Latency Prediction for 6G Dependable Networking

Dependable6G Summer School

Outline

- **Introduction + Motivation**
- Predictability of Communication Systems
- From Theory to Practical Latency Prediction
- Conclusions & Future Work

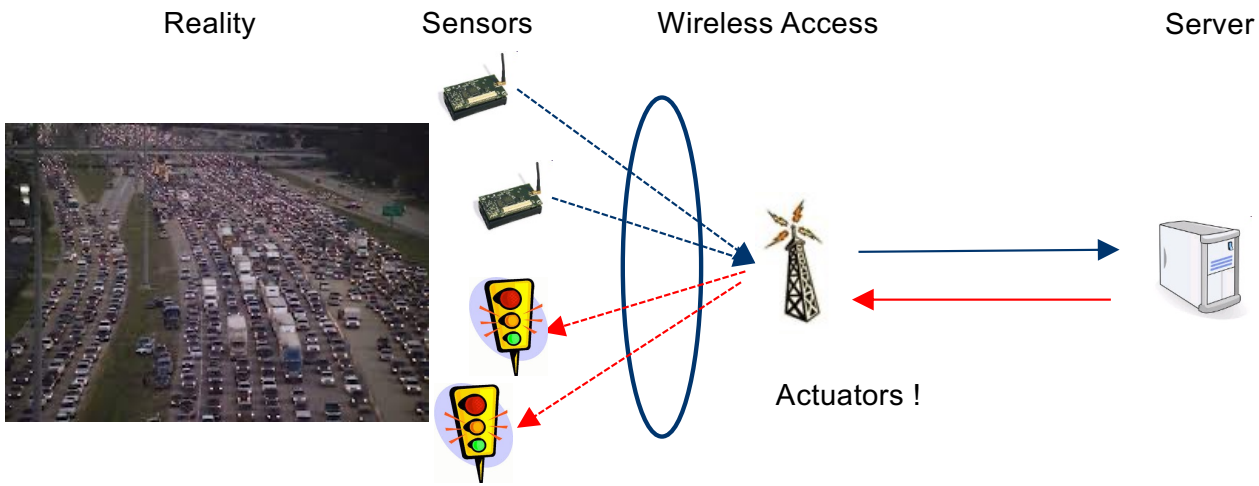
The Rediscovery of Latency during the 2010s

- Finite Blocklength Approximations
- Age of Information

WHY?

- TSN
- URLLC
- Edge Computing

Networked Cyber-Physical Systems



- From sensing applications to closed-loop control
- **Dependability** becomes the focus (latency & reliability)

URLLC: Application Fields

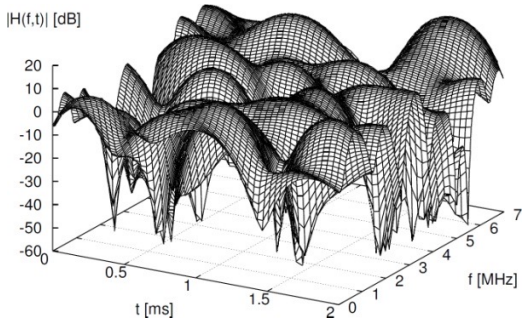
- Various application fields according to 3GPP:
  - Rail-bound mass transit
  - Building automation
  - Factory of the future / industrial automation
  - Smart living / smarty city
  - Electric power distribution & power generation

- In addition:
  - Support for autonomous devices (cars, drones, robots)
  - Human-in-the-loop applications (AR / cognitive assistance)

3GPP, TR22.804 v1.0.0, December 2017

Range of Factory Automation Requirements

- Field-Level Control
  - Cycle time: <10 ms
  - Packet sizes: < 10 byte
  - Reliability: > $1 - 10^{-6}$
- Inter-PLC Communication:
  - Cycle time: < 50 ms
  - Packet sizes: < 500 byte
  - Reliability: > $1 - 10^{-6}$



Why turn to wireless?

Reality Check

- 10 years ago expectation that
  - 5G would be driving digital transformation in automation through URLLC
  - Ubiquitous deployment of edge computing driving XR
  - TSN would have taken over the majority of field bus market

- **Today:**
  - No URLLC
  - Edge computing deployed as Telco edge, data sovereignty as driver
  - TSN has small share of field bus market (but growing)

Latency and Reliability

- Definitions
  - Latency: The time taken from when a packet arrives at the transmitter until it is successfully delivered to the receiver.
  - Reliability: The ratio of successfully delivered packets to the total number of transmitted packets.

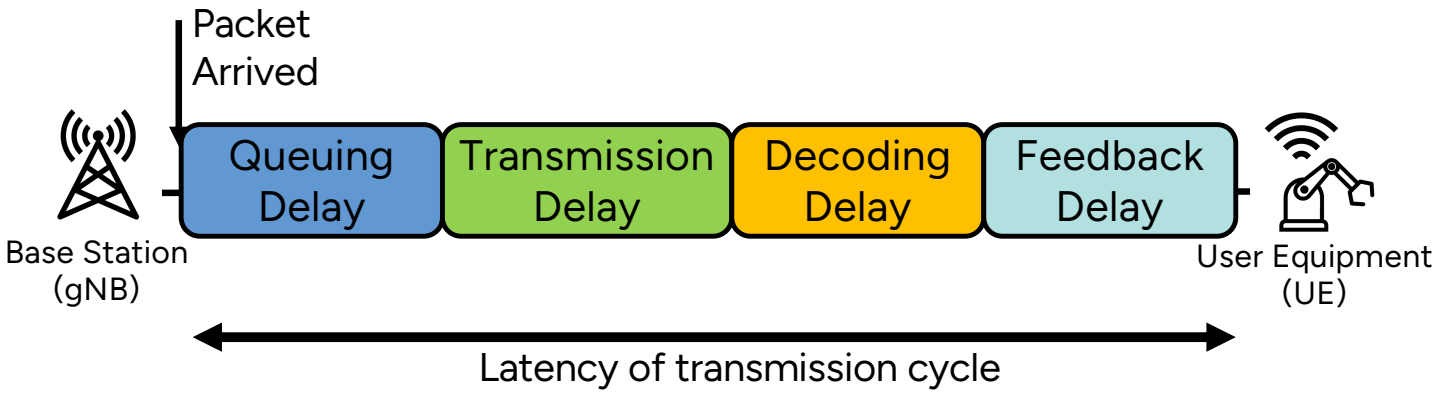
- URLLC Requirements [1]

Scenario	Reliability	Latency Requirement	Packet Size
AR/VR	99.999%	1 ms	200 Bytes
Remote driving	99.999%	3 ms	1 MB/s
Electric Power Distribution	99.9999%	3 ms	100 Bytes
Industrial Automation	99.9999%	1 ms	32 Bytes

[1] Z. Zhu *et al.*, "Research and Analysis of URLLC Technology Based on Artificial Intelligence," in *IEEE Communications Standards Magazine*, vol. 5, no. 2, pp. 37-43, June 2021, doi: 10.1109/MCOMSTD.001.2000037.

Latency in 5G

- Latency Components

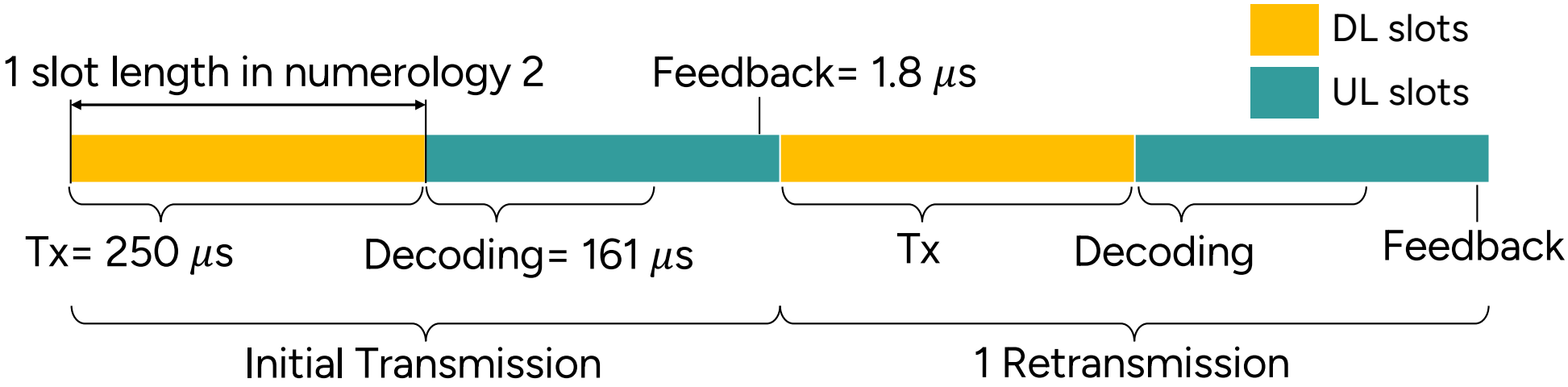


- Best-case Latency

Slot Length	Queueing	Transmission	Decoding [2]	Feedback	1 Cycle	4 Feedback Cycle	4 Repeated Cycle
1 ms	2ms	1 ms	570 us	1 ms	4.57 ms	18.28 ms	7.57 ms
0.5 ms	1 ms	500 us	357 us	500 µs	2.357 ms	9.428 ms	3.857 ms
125 us	250 us	125 us	178.5 us	125 us	0.679 ms	2.716 ms	1.054 ms

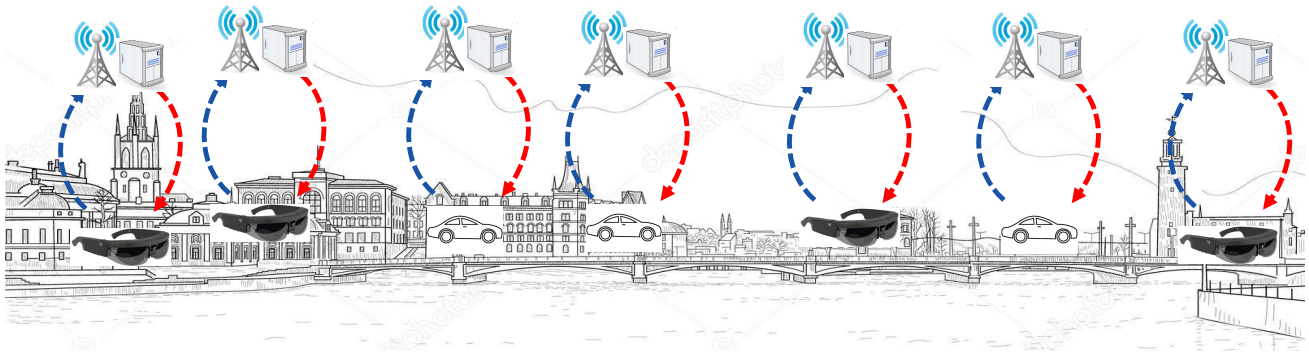
[2] "NR; Physical Layer Procedures for Data," 3GPP, TS 38.214, 03 2023, version 17.5.0.

1 ms Latency Scenario in 5G Sub6GHz band



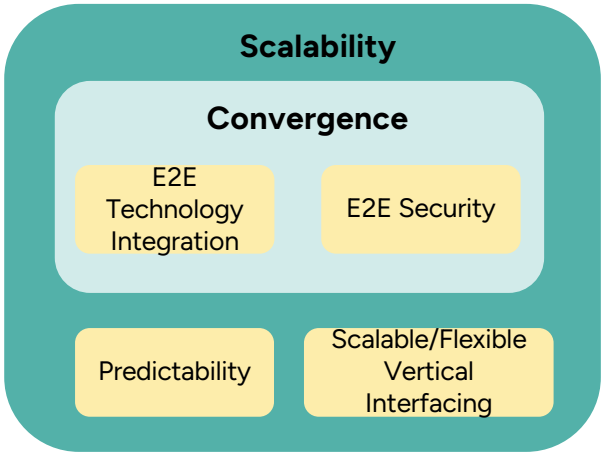
- 1 initial Tx+ 1 ReTx can be only provided with **numerology 2 within 1 ms**
  - However, numerology 2 is not mandatory
  - **Device chipset manufacturers do not realize this URLLC feature**

The 6G Story: Towards a Cyber-Physical Continuum



- Ubiquitous provisioning of CPS through mobile networks
- Last decade: Pull towards compute, latency & reliability

Key Goals of 6G Cyber-Physical Networking



- *Convergence* among different technologies to enable CPS applications
- *Scalability* of communication and compute infrastructure to support CPS applications

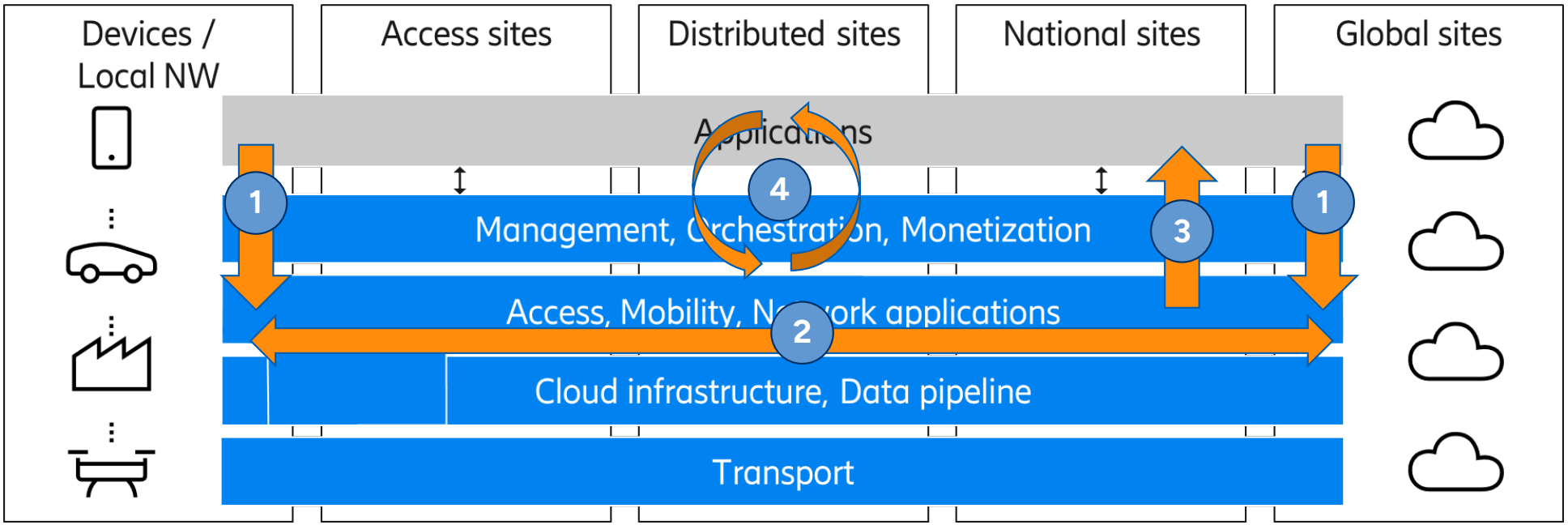
Sharma et al. "Towards Deterministic Communications in 6G Networks: State of the Art, Open Challenges and the Way Forward", *IEEE Access 2024*.

Dependable Time-Critical Communication

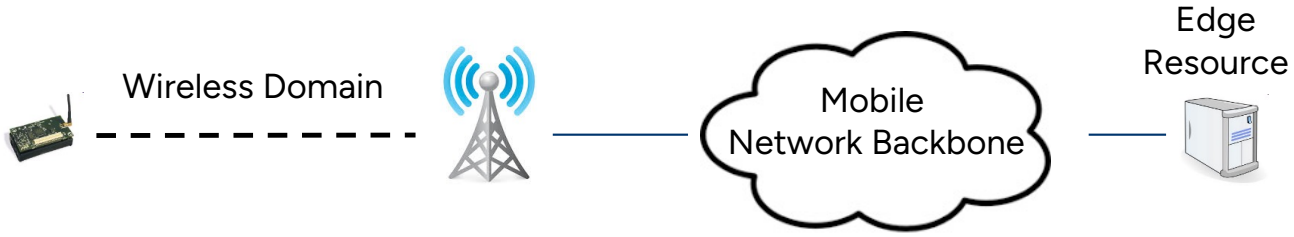
- Dependable communication
  - *Quantitatively ascertain the delivery of required service performance for the communication that are agreed*
  - *Identity upfront when these levels cannot be reached!*

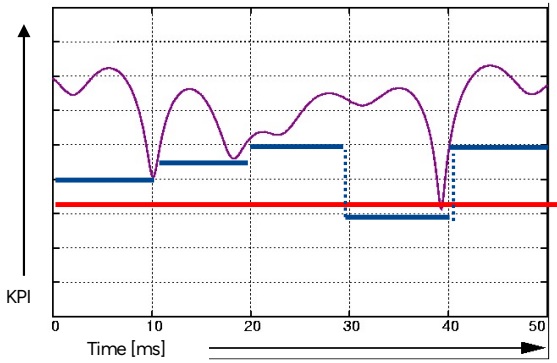
- Comprises several steps
  1. Clarity on required and agreed service performance
  2. Monitoring and prediction of delivered service performance
  3. Automated service assurance
  4. Feedback on service delivery to the application domain

Architecture Outline



Towards Predictability and Adaptation

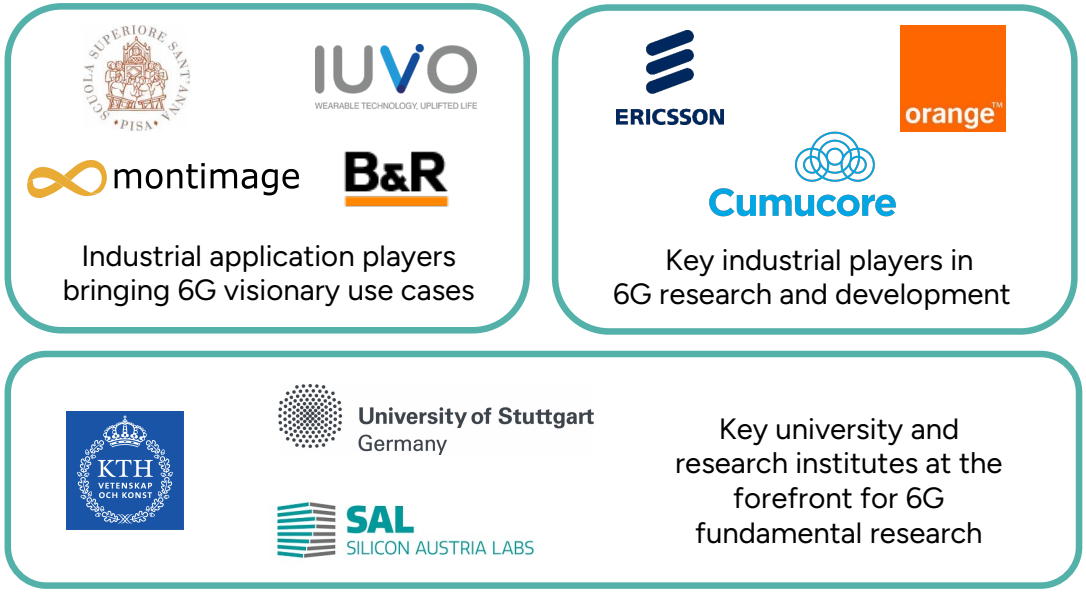




Alternative Approach: Predict network KPI and harmonize with application at run-time.

Current Approach: Static requirements from applications, translated into large footprint in the infrastructure

DETERMINISTIC6G



Leadership:
Ericsson GmbH &
KTH Stockholm

Jan 2023 – Jun
2025 (30 months)

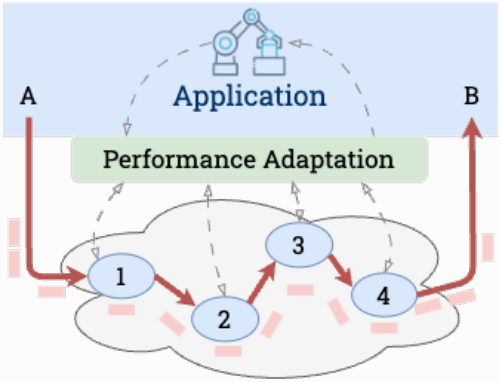
€ 5.8 M€

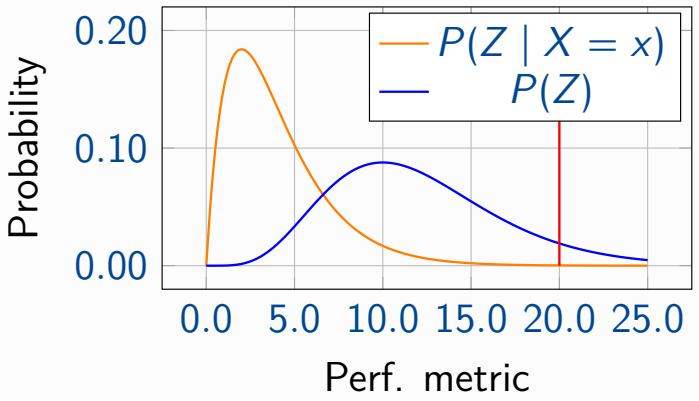
https://deterministic6g.eu/

Outline

- Introduction + Motivation
- **Predictability of Communication Systems**
- From Theory to Practical Latency Prediction
- Conclusions & Future Work

Introduction





Predicting QoS KPIs such as end-to-end delay in advance

- Enables proactive adaptation
- Probabilistic guarantees on end-to-end performance e.g. 99.999% reliable

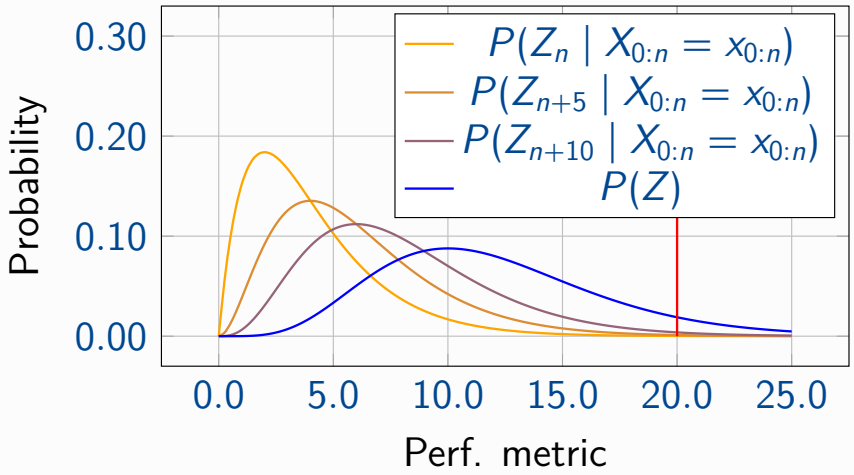
Performance (denoted by r.v. Z): data rate or delay
Observations/conditions (denoted by r.v. X)

Distribution of performance for L time steps in future, given all observations until n

$$P(Z_{n+L} \mid X_{0:n} = x_{0:n})$$

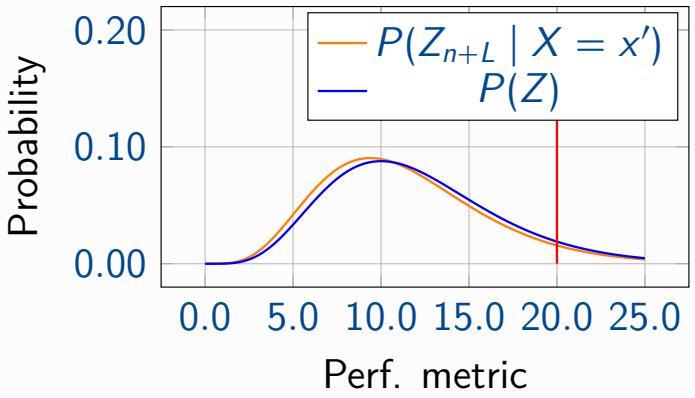
we call **forecast distribution**.

- As $L \to \infty$, for any system, $P(Z_{n+L} \mid X_{0:n} = x_{0:n}) \to P(Z_{n+L})$

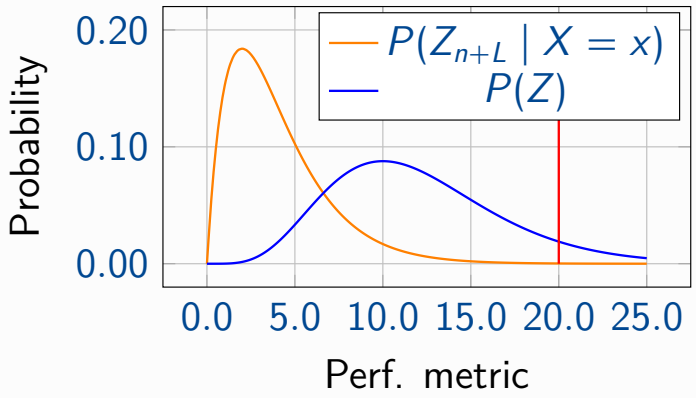


Title	Research Field	Year	Predictability Measure
Are US stock index returns predictable? Evidence from automatic autocorrelation based tests	Finance	2013	Autocorrelation
Model-free quantification of time-series predictability	Time Series Forecasting	2014	Permutation Entropy
On the predictability of infectious disease outbreaks	Epidemiology	2019	Permutation Entropy
Limits of Predictability in Human Mobility	Human Mobility	2010	Entropy Measures
Limits of Predictability for Large-Scale Urban Vehicular Mobility	Transportation	2014	Entropy (Fano's inequality)
On the Limits of Predictability in Real-World Radio Spectrum State Dynamics	Communication Networks	2015	Entropy (Fano's inequality)
Predictability and Information Theory: Measures of Predictability	Atmospheric Sciences	2004	Predictive Information, Mutual Information, etc

Table 1: Summary of Related Works on Predictability Measures



System I: Unpredictable



System II: Predictable?

Idea: The system is unpredictable if consideration of the observations makes no difference in the forecast distribution.

Mostafavi et al. "Predictability of Performance in Communication Networks under Markovian Dynamics," *IEEE Trans. VT, 2025.*

Definition: A system is unpredictable if

$$Pr(Z_{n+L}|X_{0:n} = x_{0:n}) = Pr(Z_{n+L}), \tag{1}$$

when the future state Z_{n+L} is statistically independent of the observations $x_{0:n}$.

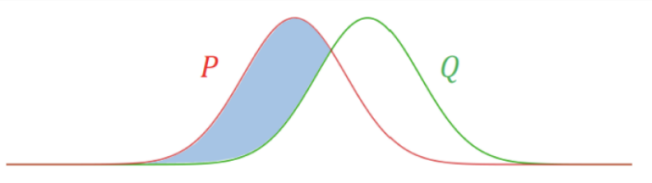
Predictability is a combined property of **the system** and **the observations**.

[1] T. DelSole,
Predictability and Information Theory. Part I: Measures of Predictability,
Journal of the Atmospheric Sciences, vol. 61, no. 20, Oct. 2004.

[2] T. DelSole and M. K. Tippett,
Predictability: Recent insights from information theory,
Reviews of Geophysics, vol. 45, no. 4, Dec. 2007.

Cont.: Predictability measure is defined as the total variation distance between the forecast and marginal distributions as

$$D_n(L) = |Pr(Z_{n+L}|X_{0:n} = x_{0:n}) - Pr(Z_{n+L})|_{\text{TV}}. \tag{2}$$



Total variation distance example

For pmfs p and q, total Variation distance (TV) is a statistical metric distance defined by

$$\text{TV}(p, q) := \sup_{A \subset \mathcal{Z}} |p(A) - q(A)| = \frac{1}{2} \sum_{z \in \mathcal{Z}} |p(z) - q(z)|. \tag{3}$$

System Model

Discrete time system with time n

Subsystem m with Markov chain conditions: $X_n^{(m)}$

- $P(x, y) = \Pr(X_{n+1} = y \mid X_n = x)$
  Transition probability from state x to state y.
- L step state transition probability $P^L(x, y)$.
- $P^L(x, y) \to \pi(y)$ as $L \to \infty$.

Observability defects:

- Delayed observations, partial observations, aggregated states e.g. $O_n = X_{n-d}$

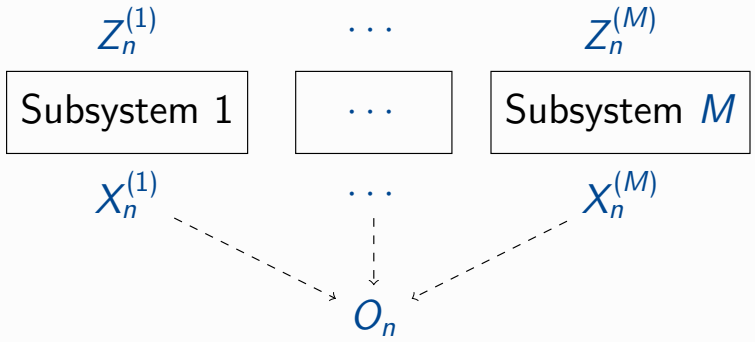


Figure 1: Multi-hop communication system model with observable measures being conditions and performance.

Predictability Analysis

Theorem 1: Predictability of a Markov-modulated process Z_n with Markov chain probabilities $\{P, \pi\}$ and posterior distributions $r_y(z)$:

$$D_n(L) = \frac{1}{2} \sum_{z \in \mathcal{Z}} \left| \sum_{y \in \mathcal{X}} (P^L(x, y) - \pi(y)) r_y(z) \right|. \tag{4}$$

Lemma: (Subadditivity of Predictability) The predictability of independent tandem multi-hop systems, is upper-bounded via the sum of predictability of each hop as

$$D_n(L) \leq \sum_{m=1}^{M} D_n^{(m)}(L). \tag{5}$$

Goals

- Assess predictability under conditions of imperfect observations.
- Determine how the randomness of the condition transitions influences predictability.
- Derive solutions for the predictability of sojourn time in Geo/Geo/1/K queues.

Geo/Geo/1/K queue
Performance metric: the sojourn time Z_n
State: system size X_n
with μ and λ as service and arrival probabilities.

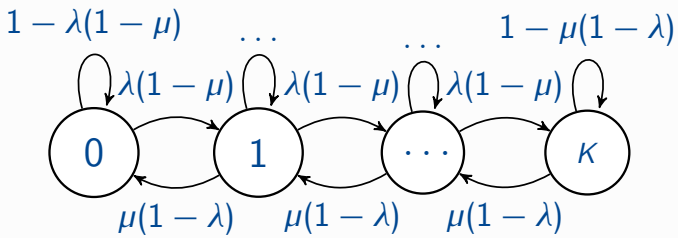


Figure 2: Geo/Geo/1 Markov chain states and transition probabilities

Numerical Evaluations

A Geo/Geo/1/K system with $\mu = 0.5$, $\rho = 0.85$, and $K = 128$.

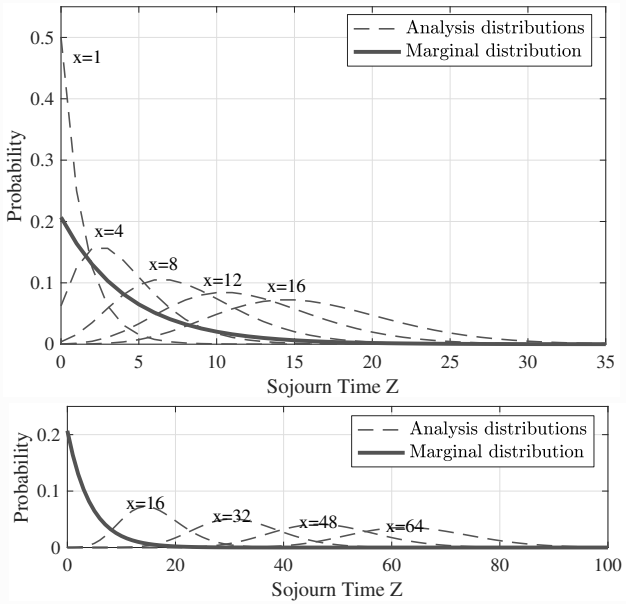


Figure 3: Posterior (analysis) and prior (marginal) distributions

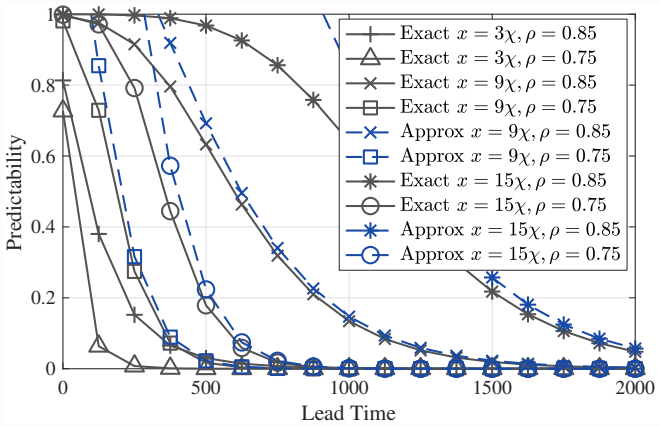
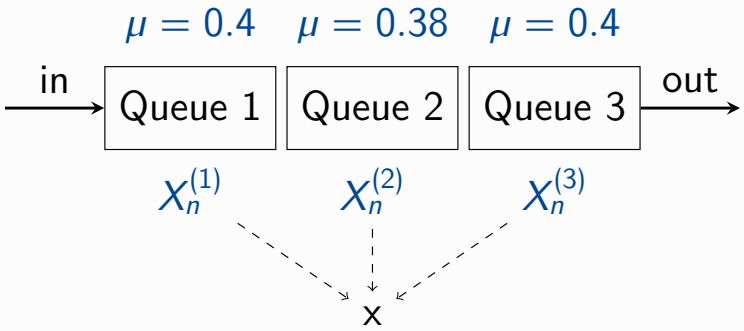


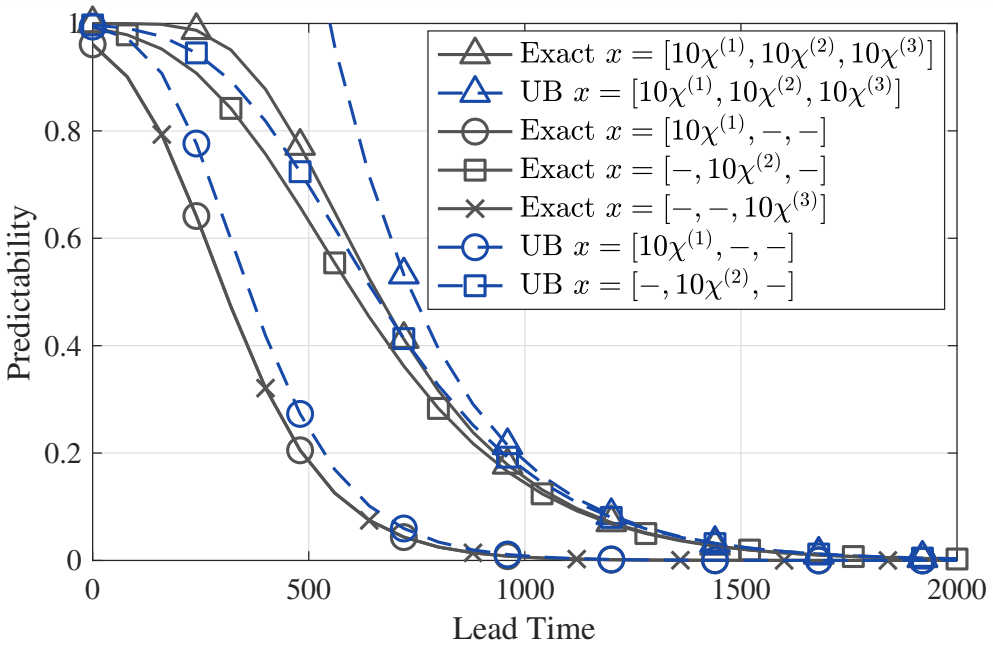
Figure 4: Exact and approximate predictability evaluations



Geo/Geo/1 queues with $\lambda = 0.34$.

UB: **Subadditivity of Predictability**.

Curves that extend further to the right represent higher predictability.

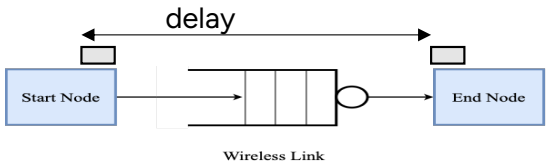


Outline

- Introduction + Motivation
- Predictability of Communication Systems
- **From Theory to Practical Latency Prediction**
- Conclusions & Future Work

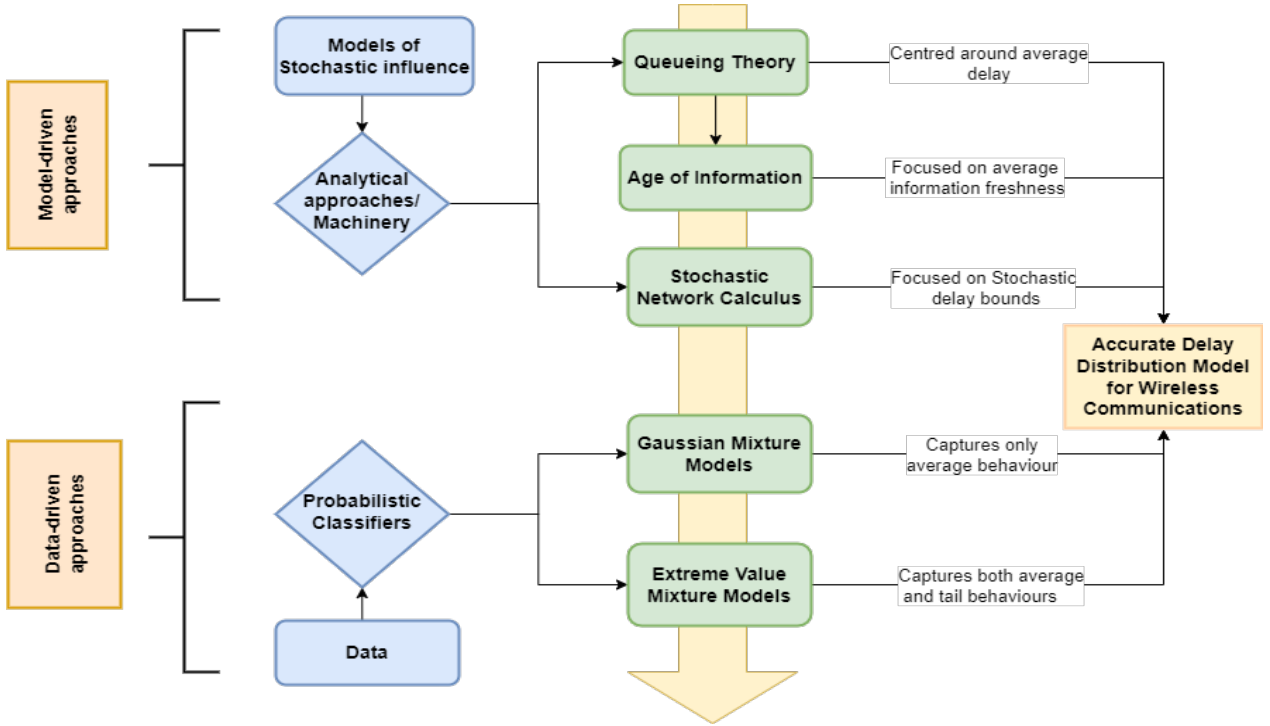
Delay Prediction Problem

- Focus on a single wireless link (for the beginning)
- Delay is a stochastic metric of a queuing system

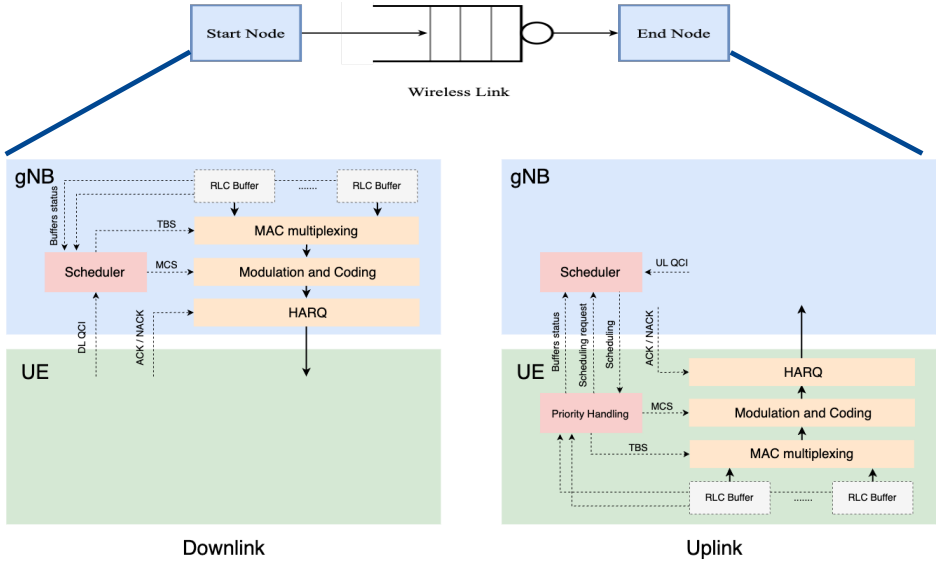


- We are interested in a prediction of delay over some time horizon -> requires a stochastic characterization of the system delay over time horizon given current conditions!

Fundamental Approaches



Model-driven Approach Realistic?



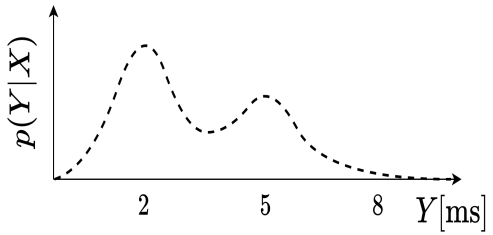
- Predictability is needed for real systems -> Cannot be captured by model-driven approaches.

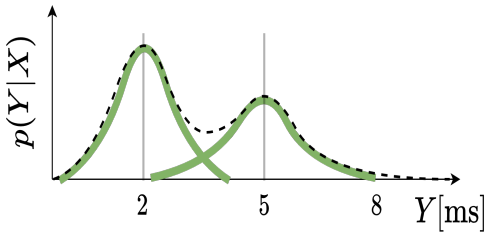
State-of-the-art: Data-driven Delay Prediction

Paradigms	Use cases and Techniques	Maturity & Specialization	References
Statistical frameworks	• Simple time-series prediction • ARIMA,SARIMA, etc	• Mature as baselines • Effective for simple trends • Struggles with non-linearity	Moreira et al. [2020] Lv et al. [2025],
Unsupervised ML	• Feature engineering and Anomaly detection • PCA, Autoencoders, K-means	• Preprocessing step • Not used for direct forecasting	Han et al.[2022]
Classical ML	• Model non-linear relationships and classification of KPI into bins • Regression, SVM combined with various boosting	• Mature & strong baselines • Interpretable & efficient • Limited for non-sequential dynamics	Khatouni et al. [2019], Flinta et al. [2020], Moreira et al [2020]. Ahmad et al [2021],
Temporal models	• Sequence modeling of delay using historical KPI data • RNN, LSTM, GRU, Transformer	• Mature & widely adopted • Captures temporal dependencies • Backbone of delay prediction • Limited Interpretability and extreme, data and compute requirements	Barmpounakis et al. [2021], Mostafavi et al. [2025], Dang et al. [2023], Kai et al. [2024], Zhou et al. [2024],

Data-Driven Approach

- Collect data from the system!
- Interpret the problem definition as conditional density estimation problem : $\hat{p}(Y|X = \boldsymbol{x}) \approx \mathbb{P}(Y|X = \boldsymbol{x})$
- Requirement: Fit a parametric density function $\hat{p}_\theta$ with parameter θ to the conditional data set

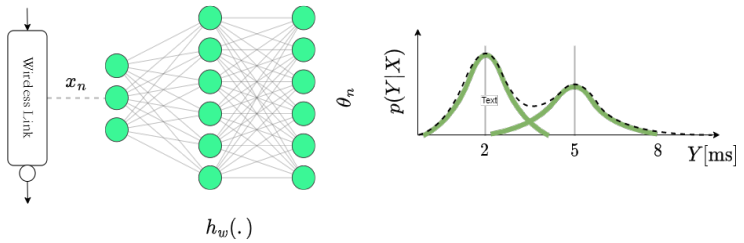




Leverage Mixture Models!

Mixture Density Networks (MDN)

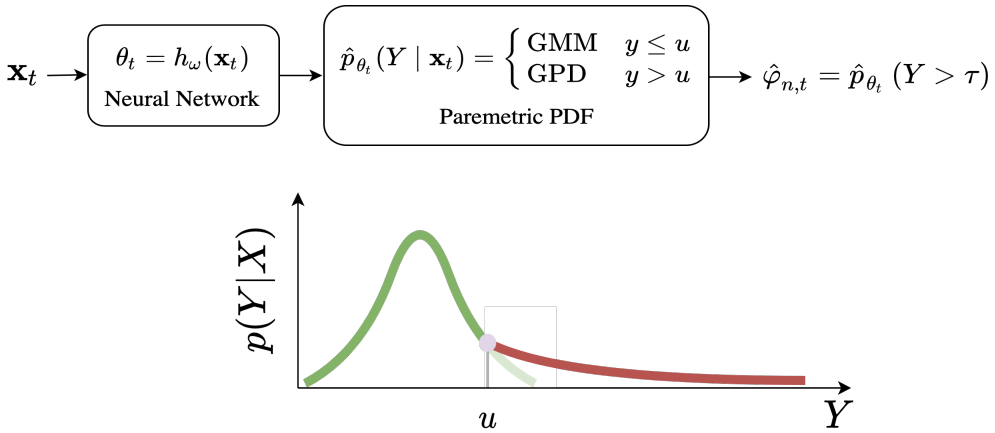
- How to find the parameters based on the conditions?
- Use a neutral network h_ω to map X to θ, i. e. $\theta_t = h_\omega(x_t)$



Through neural network training the conditional likelihood of the i.i.d samples $\{(x_1, y_1), (x_2, y_2) \ldots (x_N, y_N)\}$ is maximized and $\hat{\varphi}_{n,t} = \hat{p}_\theta(Y > \tau | X_t)$

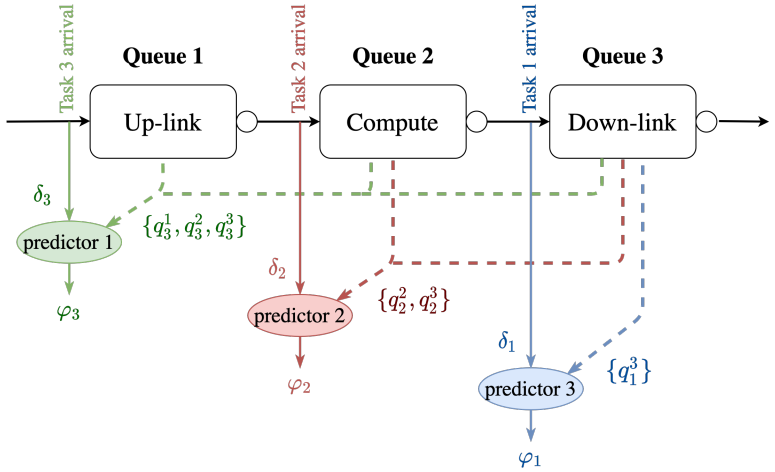
Extreme Value Theory Mixture Models

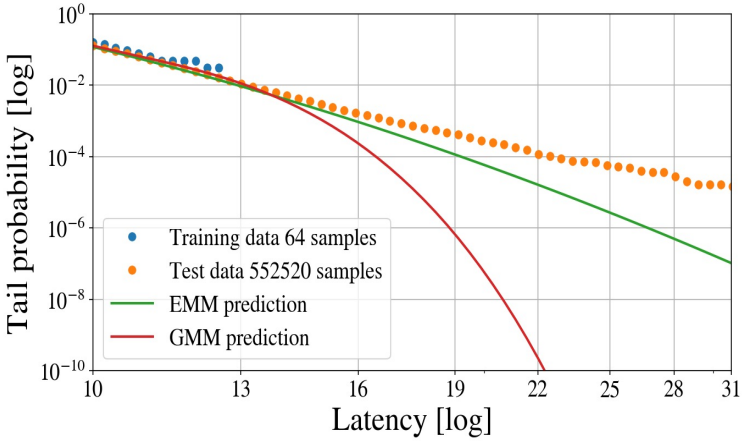
- Remaining open issue: How to devise the mixture, i.e. how many centers (random variables) and which ones?
- Choice heavily influenced by tail behavior:



Mostafavi et al. "Data-Driven End-to-End Delay Violation Probability Prediction with Extreme Value Mixture Models," *IEEE/ACM SEC 2021*.

A Synthetic (Extreme) Evaluation Example



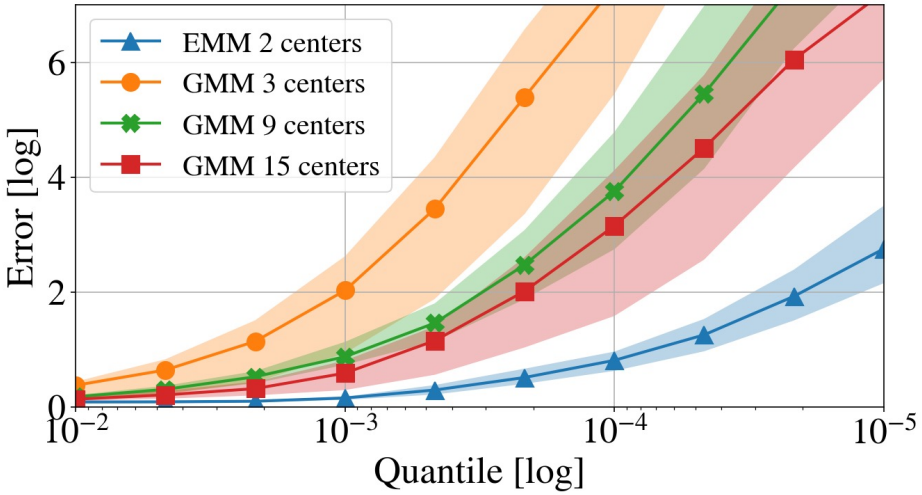


- 3-hop queueing system, system state characterized by queue length
- Gamma-distributed service

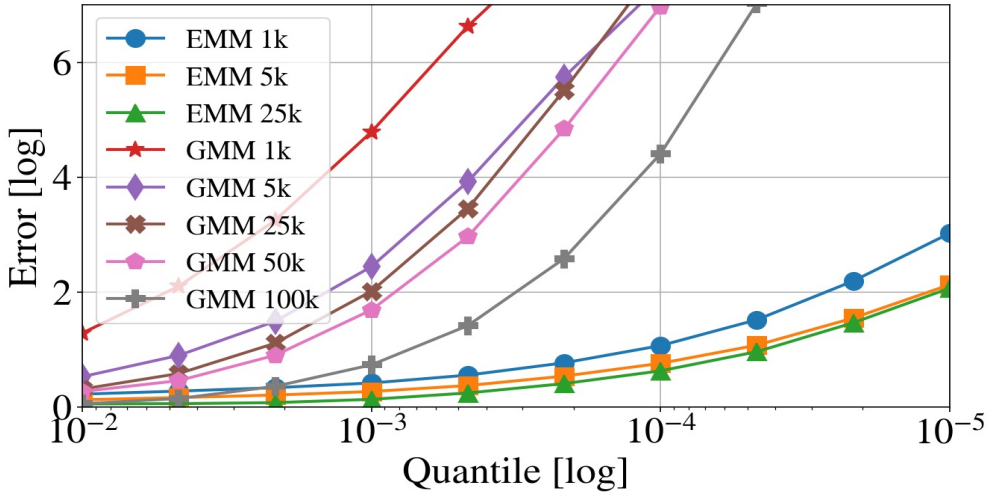
Choice of centers plays huge role!

Evaluation Example Cont.

Does adding Gaussian centers help?



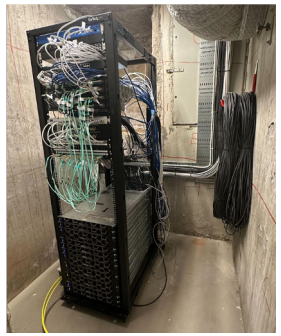
Does leveraging more data help?



Indoor Testbed

Container-based cloud and radio testbed

- 12x servers
  - PTP clock synchronized
- 24x radios
  - 10x Software-Defined Radios (SDRs)
  - 10x 5G Advantech routers as UE
  - Ericsson private 5G - 4x radio dots
- 10 TB storage
  - Block and object store



Software framework focus: Automation and Reproducibility

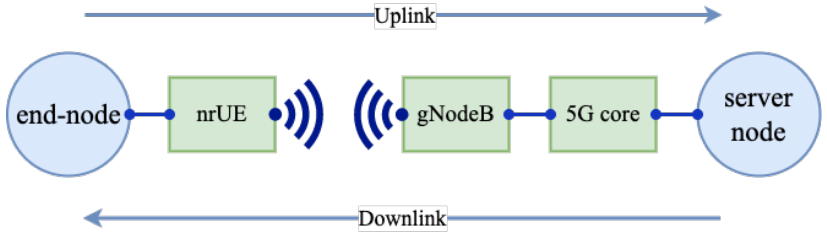
Stiftelsen för Strategisk Forskning

VINNOVA

github.com/KTH-EXPECA

Methodology

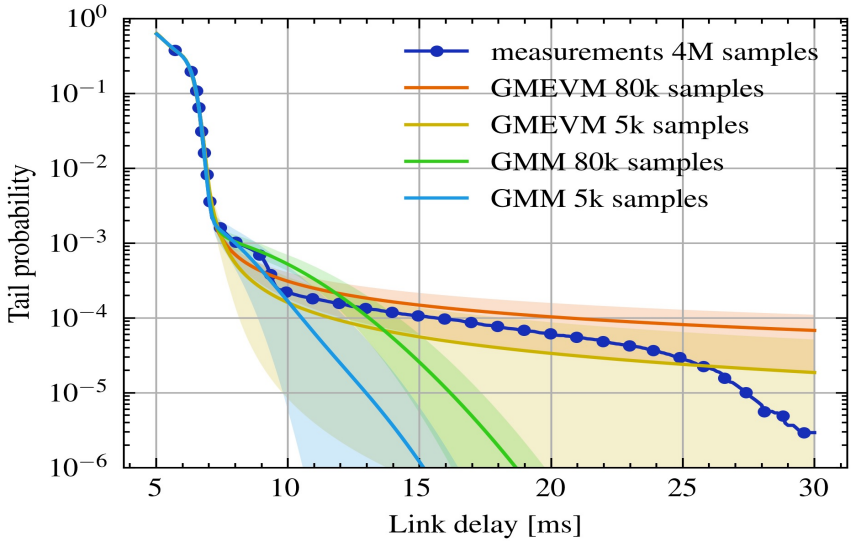
- Measurements taken on COTS and OAI 5G setups through EDAF
- Latencies (UL, DL and RTT) measured between the end node and the edge server via IRTT
- UE position and RSRP measured for the COTS 5G setup
- MCS observed in the OAI 5G setup



S. Mostafavi et al., "EDAF: An End-to-End Delay Analytics Framework for 5G and Beyond Networks", IEEE Infocom Workshop. 2024.

Latency Prediction COTS 5G

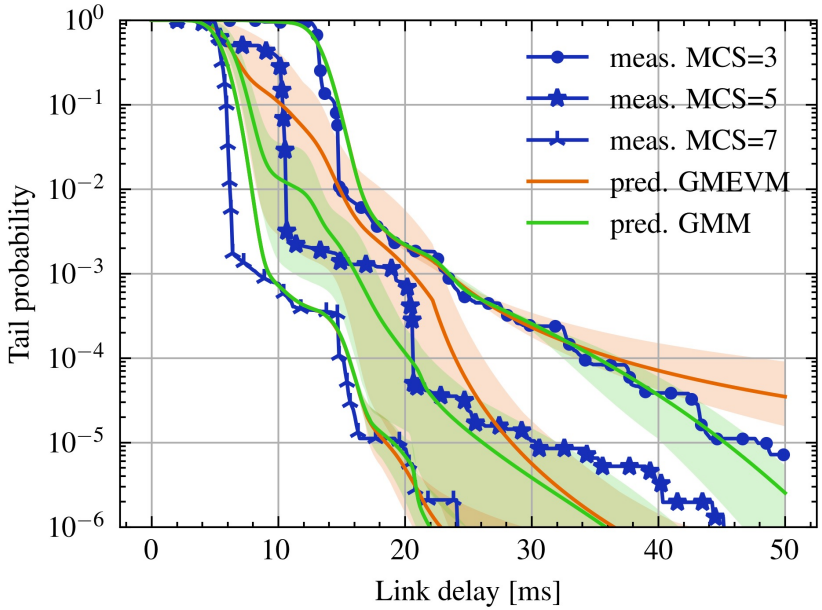
MDN	Gaussian Mixture Model (GMM) and Gaussian Mixture Extreme Value Model (GMEVM)
Neural Network	4 hidden layers ([10, 50, 50, 40])), 15 Gaussian centers
Training samples	4M samples (66 min), 80k (17 min) and 5k (5 min)
Platform (OS, hardware)	Intel(R) Core(TM) i9-10980XE CPU @ 3.00GH, 125GB RAM, 28 cores assigned



GMEVM provides better predictions than GMM

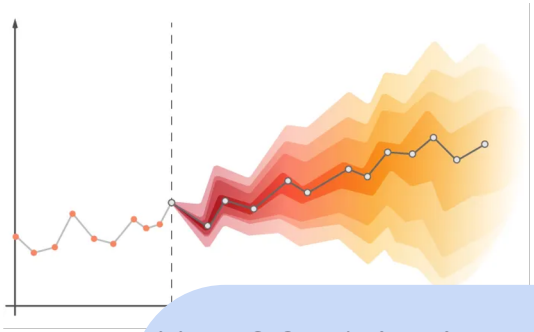
Mostafavi et al. "Data-Driven Latency Probability Prediction for Wireless Networks: Focusing on Tail Probabilities," IEEE Globecom 2023.

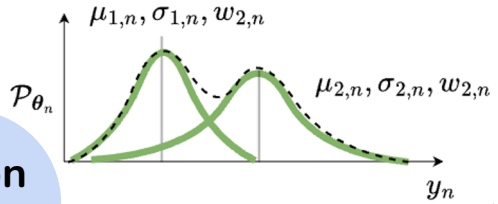
Latency Prediction of SDR 5G



- Less clear situation for SDR 5G … but good fit in both cases.

Temporal Modeling





Probabilistic data-driven prediction

Mixture density networks (MDN)
Useful predictions, uncertainty included

Use SOTA in time-series forecasting

RNNs, LSTMs, Transformers
Enable large-scale, temporal predictions

⇩

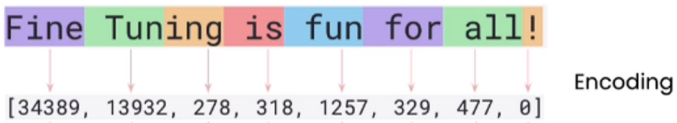
Best tools to capture
temporal dependencies

Manage data dimensions (network insights & traffic properties)

Tokenization techniques from LLMs

⇩

Compactly representing
high-dimensional, heterogeneous data

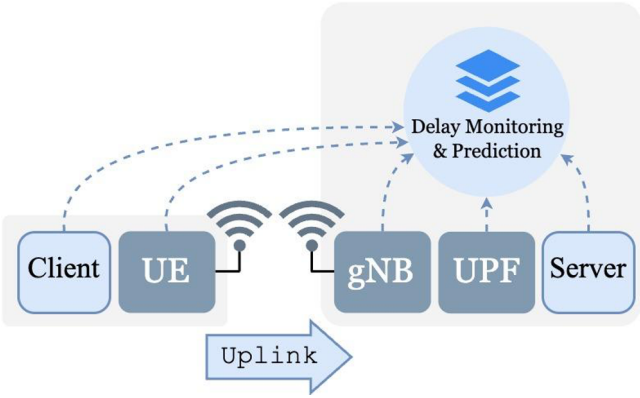


System Description

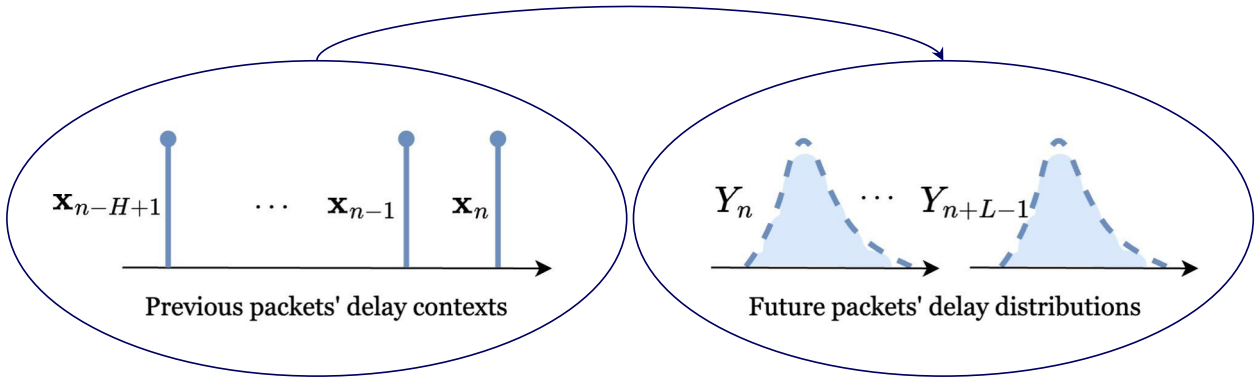
Periodic packets on 5G uplink (T_s, B_s)

Packet n's latency random variable Y_n

Key sources of packet latency include:



- Channel-Induced Delays:
  - HARQ retransmissions, RLC retransmissions if HARQ repeatedly fails
- Scheduling-Induced Delays:
  - Packets may be queued if they arrive at times unfriendly to the 5G TDD pattern or if resources are limited.



Inputs: (H) *packet delay context* vectors ($\mathbf{x}_n$):

- Traffic properties: packet size, periodicity, ...
- Network conditions: CQI, SINR, ...
- Retransmission counters: HARQ retx and RLC retx

$$P\left(Y_{n+l} \mid X_{n-H+1:n} = \mathbf{x}_{n-H+1:n}\right)$$

Objective: Predict the **delay distributions** for all L future packets: Y_n , … , Y_{n+L-1}

Historical data reveals trends or patterns beyond current state ($\mathbf{x}_n$).

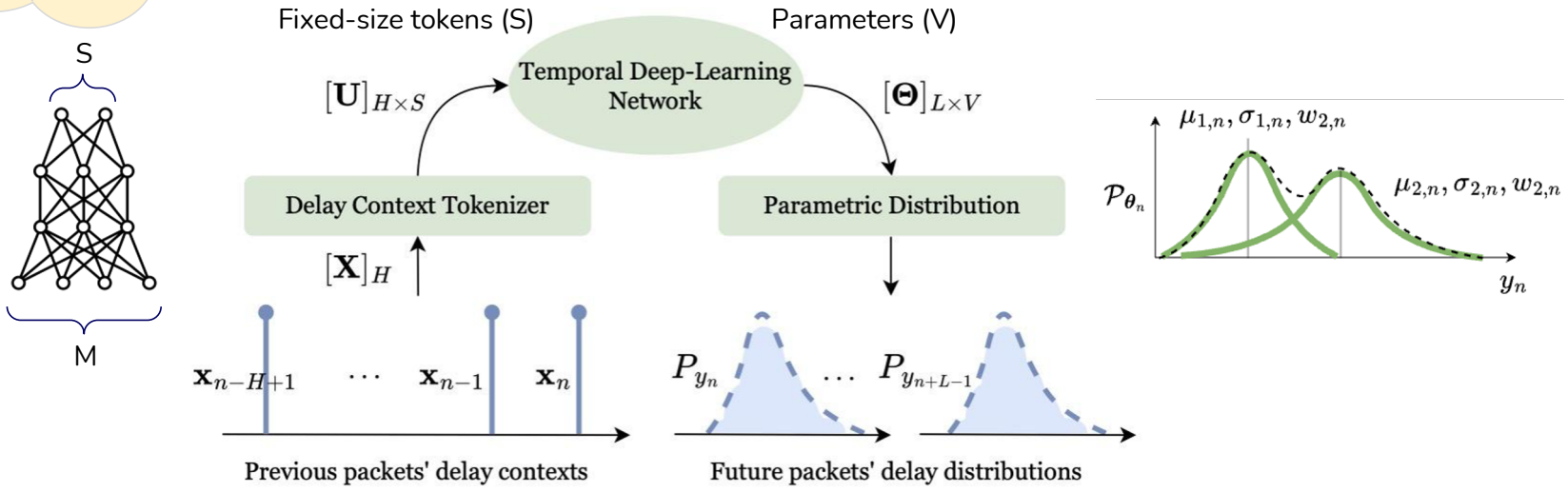
Capture the relationships between **previous observations** and future delays.

Prediction Approach

(1) Tokenization
Efficient temporal learning
Transfer learning

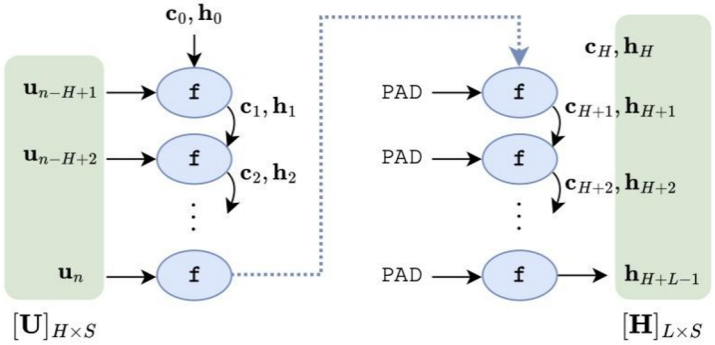
(2) LSTM or Transformer

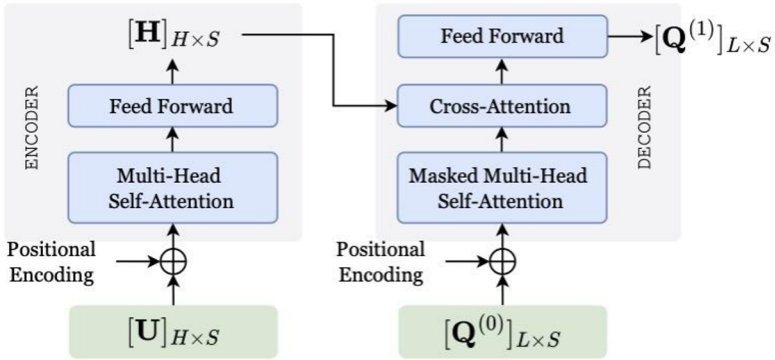
(3) Parametric Delay Density



Mostafavi et al "Probabilistic Delay Forecasting for in 5G using Recurrent and Attention-based Architectures", ArXiv 2025.

Temporal Deep Learning Methods





Recurrent Model

RNN or LSTM
Simple, but fails at longer dependencies

Attention-Based Model

Transformer
Can learn arbitrary far dependencies

Dataset Creation and Training

A traffic generator runs to capture packet-level context vectors and delays.

Each sample is a packet record containing

- A “history window” (H past packets)
- The true delay values for the following L packets.

$$\mathcal{D} = \{(\mathbf{X}_m, y_m, \ldots, y_{m+L-1})\}_{m=1}^{N}$$

The model optimizes a negative log-likelihood (NLL) loss from

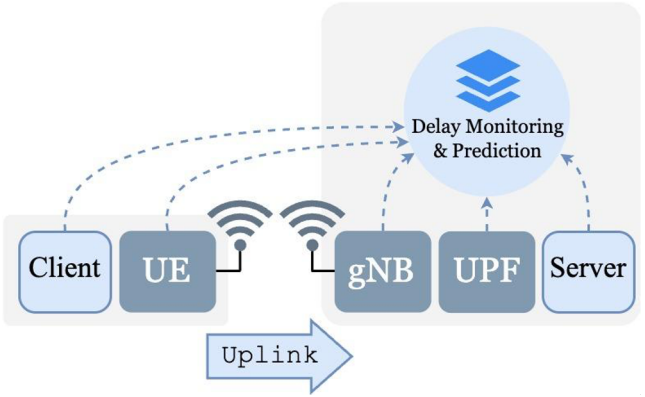
- Predicted delay distributions and the true delays

$$\mathcal{L}(\mathcal{D}) = -\sum_{m=1}^{N}\sum_{l=0}^{L-1} \ln\left(\mathcal{P}_{\boldsymbol{\theta}_{m+l}}(y_{m+l})\right)$$

Evaluation

Data collection:

- Openairinterface 5G on ExPECA
- EDAF 2.0, collecting per packet in addition to the delay:
  - Packet size (continuous)
  - Inter-arrival time (continuous)
  - Packet arrival slot (discrete)
  - MCS index (discrete)
  - HARQ retransmission count (discrete)
  - RLC retransmission count (discrete)



Experiments:

A) Single packet inter-arrival time: 50ms, and size: 200B
B) Packet size 100B with multiple inter-arrival times:
   - 10ms, 20ms, 50ms, 100ms

Training and inference:

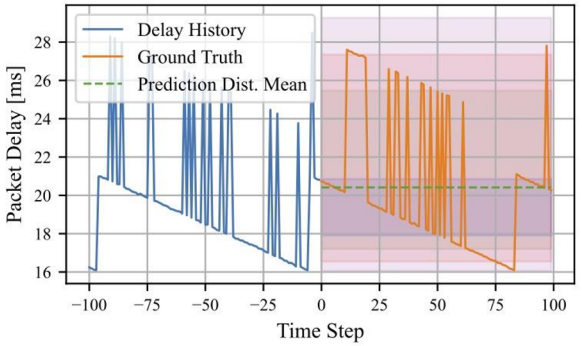
- Pytorch implementation (integrates with EDAF)
- Dell Server with Nvidia L4 GPU

Benchmarks

Single-Step Models (SOTA)

- **MLP** (37k param.)
  Fully connected feed-forward network
  Outputs a *single* delay prediction (of all future packets)

- **LSTM-SS** (33k param.)
  Incorporates past sequence information using a recurrent (LSTM) encoder
  Outputs a *single* delay prediction similar to MLP

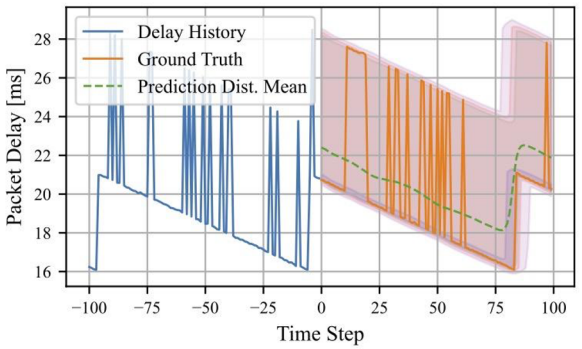


MLP 5k, H:1, L:100

Multi-Step Models (Our proposed approaches)

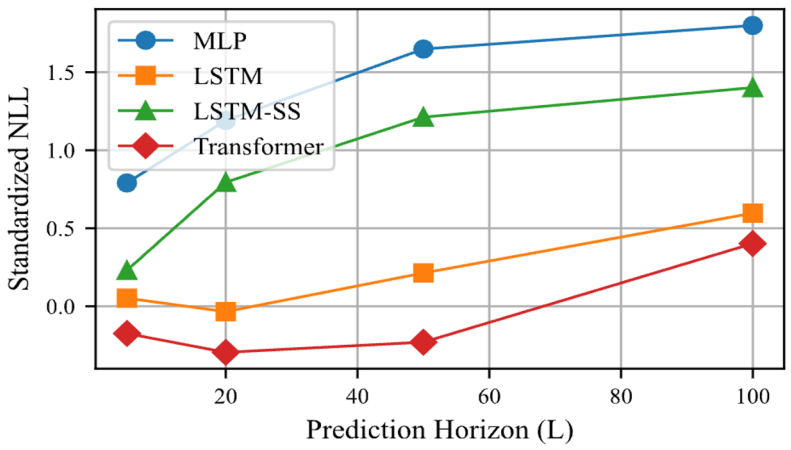
- **LSTM** (33k param.)
  Recurrent architecture *autoregresses* over future packets
  Generates delay distribution predictions for *multiple upcoming packets*

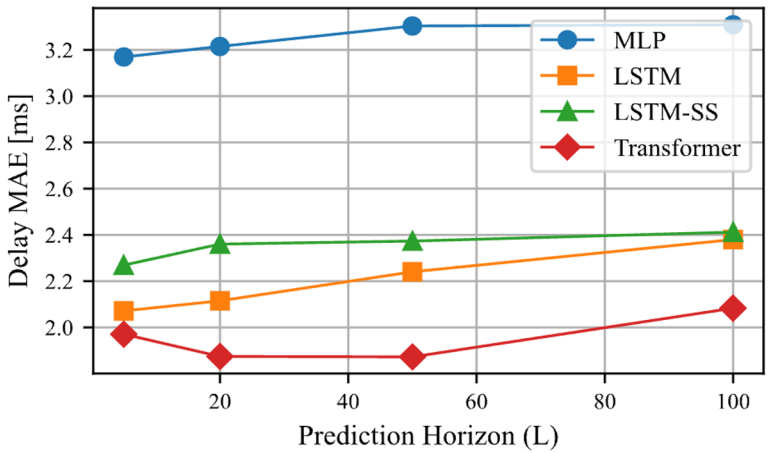
- **Transformer** (78k param.)
  Uses self-attention to process the history window
  Employs an encoder-decoder structure for *multi-step* forecasting



Transformer 5k, H:100, L:100

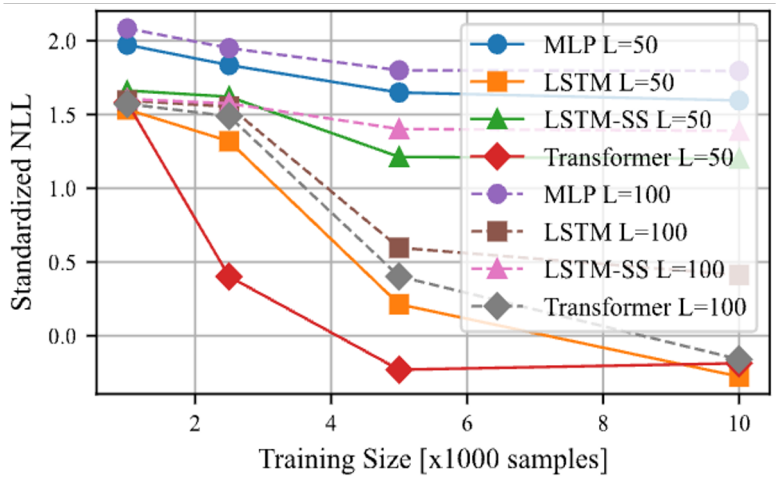
Models Accuracy Analysis



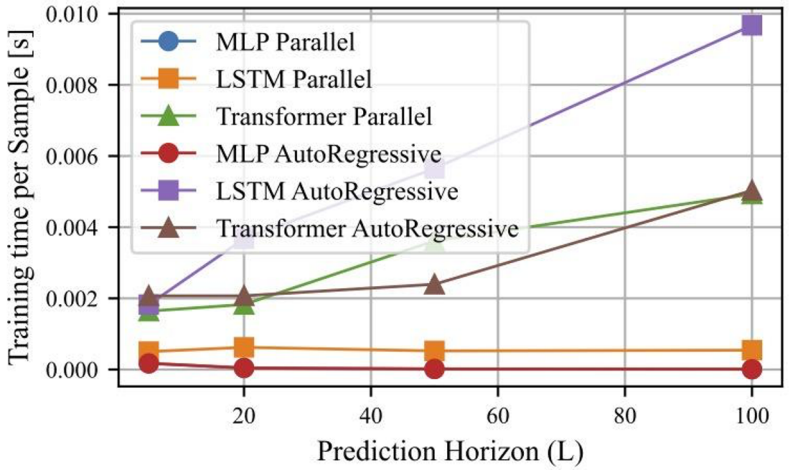


Comparison of model accuracy across different prediction horizons. All models were trained on 5k samples exp B

Models Complexity Analysis



Comparison of model accuracy across training dataset size, exp B



Training times for H:100, L:100 and different models

Outline

- Introduction + Motivation
- Predictability of Communication Systems
- From Theory to Practical Latency Prediction
- **Conclusions & Future Work**

Conclusions

- Cellular network infrastructures today are not the backbone for CPS deployment
- 6G aims at enabling the cyber-physical continuum
- One direction in enabling it:
  - Make latency evolution more transparent along the entire ‘loop’
  - Identify when degradations occur, harmonize with application
  - Performance prediction becomes key element!

- Decompose into predictability and practical prediction
  - Theoretical formulation of predictability and insights
  - Architectures for probabilistic latency prediction