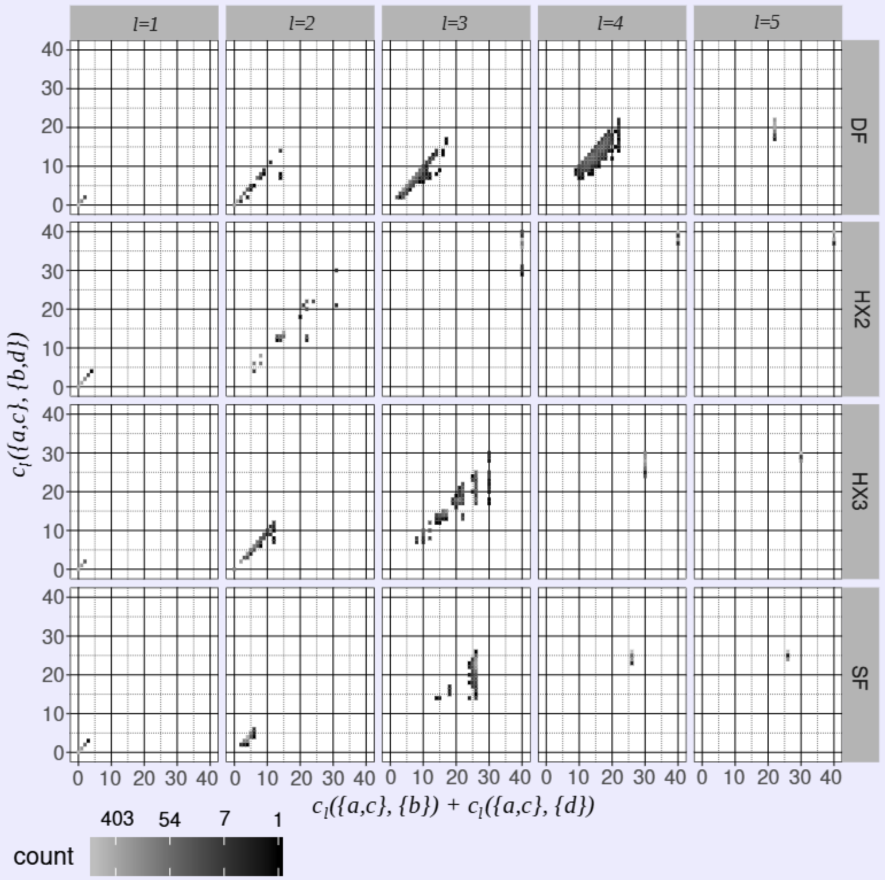


EvalNet: A Practical Toolchain for Generation and Analysis of Extreme-Scale Interconnects

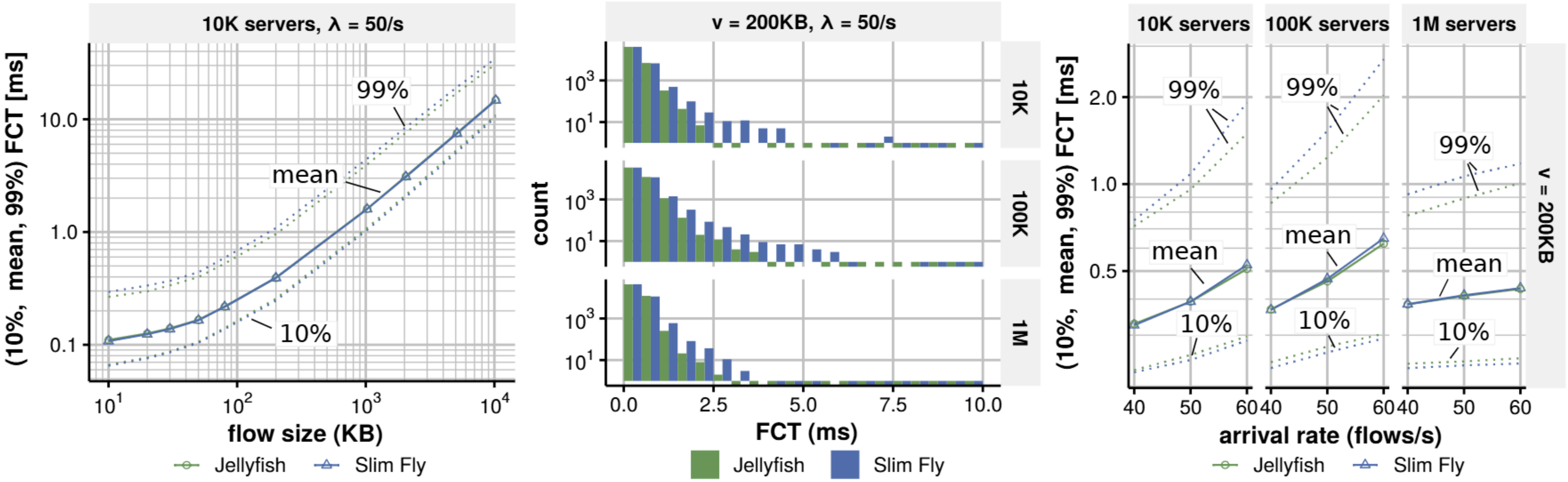
Path diversity between all router pairs, expressed as the percentage of network radix
Each point shows how well two given routers (determined by points on X & Y axes) are connected. The percentage is derived with respect to the network radix k'. Thus, 100% indicates that - for a given router pair - this pair is connected by the number of different paths that is equal to the number of available ports in each of these two routers.



2D histograms of **path interferences for shortest & non-shortest paths**
Each such histogram illustrates the counts of router pairs experiencing a given scope of interference, i.e., their actual path interference versus the potential full (non-interfering) count of edge-disjoint paths of a given length (cf. columns) in a given network (cf. row).



EvalNet: A Practical Toolchain for Generation and Analysis of Extreme-Scale Interconnects



EvalNet: A Practical Toolchain for Generation and Analysis of Extreme-Scale Interconnects

Topology	Lat.	GlB.	Ext.	C&P	MS	MnS
Slim Fly [15]	excellent	excellent	worst	good	bad	excellent
PolarFly [65]	excellent	excellent	worst	excellent	worst	excellent
Xpander [100]	good	excellent	bad	medium	bad	good
BundleFly [68]	good	good	bad	medium	bad	good
Dragonfly [64]	good	good	bad	medium	worst	medium
Cascade Dragonfly [34]	medium	good	bad	medium	bad	bad
Jellyfish [95]	medium	medium	bad	medium	bad	good
Mesh	worst	worst	good	worst	worst	bad
Torus 2D	worst	worst	good	worst	worst	bad
Torus 3D	bad	bad	good	bad	bad	medium
Torus 4D	medium	medium	good	bad	medium	good
Torus 5D	good	good	excellent	medium	good	good
Torus 6D	good	good	excellent	medium	good	excellent
Hypercube	good	good	excellent	medium	good	excellent
HyperX (2-dimensional) [2]	excellent	excellent	bad	medium	good	excellent
HyperX (3-dimensional) [2]	good	excellent	medium	medium	good	excellent
Express Mesh [54]	medium	good	good	good	medium	good
Flattened Butterfly [63]	excellent	excellent	bad	medium	good	excellent
Fat Tree (2-level) [69]	excellent	excellent	bad	good	excellent	worst
Fat Tree (3-level) [69]	good	medium	medium	medium	excellent	worst
k-ary n-tree [85]	good	excellent	excellent	medium	excellent	worst
eXtended Generalized Fat Tree [80]	good	good	excellent	medium	excellent	worst
Orthogonal Fat-Trees [60]	excellent	excellent	bad	excellent	excellent	worst
Multi-Layer Full-Mesh [60]	excellent	good	bad	medium	good	medium
Undirected Kautz [70]	medium	medium	bad	medium	bad	medium
Arrangement Network [30]	bad	medium	bad	medium	worst	bad

Table 4: A general comparison of different network topologies. Lat.: latency. GlB.: global bandwidth (i.e., the saturation point for the random uniform traffic pattern). Ext.: extensibility (i.e., how far can we extend a given concrete network with new servers?). Note that fixed diameter networks have lowest extensibility, because they have strict upper bounds on how many new servers can be attached; the lower the diameter is, the lower this bound is. Contrarily, networks such as torus can attach arbitrarily many new servers, because their diameter grows to infinity. C&P: construction cost and static power consumption. They are assessed as proportional to the total number of ports in a network, for a fixed network size N (i.e., for fixed N, networks with higher radix tend to have higher cost and power consumption). MS: diversity of shortest paths. It is assessed as a weighted average of shortest path diversities at different lengths. MnS: diversity of non-shortest paths. It is assessed as a weighted average of non-shortest path diversities at different lengths. The battery symbols serve as a rudimentary measure of relative comparison between networks. (worst): worst, (bad): bad, (medium): medium, (good): good, (excellent): excellent. They always have a positive meaning, e.g., "(excellent)" used for cost means that the cost of a given network is *very low* compared to other networks.

SC rebuttals

- EvalNet
- Higher-Order graph & LLM learning
- Sparse training/inference
- Load balancing in low-diameter networks

Major paper updates

Affordable AI Assistants with Knowledge Graph of Thoughts

CHECKEMBED: Effective Verification of LLM Solutions to Open-Ended Tasks

Multi-Head RAG: Solving Multi-Aspect Problems with LLMs

Parallel & Distributed RLMs

- More optimizations considered
  - **Hybrid Parallel Mode Switching:** Framework *re-shards* or switches parallelism scheme between training and generation to match each phase’s optimal strategy.
  - **Adaptive Pipeline Scheduling:** System treats the full RLHF loop as a pipeline-scheduling problem, exploring/optimising parallel plans across stages and mini-batches.

Parallel & Distributed RLMs

- Parallelism in RLHF: Work-Depth Analysis
  - B = number of prompt-response samples processed per iteration (batch size)
  - A = time complexity for the actor model to generate all B responses (sequentially)
  - R, V = time for reward model and value model forward passes on all B responses (if done sequentially)
  - U = time for the update stage (performing all backprop/training on the actor (and critic) for the batch)
- The baseline synchronous RLHF iteration (all stages run sequentially) has

$$\text{Work} = O(A + R + V + U)$$
$$\text{Depth} = O(A + R + V + U)$$

Parallel & Distributed RLMs

Parallel Technique	Total Work per Iteration	Critical Path Depth per Iteration
Baseline (Sequential)	$O(A+R+V+U)$	$O(A+R+V+U)$ (no parallelism)
Disaggregated Model Placement	$O(A+R+V+U)$	$O(A+max(R,V)+U)$ reward & value runs in parallel
Off-Policy RLHF	$O(A+R+V+U)$	$O(max(A+R+V,U))$ generation vs. training overlapped

Parallel & Distributed RLMs

- Asynchronous RLHF

KV cache optimizations

Quantization Hurts Reasoning? An Empirical Study on Quantized Reasoning Models

- Depends on Model Size → Smaller models are the one that suffer the most (Feels counterintuitive)
- High Impact on more difficult tasks

- ... Actually, not that unintuitive

KV cache optimizations

- Exploring a very interesting direction

Bottlenecked Transformers: Periodic KV Cache Abstraction for Generalised Reasoning

Adnan Oomerjee
UCL Centre for AI, UK
Huawei Noah's Ark, UK
adnan.oomerjee.22@ucl.ac.uk

Zafeirios Fountas
Huawei Noah's Ark, UK
zafeirios.fountas@huawei.com

Zhongwei Yu
Hong Kong University of Science and Technology, HK
zhongwei-yu@outlook.com

Haitham Bou-Ammar
Huawei Noah's Ark, UK
University College London, UK
haitham.ammar@huawei.com

Jun Wang
UCL Centre for AI, UK
jun.wang@ucl.ac.uk