URLLC: System Design Perspectives through Queuing Analysis

TUM URLLC Workshop, Zugspitze, July 2018

joint work with S. Schiessl, H. Al-Zubaidy

James Gross

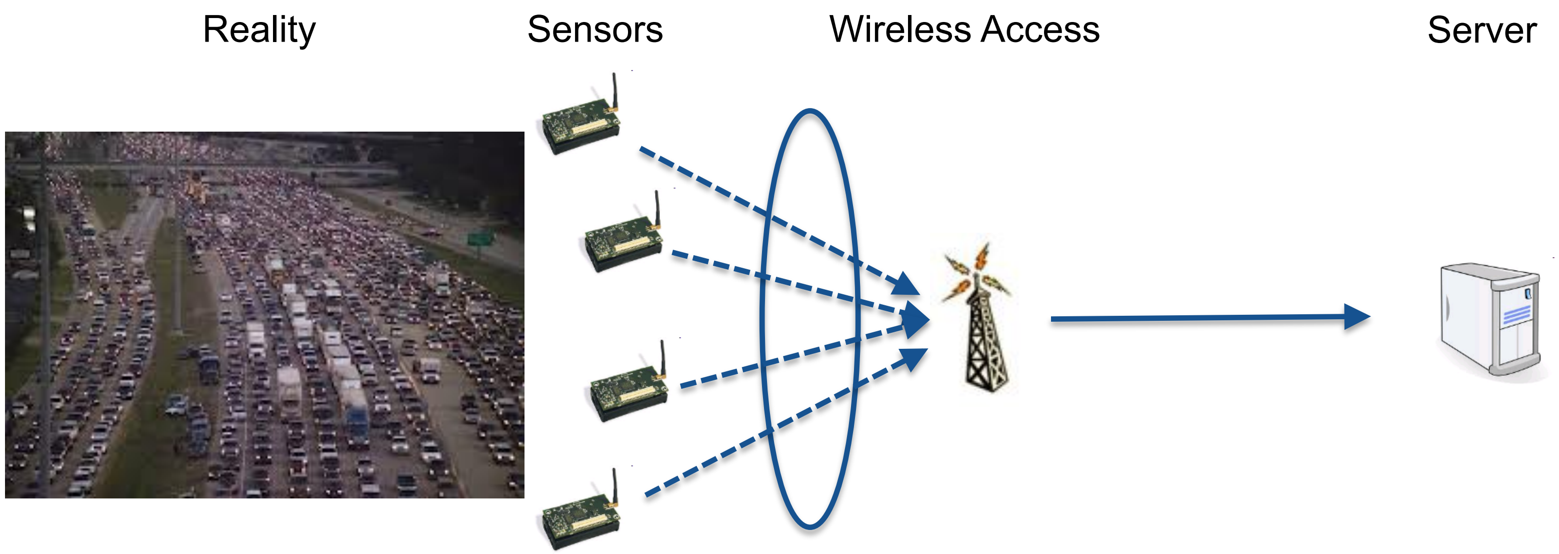
- Associate professor at KTH Stockholm (since 2012)
- Assistant professor at RWTH Aachen University (2008-12)
- PhD from TU Berlin in 2006
- Co-Founder of R3 Communications GmbH/Berlin

- Research focus:
  - Cellular networks
  - Critical machine-type communications
  - Theoretical network performance models
  - Edge computing and artificial intelligence

Outline

- URLLC: Motivation and Requirements
- Queuing Analysis Approaches
- Achieved Results:
  - Interference Channel
  - FBL and CSI Accuracy
  - MISO Downlink
- Discussion and Outlook

Machine-Type Communications: Origins

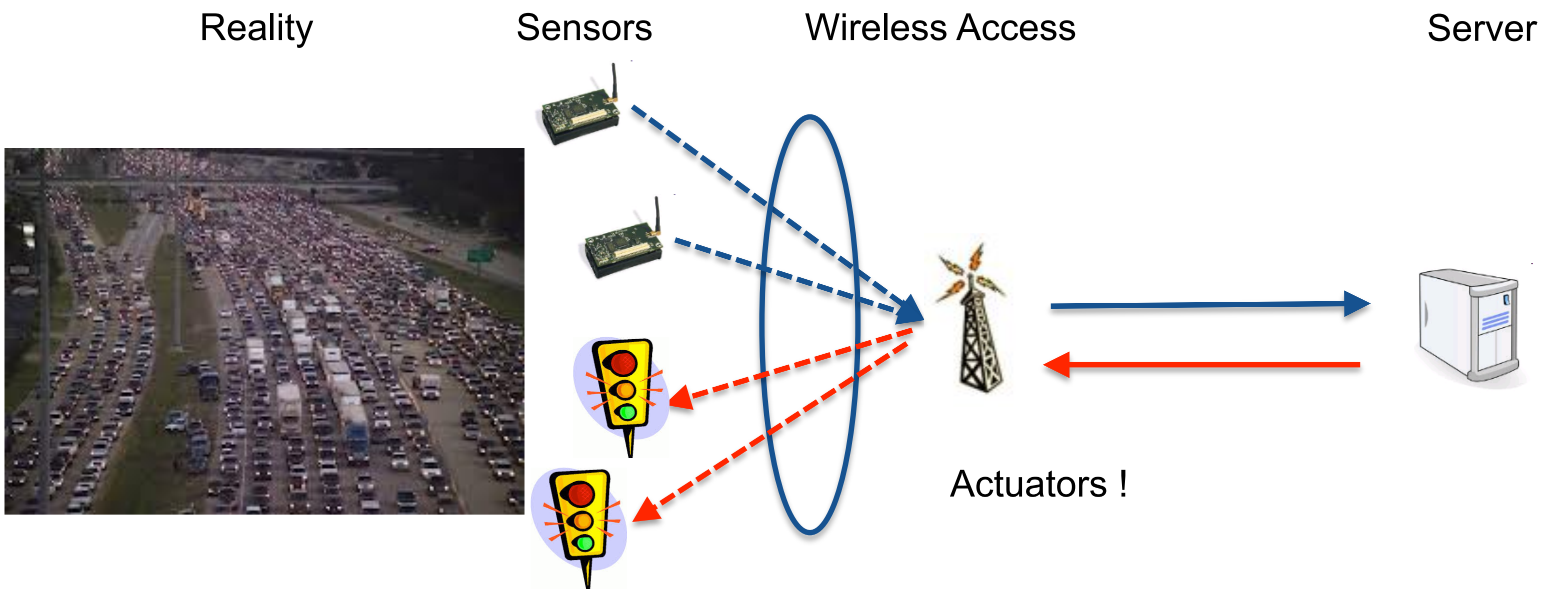


Autonomous monitoring & metering purpose

- End of 90s: First research on “sensor networks”
- Mid 2000: First standards (802.15.4, 6LowPAN)
- ~2010: Picked up by cellular networking industry (M2M business)

➔ Massive machine-type communications

Closing the Loop …



- Closed-loop control (driven by autonomy trend)
- Dependability becomes the focus

  ➔ Critical machine-type communications!

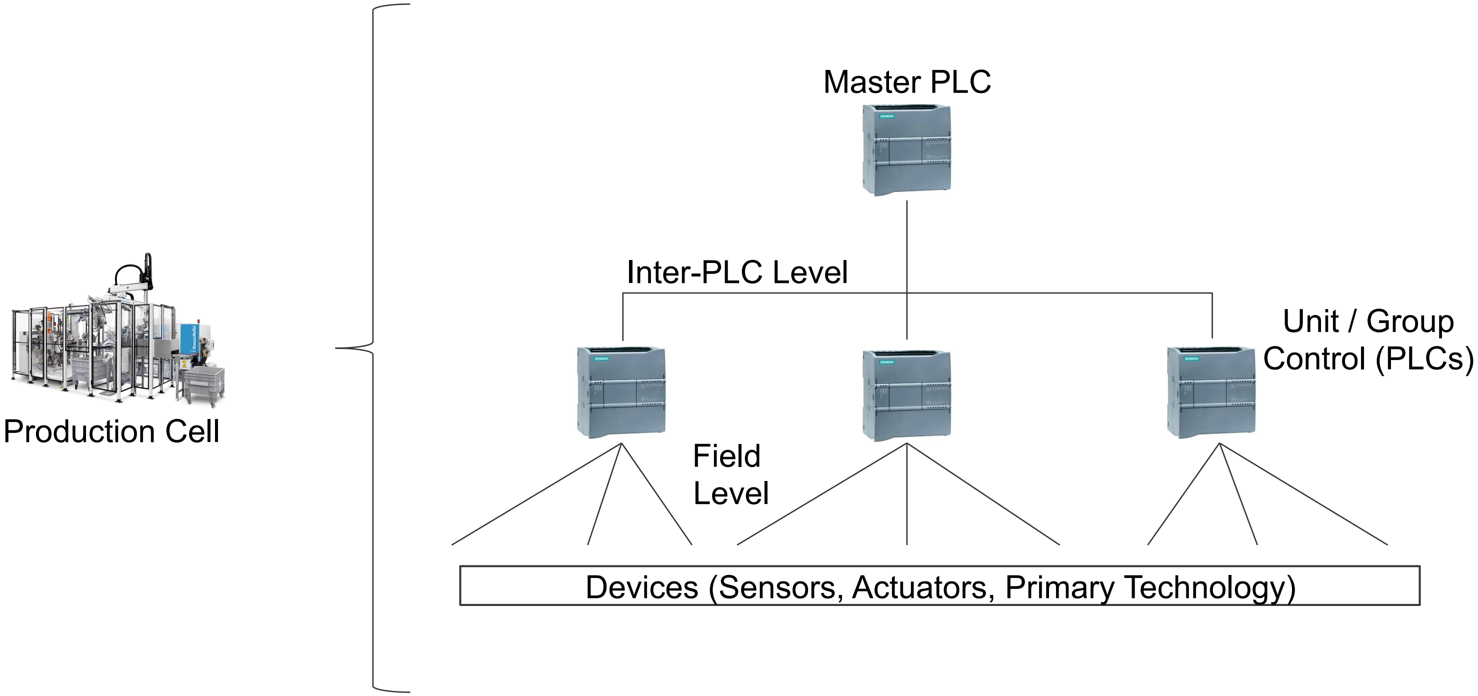
Critical MTC: Application Fields

- Various application fields according to 3GPP [1]:
  - 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

Critical MTC: Factory Automation

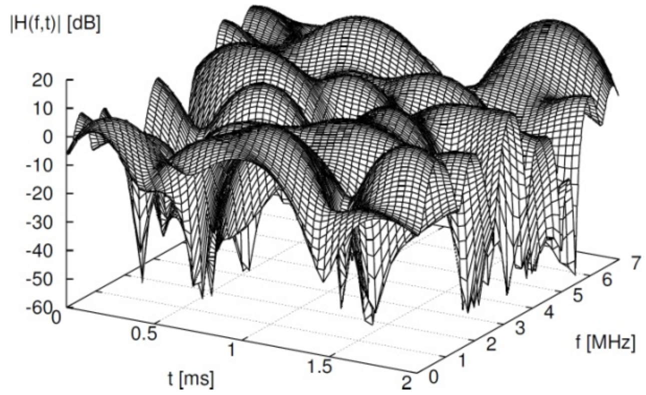


Range of Factory Automation Requirements

- Dependability: Availability + Reliability + Security

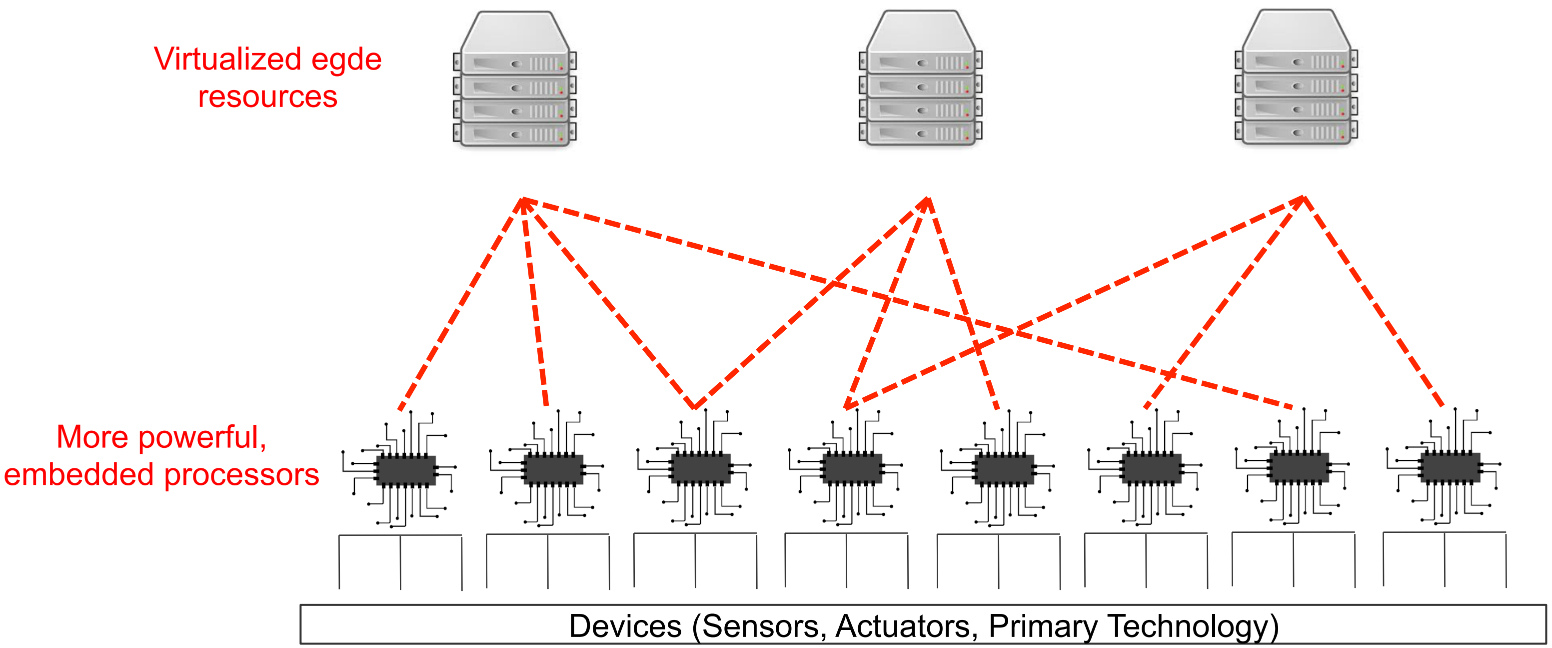
- Field-Level Control
  - Cycle time: <10 ms
  - Packet sizes: < 10 byte
  - Reliability: > $1 - 10^{-6}$

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



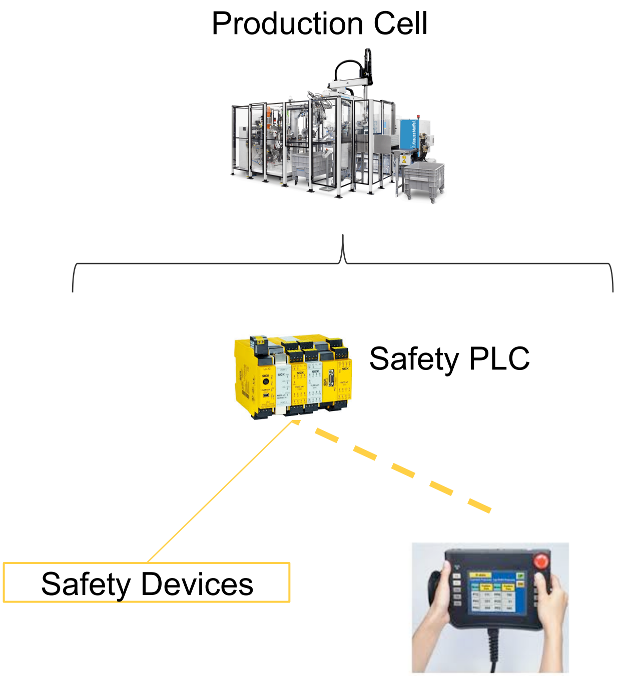
Why turn to wireless?

Visionary Reasoning: Flexibility

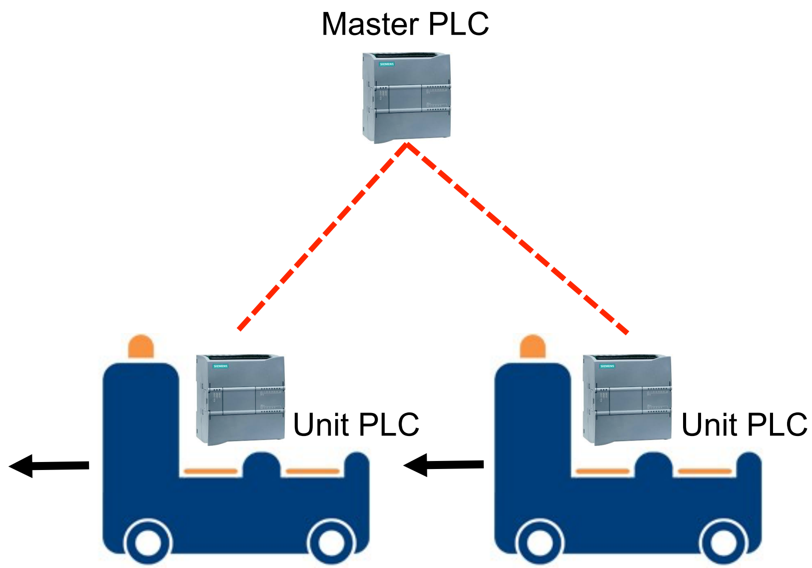


Realistic Use Cases: Mobility-Driven

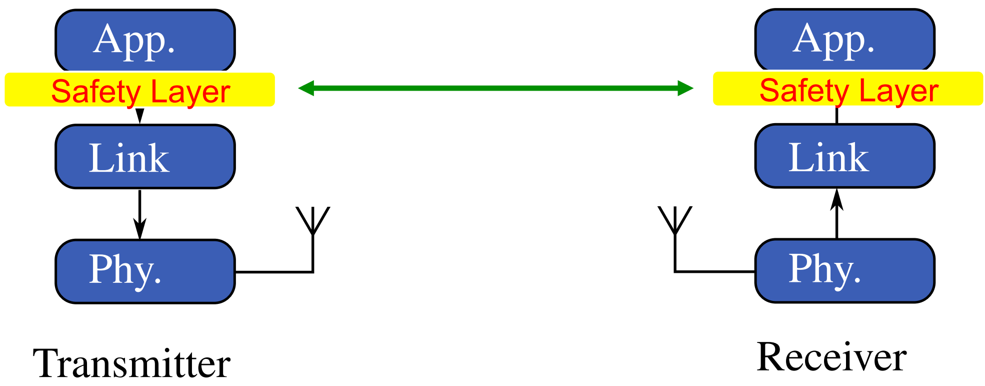
Safety Cases



Logistics Cases

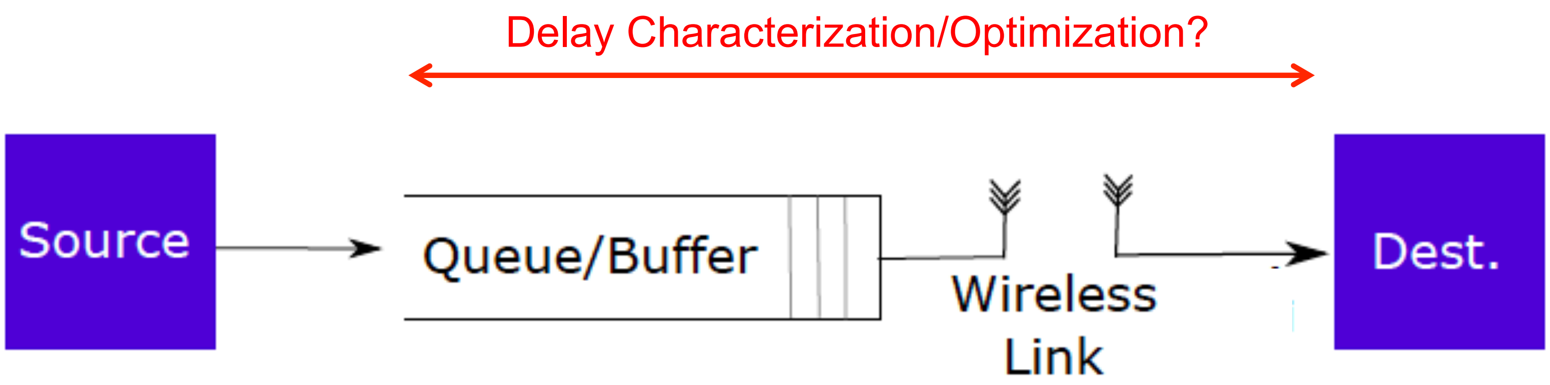


Systems & Safety Layers



- Black channel principle
- Periodic message exchange, >10 ms cycle time
- Small PDUs, about 10 byte
- **Turns link reliability issues into availability issues of the system**

Queuing-Theoretic Problem Formulation



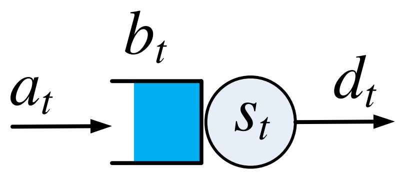
- Deterministic arrivals
- Random service: Fading, interference, cross-traffic

Outline

- URLLC: Motivation and Requirements
- **Queuing Analysis Approaches**
- Achieved Results:
  - Interference Channel
  - FBL and CSI Accuracy
  - MISO Downlink
- Discussion and Outlook

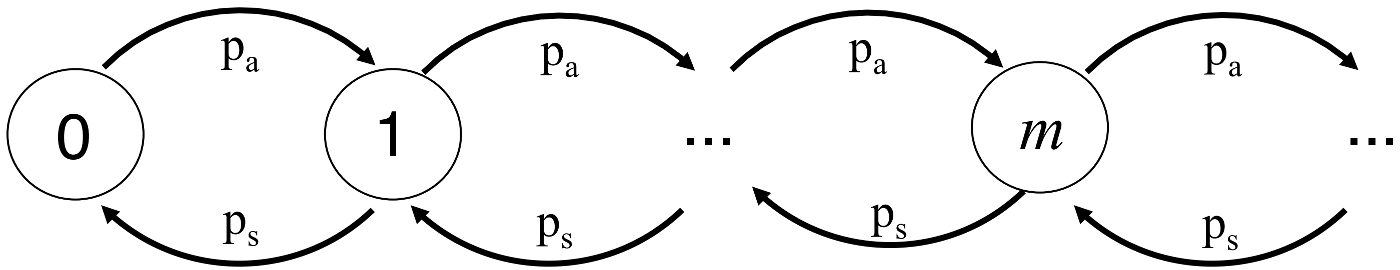
Modeling Assumptions

- Discrete time t
- Queue has infinite size
- Work-conserving server
- FIFO service order
- a_t, s_t, d_t : Arrival, service and departure of slot t
- Arrival & service process are independent and stationary
- b_t : Backlog at slot t



Traditional Approach: DTMCs

- Per slot system size grows/decreases by 1, or stays the same
- Markov property of arrival and service process: With probability p_s system size decreases by 1 regardless of previous evolution (p_a : increases by 1)
  ➔ Homogeneous discrete-time birth-death Markov chain, steady state exists under certain conditions (stability criteria)



Steady-state analysis: $\vec{\pi} = \vec{\pi} \cdot \mathbf{P} \quad \& \quad \sum_{\forall i} \pi^i = 1$

Traditional Approach: Pros & Cons

- Difference equation approach (balance equations)
- **Pros:**
  - 100 years of research: Lots of results, well understood
  - Typically provides exact results
- **Cons:**
  - Simplicity hinges on Markov property / single packet event
  - Quickly becomes intractable (concatenated systems, cross-traffic, scheduling)

Cumulative System View

Define the following cumulative processes:

$$A_{s,t} = \sum_{i=s}^{t} a_i, \quad S_{s,t} = \sum_{i=s}^{t} s_i, \quad D_{s,t} = \sum_{i=s}^{t} d_i$$

Let us assume that new arrivals can be served instantly.

Denote the backlog at time t as b_t, we have (Lindley) :

$$b_t = \max(0, b_{t-1} + a_t - s_t)$$

As the system is lossless, we also have:

$$b_t = A_{0,t} - D_{0,t}$$

Exercise: From Lindley to Reich!

Work through the recursion of Lindley's equation (use $b_0 = 0$)

$$
\begin{aligned}
b_t &= \max\left(0, b_{t-1} + a_t - s_t\right) \\
&= \max\left(0, \max\left(0, b_{t-2} + a_{t-1} - s_{t-1}\right) + a_t - s_t\right) \\
&= \max\left(0, \max\left(a_t - s_t, b_{t-2} + a_t + a_{t-1} - s_t - s_{t-1}\right)\right) \\
&= \max\left(0, A_{t,t} - S_{t,t}, b_{t-2} + A_{t-1,t} - S_{t-1,t}\right) \\
&= \max_{0 \le i \le t}\left(0, A_{i,t} - S_{i,t}\right) \\
&= \max_{0 \le i \le t}\left(A_{i,t} - S_{i,t}\right)^+
\end{aligned}
$$

Min,+ System Theory of Queuing Systems

What does Reich's equation mean for the system output?

$$\begin{aligned} b_t &= A_{0,t} - D_{0,t} \Leftrightarrow \\ D_{0,t} &= A_{0,t} - b_t \\ &= A_{0,t} - \max_{0\leq i\leq t}\left(A_{i,t} - S_{i,t}\right)^+ \\ &= \min_{0\leq i\leq t}\left(A_{0,t} - A_{i,t} + S_{i,t}\right) \\ &= \min_{0\leq i\leq t}\left(A_{0,i-1} + S_{i,t}\right) \\ &= \left(A \oplus S\right)_{0,t} \end{aligned}$$

Turns out that:

$$\begin{aligned} b_t &= A_{0,t} - D_{0,t} \\ &= \max_{0\leq i\leq t}\left(A_{i,t} - S_{i,t}\right)^+ \\ &= \left(A \ominus S\right)_{t,t} \end{aligned}$$

with:

$$\left(X \ominus Y\right)_{s,t} = \max_{\tau\leq s}\left(X_{\tau,t} - Y_{\tau,s}\right)$$

Probabilistic Backlog Bound

First consider:

$$\begin{aligned}\mathbb{P}\left((X \ominus Y)_{s,t} \geq z\right) &= \mathbb{P}\left(\max_{\tau \leq s}\left(X_{\tau,t} - Y_{\tau,s}\right) \geq z\right) \\ &\leq \sum_{\tau=0}^{s} \mathbb{P}\left(X_{\tau,t} - Y_{\tau,s} \geq z\right) \\ &\leq e^{-\theta z} \cdot \sum_{\tau=0}^{s} \mathbb{M}_X(\theta, \tau, t) \cdot \mathbb{M}_Y(-\theta, \tau, s) \\ &= \epsilon\end{aligned}$$

Union Bound

Chernoff Bound

Thus:

$$\mathbb{P}\left((A \ominus S)_{t,t} \geq \max_{0 \leq \theta}\left(\frac{1}{\theta}\left(\log \sum_{\tau=0}^{t} \mathbb{M}_A(\theta, \tau, t) \cdot \mathbb{M}_S(-\theta, \tau, t) - \log \epsilon\right)\right)\right) \leq \epsilon$$

Stochastic Network Calculus: Pros & Cons

- Moment-bounds on system variables

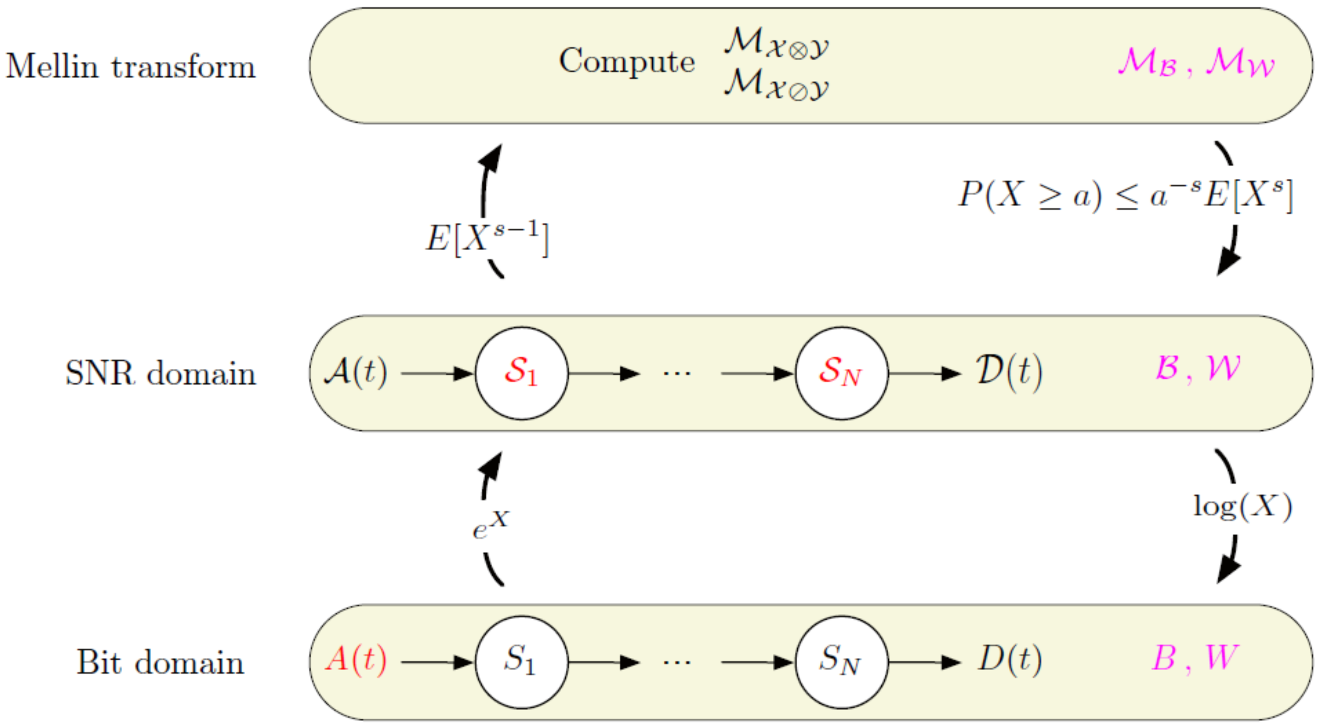
- **Pros:**
  - Applicable for arbitrary arrival and service processes
  - Strict upper bound on system performance
  - Works also for concatenated systems

- **Cons:**
  - Best for stationary processes with independent increments
  - Upper bound is not tight in general

Outline

- URLLC: Motivation and Requirements
- Queuing Analysis Approaches
- **Achieved Results:**
  - **Interference Channel**
  - FBL and CSI Accuracy
  - MISO Downlink
- Discussion and Outlook

From Bit-Domain SNC to SNR-Domain SNC



H. Al-Zubaidy et al. “Network-layer Performance Analysis of Multi-hop Fading Channels,” Transactions on Networking, 24/1, 2016

SISO Interference Channel

Signal-of-interest and interference signals are fading.

$$\gamma_t = \frac{P_0|h_{0,t}|^2}{\sum_i P_i|h_{i,t}|^2 + \sigma^2}$$

Service in time slot t in bits:

$$S_t = n\log_2(1+\gamma_t)$$

w.l.o.g., assume n/log(2)=1.

➔ Service in the **SNR-domain:**

$$\mathcal{S}_t = e^{S_t} = 1 + \gamma_t$$

SISO Interference Channel

For the queueing analysis, we must find

$$\mathcal{M}_{\mathcal{S}}(\theta) = \mathbb{E}\left[\mathcal{S}^{\theta-1}\right] = \int_0^\infty (1+\gamma)^{\theta-1} f(\gamma) d\gamma$$

For K interferers, we get K integrals of the form

$$\int_0^\infty \frac{(1+\gamma)^{\theta-2}}{\gamma+a} e^{-\gamma} d\gamma = \int_1^\infty \frac{z^{\theta-2}}{z+a-1} e^{-z+1} dz$$

S. Schiessl et al. “On the Delay Performance of Interference Channels,” *IFIP Networking*, 2016.

SISO Interference Channel

Solution:

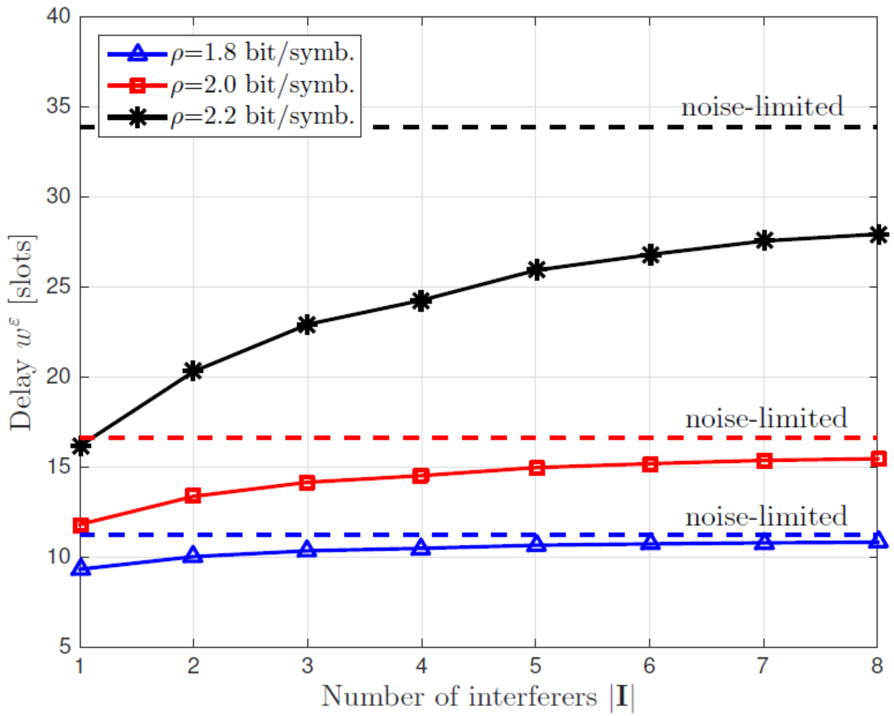
- Split the integral into two parts: $z < a$-1 and $z > a$-1
- For the second part with $z > a$-1:

$$\frac{z^{\theta-3}}{1+\frac{a-1}{z}} = z^{\theta-3}\sum_{n=0}^{\infty}\left(\frac{1-a}{z}\right)^{n}$$

- For the first part: similar solution
- ➔ Can determine $\mathcal{M}_S(\theta)$ in closed form (as a series of incomplete gamma functions)

S. Schiessl et al. "On the Delay Performance of Interference Channels," *IFIP Networking*, 2016.
F. Naghibi et al. "Performance of Wiretap Rayleigh Fading Channels under Statistical Delay Constraints," *IEE ICC*, 2017

SISO Interference Channel: Main Result



- Main av. signal power: 15 dB
- Total av. interference power is constant (8 dB).
- What is the max. delay w such that $p_v(w) < 10^{-6}$?

Result:

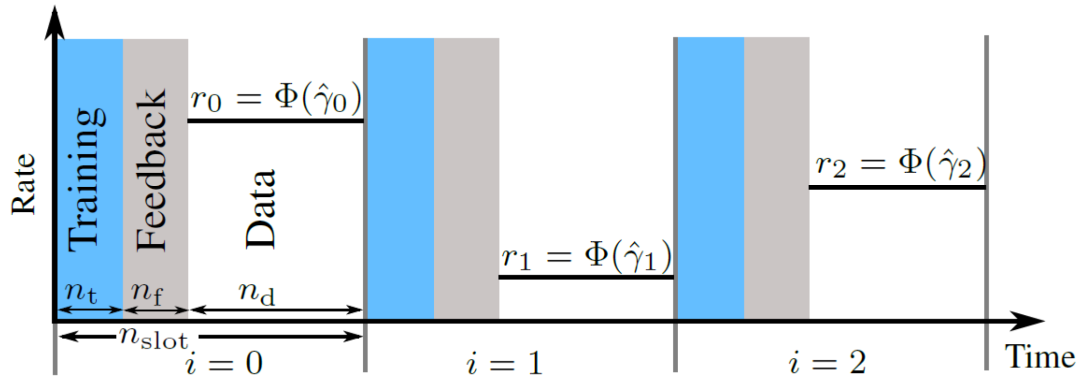
It is better to have one interf. with av. P=8 dB than two interf. with av. P=5 dB each.

Reason: signal from the one interferer is often weak, allowing high data rates

Outline

- URLLC: Motivation and Requirements
- Queuing Analysis Approaches
- **Achieved Results:**
  - Interference Channel
  - **FBL and CSI Accuracy**
  - MISO Downlink
- Discussion and Outlook

Finite Blocklength and Imperfect CSIT



- SISO set-up, focus on impact of CSI at transmitter:
  - Trade-off 1: Training symbols n_t ⇔ Data symbols n_d
  - Trade-off 2: Rate r ⇔ Error probability ε

➔ Errors are bad, but low r and small n_d can also increase the queueing delay!

Finite Blocklength and Imperfect CSIT

Normal approximation (Polyanskiy et al. / Yang et al.):

$$\varepsilon \approx \mathbb{E}\left[Q\left(\frac{\log_2(1+\Gamma) - r}{\sqrt{\mathcal{V}(\Gamma)/n_{\mathrm{d}}}} \right) \middle| \hat{\gamma} \right]$$

Γ : Actual SNR (unknown/random)

$\hat{\gamma}$: Estimated SNR

Too complex for queueing analysis.

Thus, we find a normal approximation for Γ and use a Taylor approximation for the FBL effects, giving:

$$\varepsilon \approx Q\left(\frac{\hat{\gamma} - (2^r - 1)}{\sigma_{\mathrm{ICSI,FBL}}} \right)$$

S. Schiessl et al. “Delay Performance of Wireless Communications with Imperfect CSI and Finite Length Coding ,” *accepted for publication Transactions on Communications, 2018.*

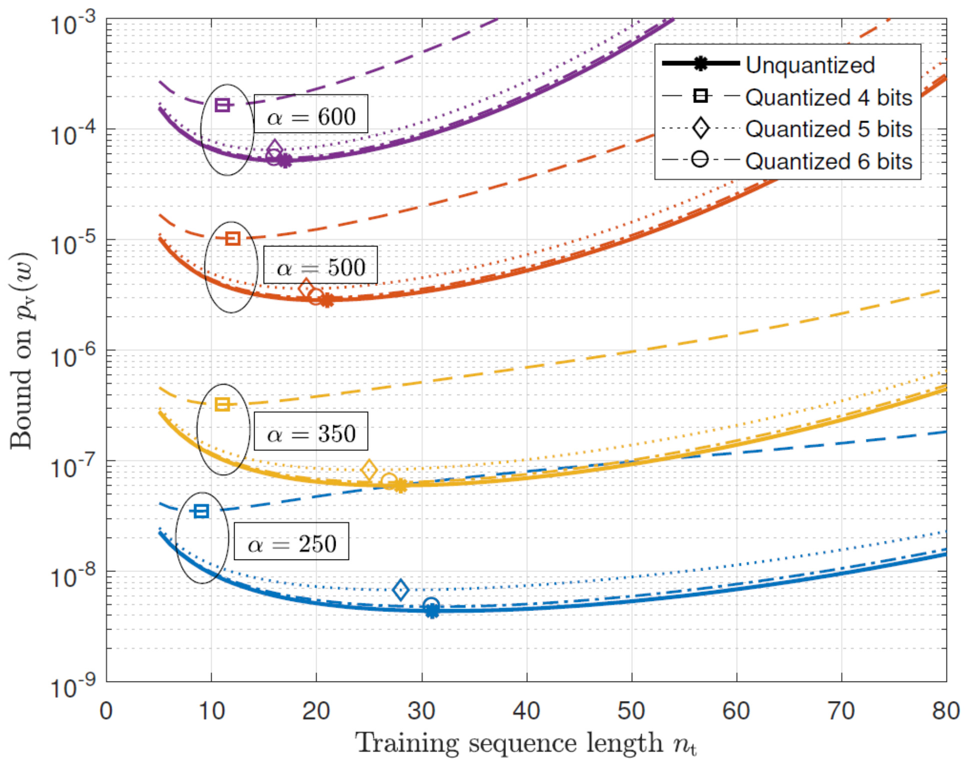
Finite Blocklength and Imperfect CSIT

To minimize the delay violation probability, minimize

$$\mathcal{M}_{\mathcal{S}}(\theta) = \mathbb{E}\left[\mathcal{S}^{\theta-1}\right] = \int_0^\infty (1+\gamma)^{\theta-1} f(\gamma) d\gamma$$

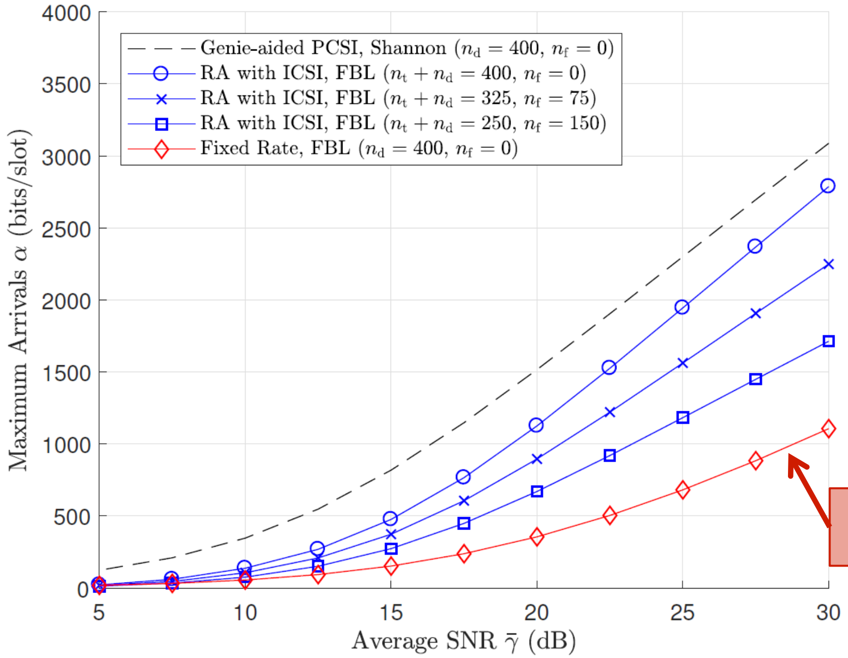
- For each estimated SNR $\hat{\gamma}$: need to solve trade-off $r \Leftrightarrow \varepsilon$
- Can be solved quickly, as the expression is convex in the approximate ε

Main Result 1: Optimal n_t



Parameters:
- n_{slot} = 250,
- n_d = n_{slot} - n_t,
- w = 5 slots,
- Avg. SNR 15 dB

Result 2: Rate Adaptation is Superior



Adaptive rate, n_d=370 – 390
$n_{feedback}$=0

Adaptive rate, n_d=220 - 240,
$n_{feedback}$=150

Fixed rate, n_d=400

- n_{slot} = 400,
- n_d = n_{slot} - n_t,
- w = 5 slots

- These results consider queueing constraints: $p_v(w=5) < 10^{-8}$
- Ignoring the queueing constraints would lead to wrong conclusions.

Outline

- URLLC: Motivation and Requirements
- Queuing Analysis Approaches
- **Achieved Results:**
  - Interference Channel
  - FBL and CSI Accuracy
  - **MISO Downlink**
- Discussion and Outlook

Multiuser MISO

Multiuser MISO with zero-forcing beamforming (ZFBF).

M antennas, K scheduled users

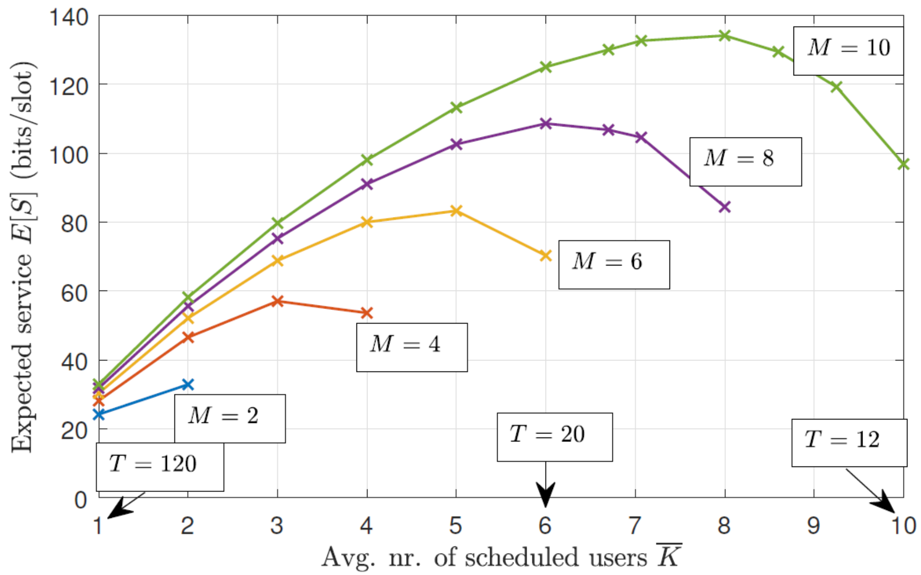
Large K: multiplexing gain ⟷ Small K: beamforming gain

What is the optimal K under delay constraints?

S. Schiessl et al. “On the Delay Performance of the Multi-user MISO Downlink,” *ArXiv preprint*, 2018.

Multiuser MISO

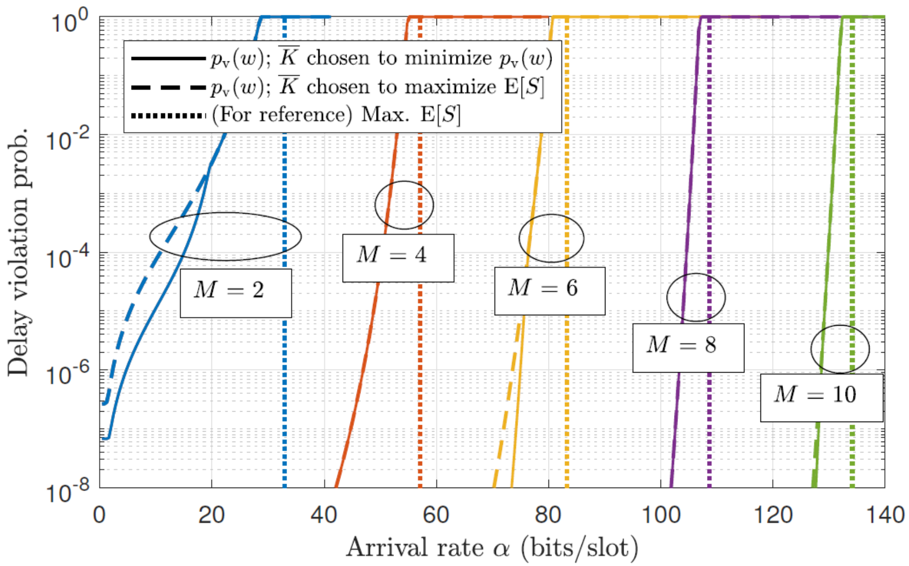
- Has been well studied with respect to ergodic sum rate, e.g., Hochwald & Vishwanath ’02.
- Choose K≈αM. Here: α ≈ 0.8



- n_{slot} = 400,
- K_{tot} = 120 users,
- P_{sum} = 20 dB

Multiuser MISO: Delay Performance

- Observation: For M ≥ 6, no queueing delay as long as expected arrival rate < 0.9 * expected service rate
- Optimal value K rarely changes under delay constraints



- n_{slot} = 400,
- K_{tot} = 120 users,
- P_{sum} = 20 dB,
- w = 120 slots

FYI: When K=2, each of the K_{tot} =120 users can be scheduled 2 times within w=120 slots.

Outline

- URLLC: Motivation and Requirements
- Queuing Analysis Approaches
- Achieved Results:
  - Interference Channel
  - FBL and CSI Accuracy
  - MISO Downlink
- **Discussion and Outlook**

Discussion

- Queuing analysis extends physical layer work towards real application layer performance
- SNC approaches can provide useful upper bounds
- Somewhat surprising findings for URLLC:
  - Have rather one strong interferer
  - Estimate channel & rate adaption
  - Relatively few antennas at transmitter lead (through channel hardening) already to almost perfect system performance

Outlook

- Transient system characterization instead of steady-state
- Analyze the entire loop through edge server
- Integrate models with control performance models