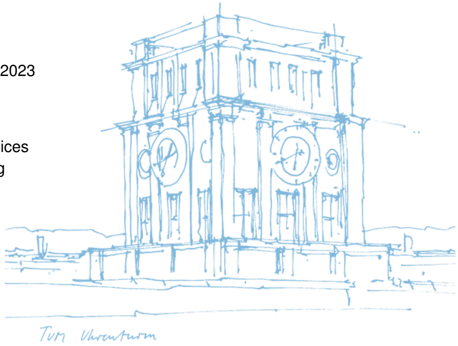
Chair of Network Architectures and Services
Department of Computer Engineering
Technical University of Munich

Predicting Latency Quantiles using Network Calculus-assisted GNNs

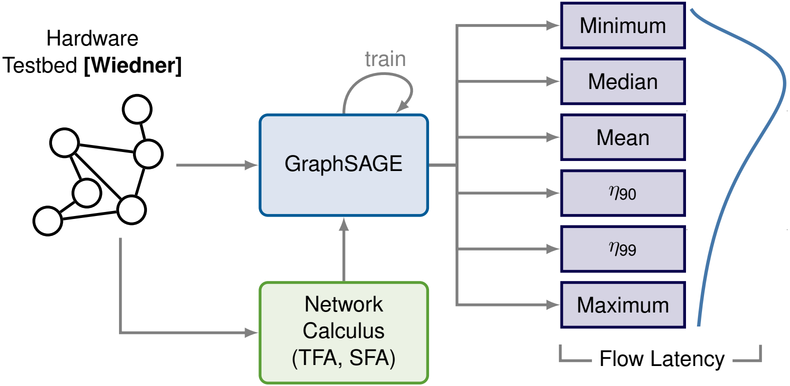
Max Helm, Georg Carle

2nd Graph Neural Networking Workshop 2023
December 8, 2023, Paris

Chair of Network Architectures and Services
Department of Computer Engineering
Technical University of Munich



What? Why?



[Wiedner] Wiedner, Florian, et al. "HVNet: Hardware-Assisted Virtual Networking on a Single Physical Host." IEEE INFOCOM 2022-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS). IEEE, 2022.

What? Why?

Questions to Answer

How well can we predict latency quantiles in a hardware setup?

Can Network Calculus bounds improve prediction performance?

Related Work

Work	Year	GNN	Formal Method as Input	Data Source	Prediction Target
Rusek et al.	2020	✓	✗	Simulation	Normal, (gamma) distribution
Ferriol-Galmés et al.	2022	✓	✗	Simulation	Mean
Wang et al.	2022	✓	✗	Simulation	Mean per timestep
Yang et al.	2022	✓	✗	Simulation	Distribution (mean, η_{99} reported)
Zhang et al.	2023	✓	✓	Simulation	Mean
Suárez-Varela et al.	2023	✓	✗	Hardware Testbed	Mean
This Work	2023	✓	✓	Hardware Testbed	Mean + Quantiles

[Rusek] Rusek, Krzysztof, et al. "Routenet: Leveraging graph neural networks for network modeling and optimization in sdn." IEEE Journal on Selected Areas in Communications 38.10 (2020): 2260-2270.

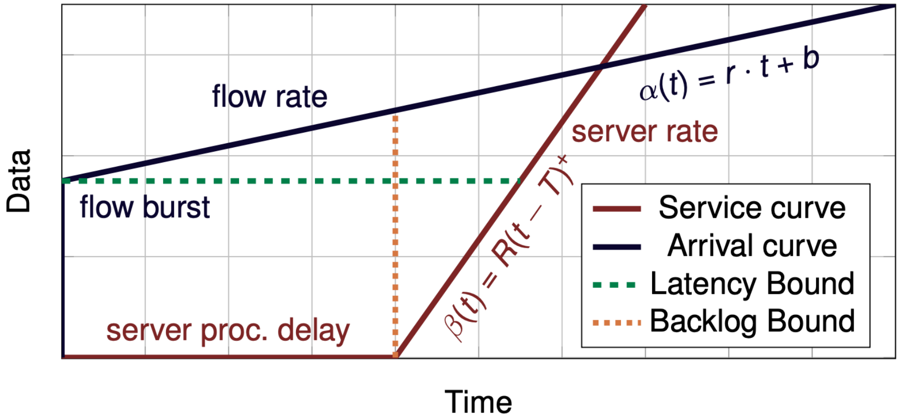
[Ferriol-Galmés] Miquel Ferriol-Galmes, Krzysztof Rusek, Jose Suarez-Varela, Shihan Xiao, Xiang Shi, Xiangle Cheng, Bo Wu, Pere Barlet-Ros, and Albert Cabellos-Aparicio. 2022. Routenet-Erlang: A Graph Neural Network for Network Performance Evaluation. In IEEE INFOCOM 2022-IEEE Conference on Computer Communications. IEEE, 2018–2027.

[Wang] Mowei Wang, Linbo Hui, Yong Cui, Ru Liang, and Zhenhua Liu. 2022. xNet: Improving Expressiveness and Granularity for Network Modeling with Graph Neural Networks. In IEEE INFOCOM 2022 - IEEE Conference on Computer Com- munications. 2028–2037.

[Yang] QingqingYang,XiPeng,LiChen,LibinLiu,JingzeZhang,HongXu,Baochun Li, and Gong Zhang. 2022. DeepQueueNet: Towards Scalable and Generalized Network Performance Estimation with Packet-level Visibility. In Proceedings of the ACM SIGCOMM 2022 Conference. 441–457.

[Zhang] LianmingZhang,BenleYin,QianWang,andPingpingDong.2023.GraphNeural Network-based Delay Prediction Model Enhanced by Network Calculus. In 2023 IFIP Networking Conference (IFIP Networking). IEEE, 1–7.

[Suárez-Varela] Jose Suarez-Varela et al. 2021. The graph neural networking challenge: a world- wide competition for education in AI/ML for networks. ACM SIGCOMM Computer Communication Review 51, 3 (2021), 9–16.


flow rate
$\alpha(t) = r \cdot t + b$
server rate
$\beta(t) = R(t - T)^+$
flow burst
server proc. delay
Service curve
Arrival curve
Latency Bound
Backlog Bound
Data
Time

Network Calculus Primer

Total Flow Analysis **[Bondorf]**:

$$D_{s_i} = \begin{cases} h\left(\alpha_{s_i}, \beta_{s_i}\right) & \text{if } |F(s_i)| = 1 \text{ (FIFO per µFlow)} \\ bp\left(\alpha_{s_i}, \beta_{s_i}\right) & \text{otherwise} \end{cases}$$

$$D^{\text{TFA}}_{P(\text{foi})} = \sum_{s_i \in P(\text{foi})} D_{s_i}$$

- Aggregate flows
- Calcluate latency bound per hop
- Sum up bounds along flow path

Separate Flow Analysis **[Bondorf]**:

$$\beta^{\text{l.o.foi}}_{s_i} = \beta_{s_i} \ominus \alpha^{x(\text{foi})}_{s_i}, \qquad \beta^{\text{l.o.}^{\text{SFA}}\text{foi}}_{P(\text{foi})} = \bigotimes_{s_i \in P(\text{foi})} \beta^{\text{l.o.foi}}_{s_i}.$$

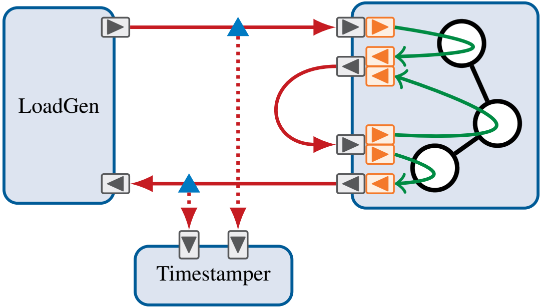
$$D^{\text{foi}} = h\left(\alpha^{\text{foi}}, \beta^{\text{l.o.}^{\text{SFA}}\text{foi}}_{P(\text{foi})}\right),$$

- Calculate left-over service curve per hop
- Convolute left-over service curves along the path
- Calculate latency bound

[Bondorf] Bondorf, Steffen. Worst-Case Performance Analysis of Feed-Forward Networks–An Efficient and Accurate Network Calculus. Diss. Technische Universität Kaiserslautern, 2016.

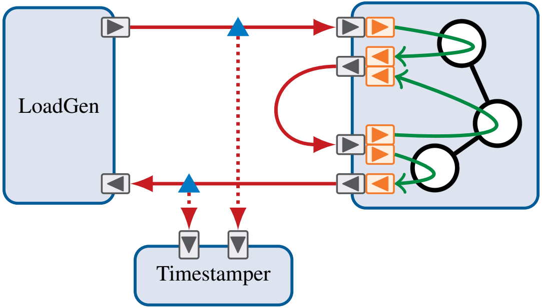
Measurements

Setup **[Gallenmüller], [Wiedner]**:

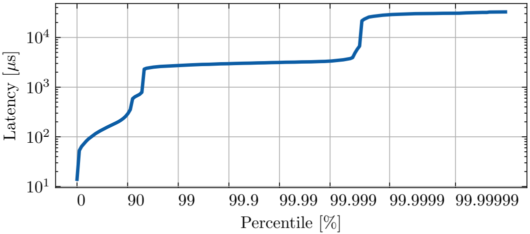


Measurements

Setup **[Gallenmüller], [Wiedner]**:



Results **[Wiedner]**:

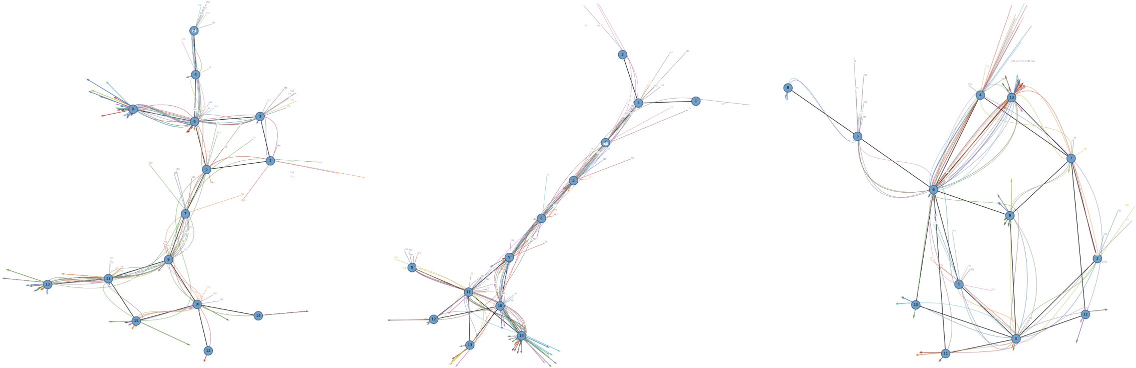


- Measured 14 billion latency values

[Gallenmüller] Gallenmüller, Sebastian, et al. "How Low Can You Go? A Limbo Dance for Low-Latency Network Functions." Journal of Network and Systems Management 31.1 (2023): 20.

[Wiedner] Wiedner, Florian, et al. "HVNet: Hardware-Assisted Virtual Networking on a Single Physical Host." IEEE INFOCOM 2022-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS). IEEE, 2022.

Example Network Topologies

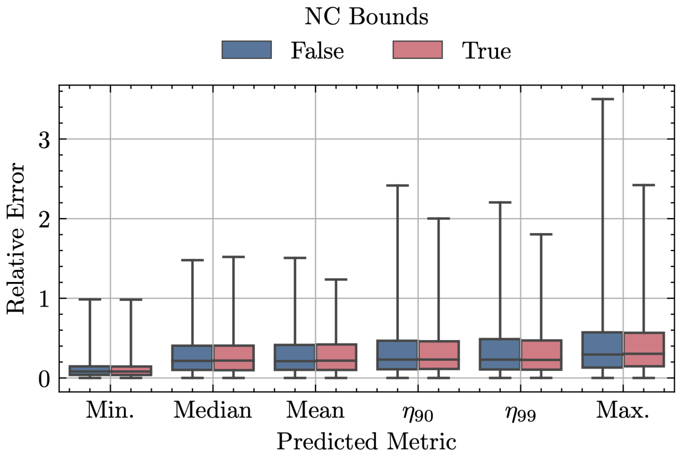


(a) Network I (b) Network II (c) Network III

Measurements on 100 random, non-isomorph network topologies

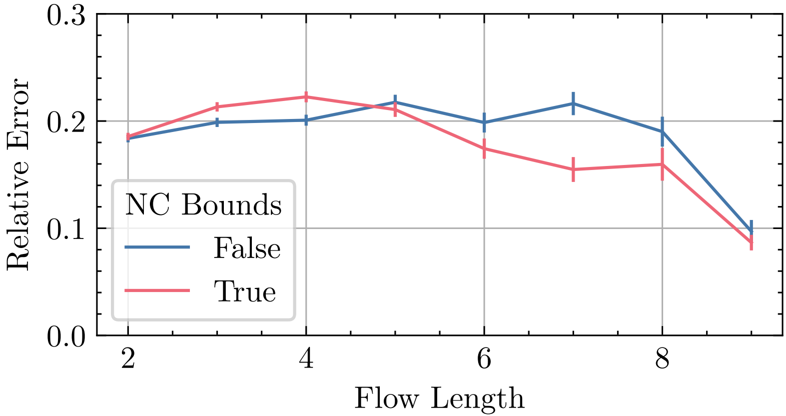
Prediction Evaluation

Latency Quantile Point Predictions



Prediction Evaluation

Latency Quantile Point Predictions



Prediction Evaluation

Importance of Network Calculus Results

Network Calculus analysis methods:

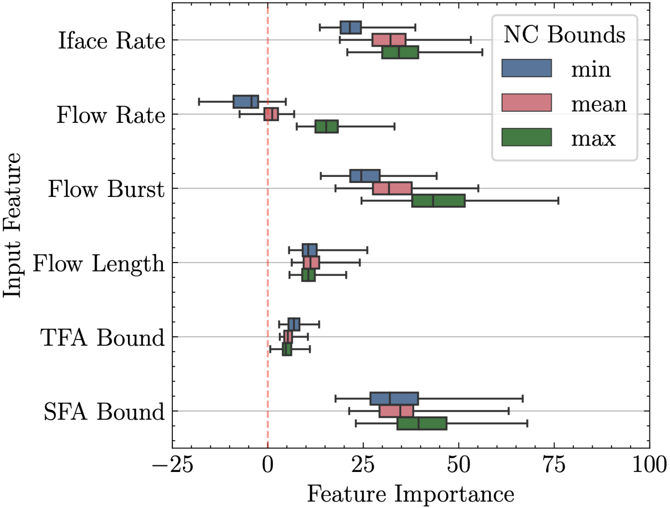
- Total Flow Analysis (**TFA**): Bounds on flow aggregates on per-hop basis
- Separate Flow Analysis (**SFA**): Bounds per flow using left-over service curves and service curve convolutions

Prediction Evaluation

Importance of Network Calculus Results

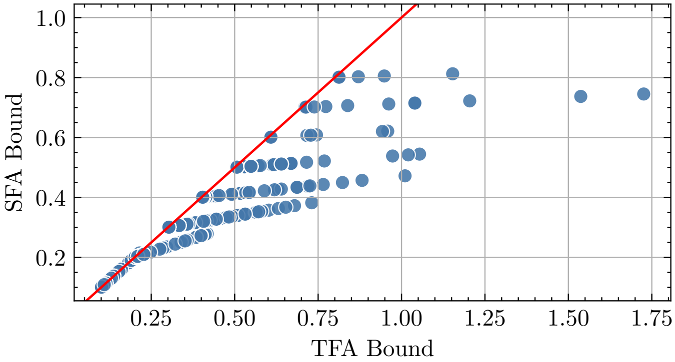
Network Calculus analysis methods:

- Total Flow Analysis (**TFA**): Bounds on flow aggregates on per-hop basis
- Separate Flow Analysis (**SFA**): Bounds per flow using left-over service curves and service curve convolutions



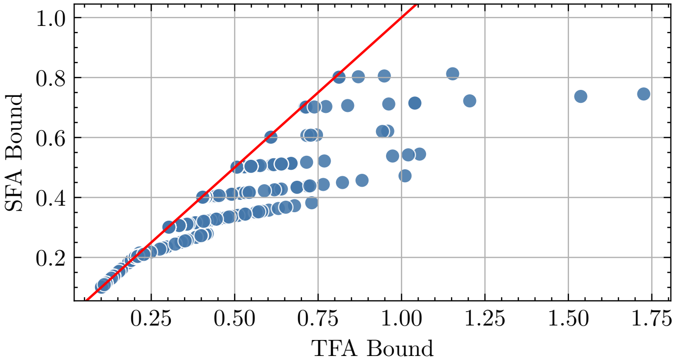
Prediction Evaluation

Importance of Network Calculus Results



Prediction Evaluation

Importance of Network Calculus Results



- SFA bounds always as tight or tighter than TFA bounds
- TFA provides worse upper bounds

Conclusion

Data:

- Simple GNN for quantile point predictions
- Based on hardware measured latencies
- 100 different network topologies
- **Limitations:** queueing, network size, sample size

Results:

- Network Calculus bounds reduce large prediction errors
- Network Calculus more useful for higher quantiles
- GNN is able to include bound tightness into decision making
- **Limitations:** GNN architecture, Network Calculus analysis methods

Paper:

Data and Code:

Questions?: helm@net.in.tum.de