IoT connectivity, massive access, and URLLC

Petar Popovski
Connectivity Section
Department of Electronic Systems
petarp@es.aau.dk

AALBORG UNIVERSITY
DENMARK

SPS summer school on 6G @ Linköping, August 29, 2022

outline

6G and IoT connectivity

- massive access
  - model for correlated access
  - user identification in unsourced access
  - massive downlink ACK
- ultra-reliable low-latency communications
  - latency and general timing
  - how to guarantee reliability
  - URLLC in action

outline

6G and IoT connectivity

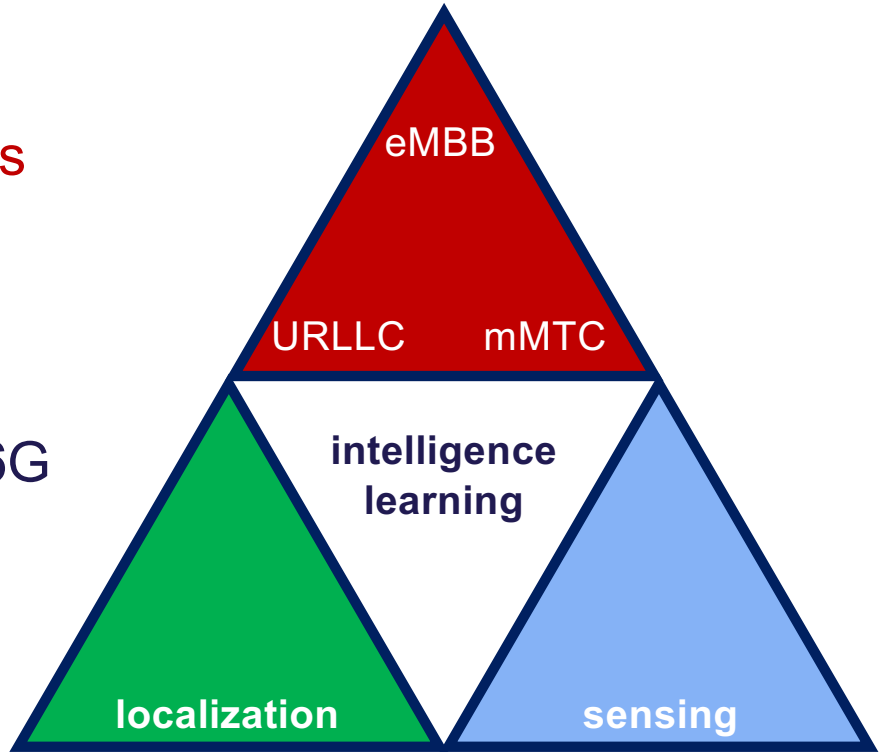
- massive access
- model for correlated access
- user identification in unsourced access
- massive downlink ACK
- ultra-reliable low-latency communications
- latency and general timing
- how to guarantee reliability
- URLLC in action

6G: more than communications

the 5G triangle was about communications

this triangle is about to be augmented in 6G

- potentially generalize the concept of slicing



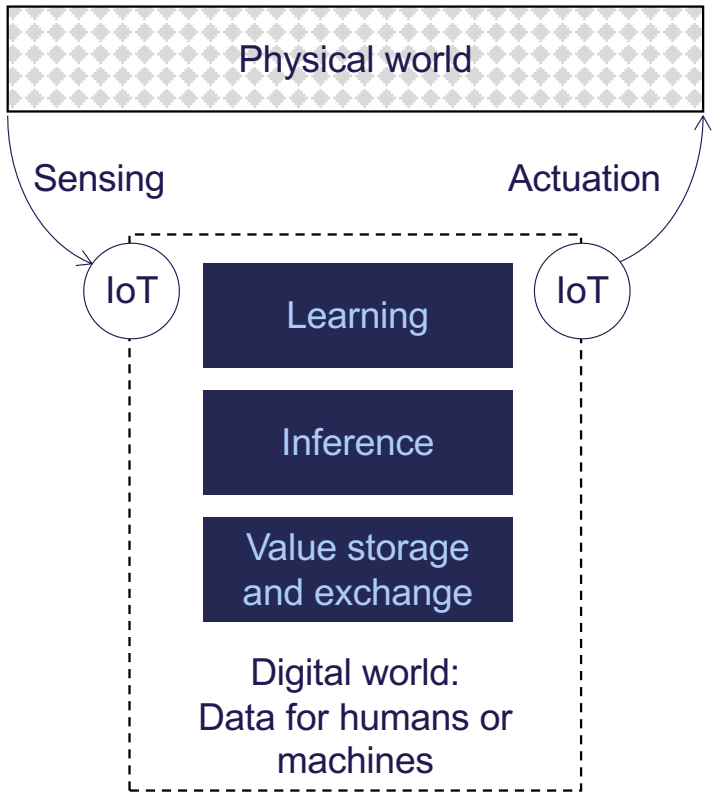
Internet of Things (IoT) in 6G

IoT as a micro-tunnel between the physical and digital world

- physical information → digital data
- data+algorithms → physical actions

data used in three principal ways

- learning and training of AI models
- inference and command actuation
- value storage and exchange



P. Popovski, F. Chiariotti, V. Croisfelt, A. E. Kalør, I. Leyva-Mayorga, L. Marchegiani, S. R. Pandey, B. Soret, "Internet of Things (IoT) Connectivity in 6G: An Interplay of Time, Space, Intelligence, and Value", available on ArXiv, https://arxiv.org/pdf/2111.05811.pdf, 2021.

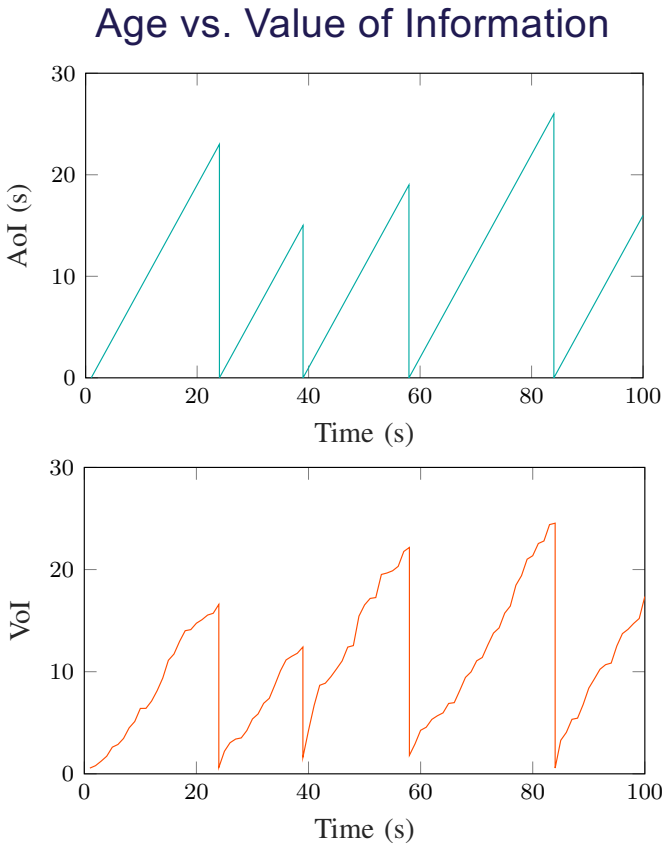
time in 6G

relativity of real-time

- Human-to-Human (H2H)
- Human-to-Machine (H2M) and human in the loop
- Machine-to-Machine (M2M)

generalized timing objective:
distributed decision or estimation within a given time frame

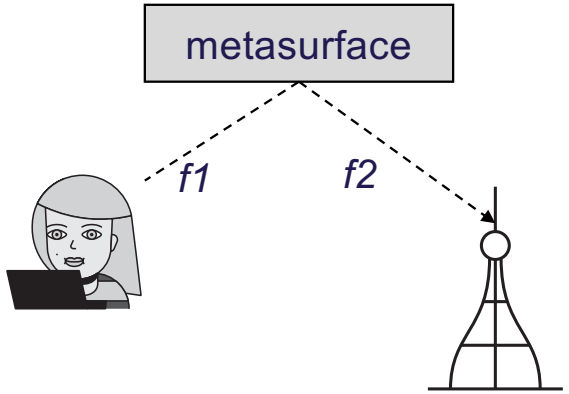
Value of information relates to semantics-oriented communication.



P. Popovski, F. Chiariotti, K. Huang, A. E. Kalør, M. Kountouris, N. Pappas, and B. Soret, "A Perspective on Time towards Wireless 6G," in Proceedings of IEEE, accepted, 2022.

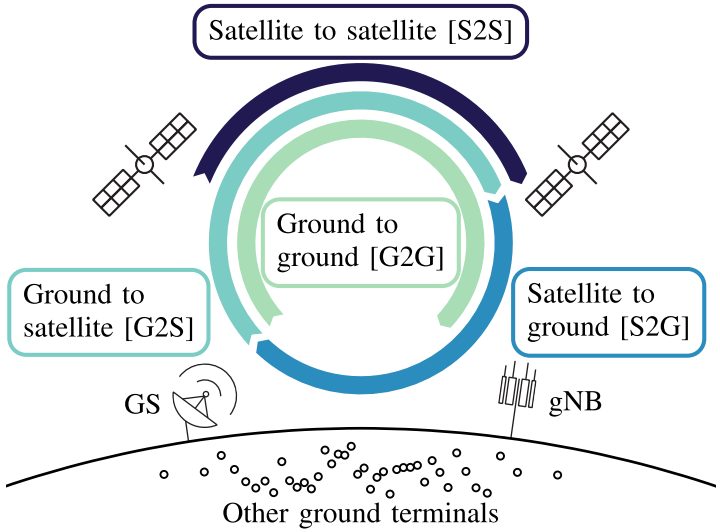
space in 6G

intelligent propagation space



- Reconfigurable Intelligent Surfaces
- Large Intelligent Surfaces
- what if the surface does a nonlinear transformation of signals?

NewSpace

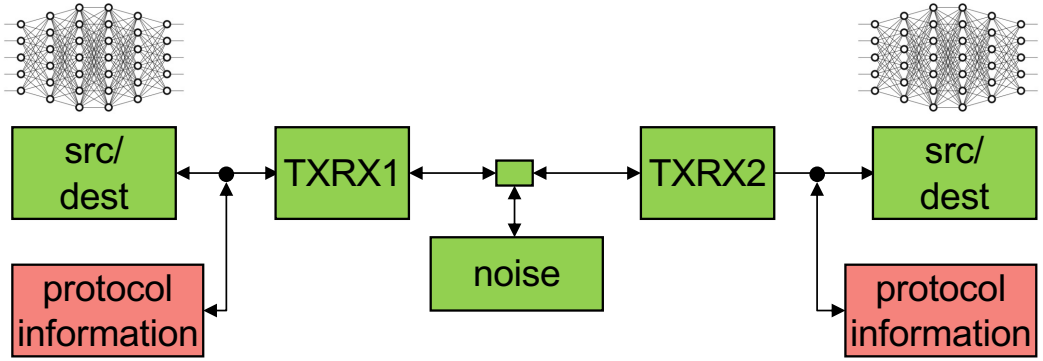


[*] E. Björnson, H. Wymeersch, B. Matthiesen, P. Popovski, L. Sanguinetti, and E. de Carvalho, "Reconfigurable Intelligent Surfaces: A Signal Processing Perspective With Wireless Applications," in IEEE Signal Processing Magazine, accepted, 2021.

[**] I. Leyva-Mayorga, B. Soret, M. Röpper, D. Wübben, A. Dekorsy, and P. Popovski, "LEO Small-Satellite Constellations for 5G and Beyond-5G Communications," in IEEE Access, vol. 8, pp. 184955-184964, 2020.

intelligence in 6G

how are the communication protocols affected
by the growing intelligence in the nodes?



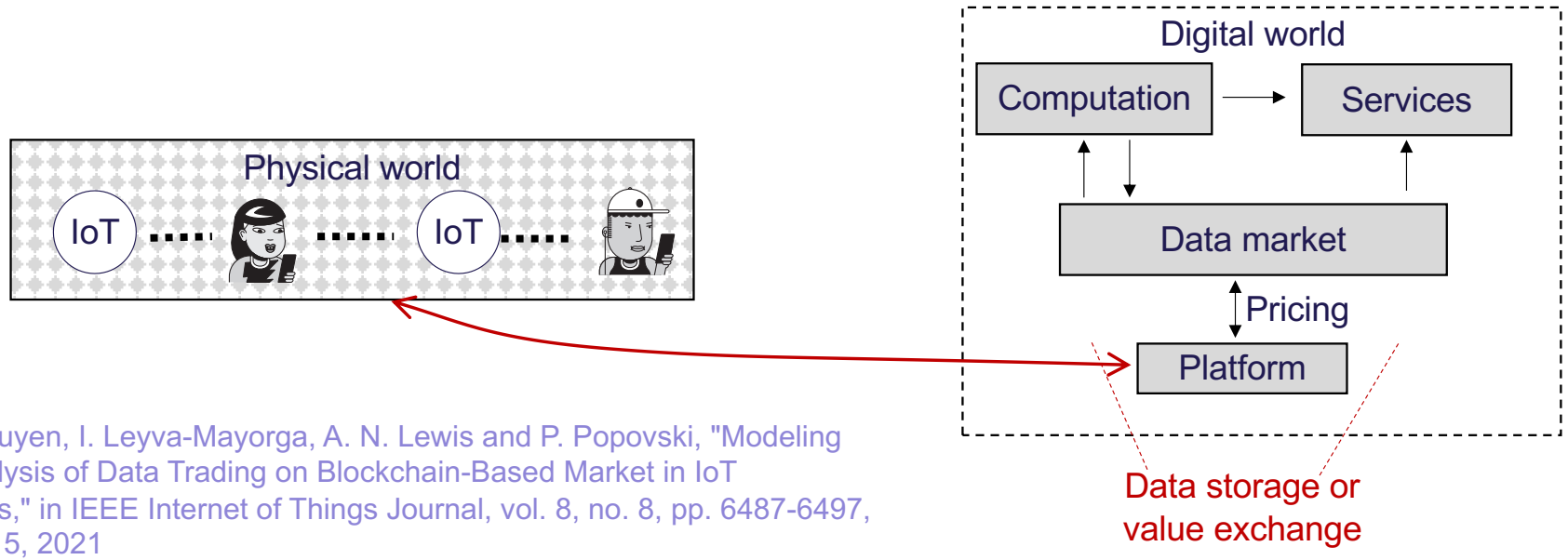
- distributed learning and distributed (edge) inference

[*] P. Popovski, O. Simeone, F. Boccardi, D. Gündüz, and O. Sahin, "Semantic-Effectiveness Filtering and Control for Post-5G Wireless Connectivity", Journal of the Indian Institute of Science, invited paper, 2020.

[**] Q. Lan, D. Wen, Z. Zhang, Q. Zeng, X. Chen, P. Popovski, and K. Huang, "What is Semantic Communication? A View on Conveying Meaning in the Era of Machine Intelligence", Journal of Communications and Information Networks (JCIN), invited paper, accepted, 2021.

value in 6G

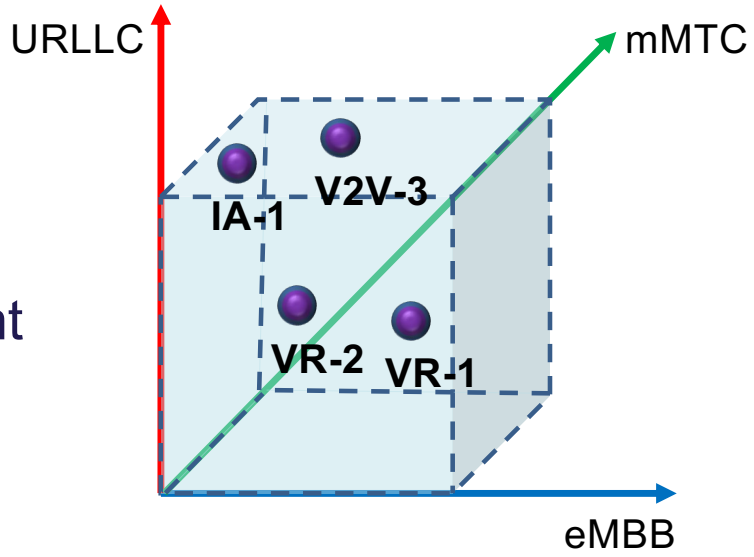
- enormous data amounts used in various inference and learning tasks
- privacy vs. economic value of data
- future IoT devices may become autonomous sellers and buyers of data



L. D. Nguyen, I. Leyva-Mayorga, A. N. Lewis and P. Popovski, "Modeling and Analysis of Data Trading on Blockchain-Based Market in IoT Networks," in IEEE Internet of Things Journal, vol. 8, no. 8, pp. 6487-6497, 15 April15, 2021

IoT in 5G

5G has a platform approach to connectivity:
define three generic services and
represent any concievable connectivity requirement
as their combination.



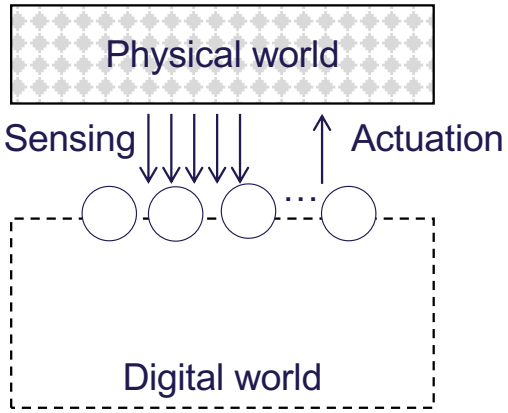
natural to ask:
what IoT connectivity requirements
may not be addressed by 5G in an efficient way?

two examples

massive MTC

- **original problem:**
  throughput maximization
  from an unknown random subset

- **problem with correlated sources:**
  collect sufficient data to carry out inference

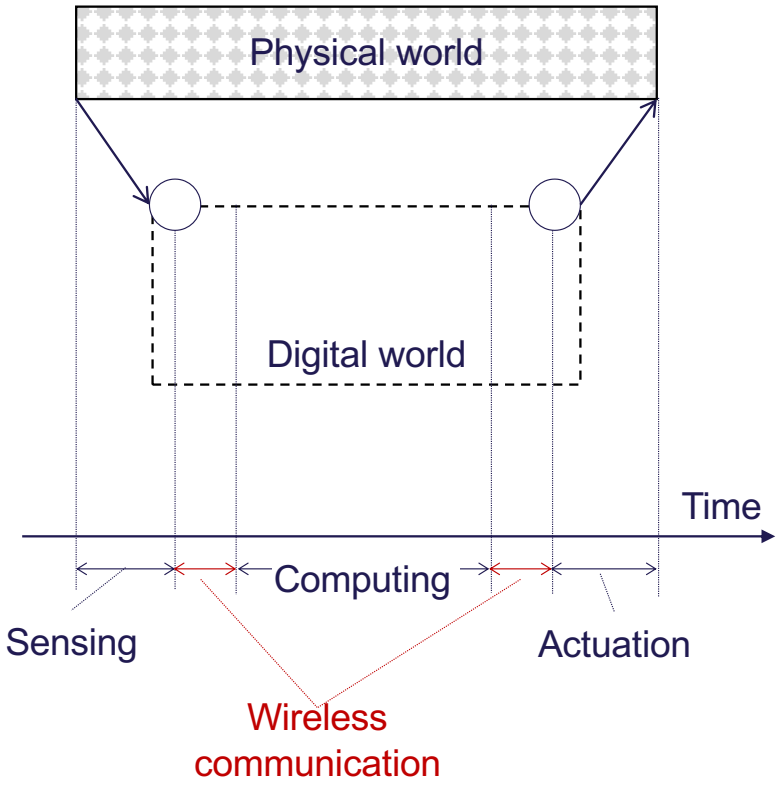


two examples

URLLC

- **original problem:**
  deliver the data of
  size X within Y milliseconds with reliability Z

- **modified problem:**
  work with a generalized
  timing requirement



outline

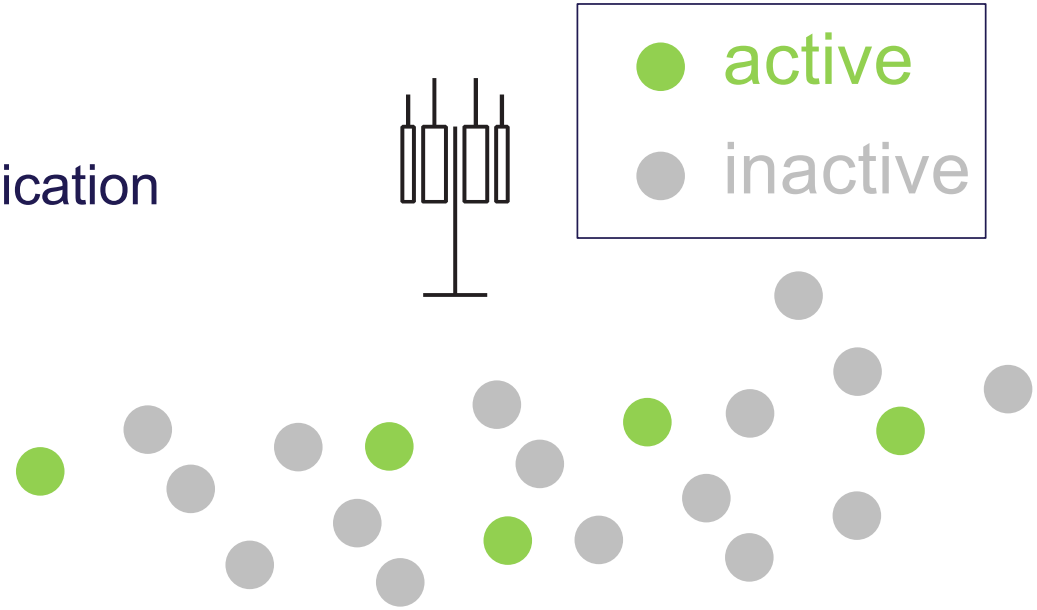
6G and IoT connectivity

- massive access
- model for correlated access
- user identification in unsourced access
- massive downlink ACK

- ultra-reliable low-latency communications
- latency and general timing
- how to guarantee reliability
- URLLC in action

the massive access problem

mMTC:
massive Machine Type Communication



the **set of nodes** that
want to send to the BS in the **uplink** is **unknown**

communication model

this is a slotted model

single BS received signal in the k –th slot:

$$y_k = \sum_{n=1}^{N} h_{kn} a_{kn} x_{kn} + z_k$$

N - total number of users;

h_{kn} - wireless channel coefficient during the k –th slot;

a_{kn} - the activity of n–th user during the k –th slot,
$a_{kn} = 1$ if the user is active and $a_{kn} = 0$ otherwise;

x_{kn} - packet/symbol sent by the n–th user during the k –th slot;

z_k - noise in the k –th slot.

sources of uncertainty

$$y_k = \sum_{n=1}^{N} h_{kn} a_{kn} x_{kn} + z_k$$

all of the following:

$$h_{kn}\,, a_{kn}, x_{kn}\,, z_k$$

the central role in mMTC is played by the uncertainty contained in the activity variable a_{kn}

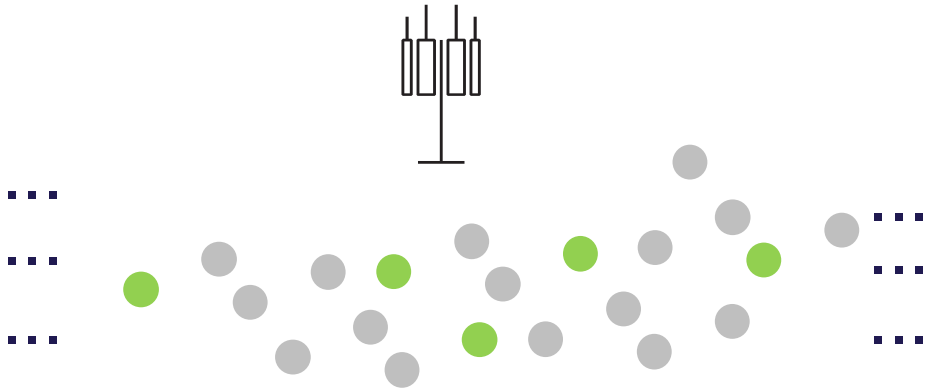
for modeling purpose we can make a_{kn}

- part of the channel $h'_{kn} = h_{kn} a_{kn}$, leading e.g. to Bernoulli-Gauss distribution for Rayleigh channel
- part of the transmitted symbol $x'_{kn} = x_{kn} a_{kn}$, leading to having a special ”empty” symbol.

random access: two classical assumptions

1. packet is an atomic unit of information

2. users are activated independently



packet as an atomic unit

convenient from the protocol viewpoint

used since the earliest days of ALOHA

infinite population assumption $N \to \infty$

allows asymptotic analysis

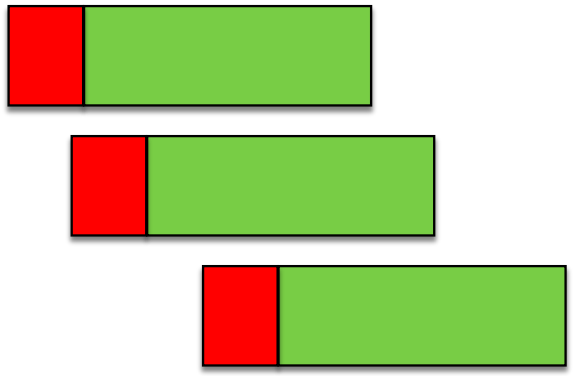
allows to model a ”once-in-a-lifetime” activation of a user, independent of the other users

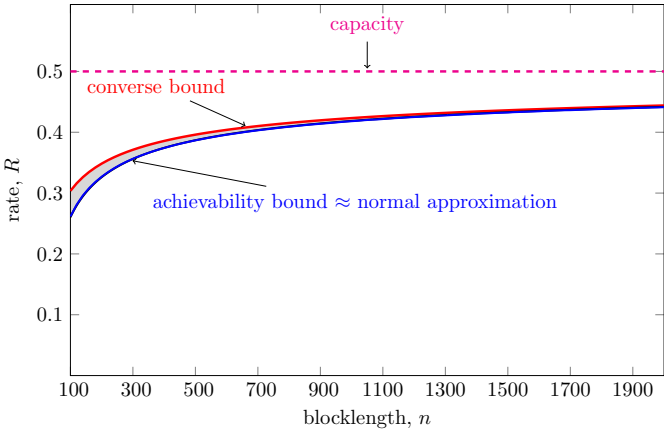
packet as an atomic unit

problems with the infinite population assumption

the packet cannot have a finite length
as the address size of a user $\log_2 N$ also goes to infinity.

finite blocklength effects in short packets





many-access channel (MnAC)

recent information-theoretic model suited for mMTC

$$\boldsymbol{Y} = \sum_{k=1}^{\ell_n} \boldsymbol{S}_k(w_k) + \boldsymbol{Z}$$

each user accesses the channel independently with α_n

if not accessing, a special zero codeword is sent

if accessing, one of the M messages is sent

number of possible users l_n tied to the packet blocklength n

X. Chen, T. Y. Chen and D. Guo, "Capacity of Gaussian Many-Access Channels," in IEEE Transactions on Information Theory, vol. 63, no. 6, pp. 3516-3539, June 2017.

unsourced massive access (U-RA)

Gaussian MAC with $h_{kn} = 1$.

K active users

D non-zero messages

all users have one and the same codebook with D codewords

- no user identification possible!

an active user chooses its message uniformly at random, independently of any other user

decoding is done up to a permutation of transmitted messages

Y. Polyanskiy, "A perspective on massive random-access," 2017 IEEE International Symposium on Information Theory (ISIT), Aachen, 2017, pp. 2523-2527.

further developments in U-RA

U-RA gained importance as a model for massive IoT communciation

potential for significant spectral and power efficiency gains compared to other dedicated IoT-driven schemes such as LoRA and even LTE cellular systems (in terms of bit/s/Hz per sector) *.

however, eventual practical use of U-RA requires capability to identify and authenticate users.

* Alexander Fengler, Giuseppe Caire, Peter Jung, and Saeid Haghighatshoar, “Massive MIMO Unsourced Random Access”

random access: two classical assumptions

1. packet is an atomic unit of information
2. users are activated independently

a_n - the activity of n–th user during the contention round
$a_k = 1$ if the user is active and $a_k = 0$ otherwise;

ALOHA-type: $\Pr(a_1, a_2, a_3, \dots, a_N) = \prod_{i=1}^{N} \Pr(a_i)$

general activation pattern has a general joint distribution
$\Pr(a_1, a_2, a_3, \dots, a_N)$

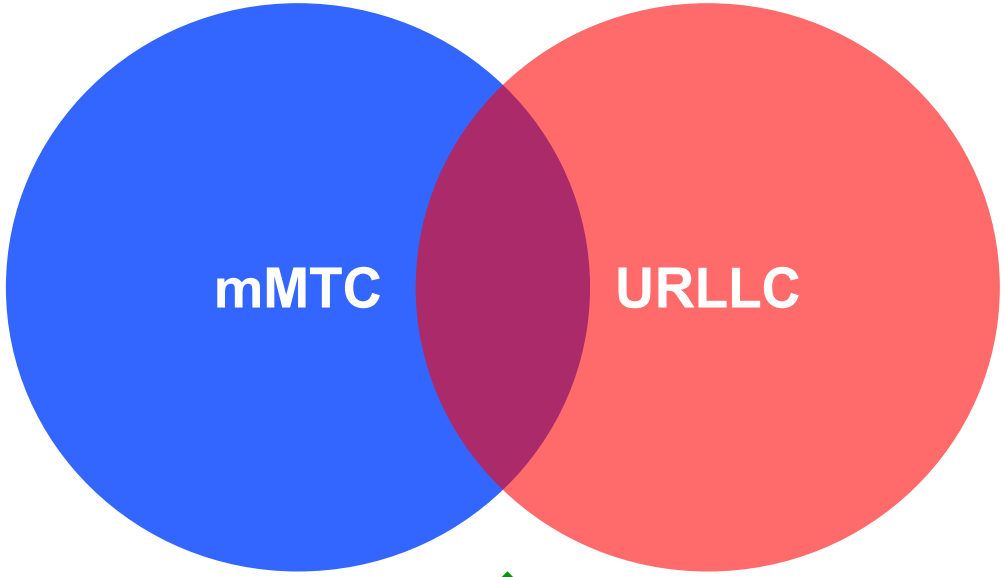
outline

6G and IoT connectivity

- massive access
- model for correlated access
- user identification in unsourced access
- massive downlink ACK

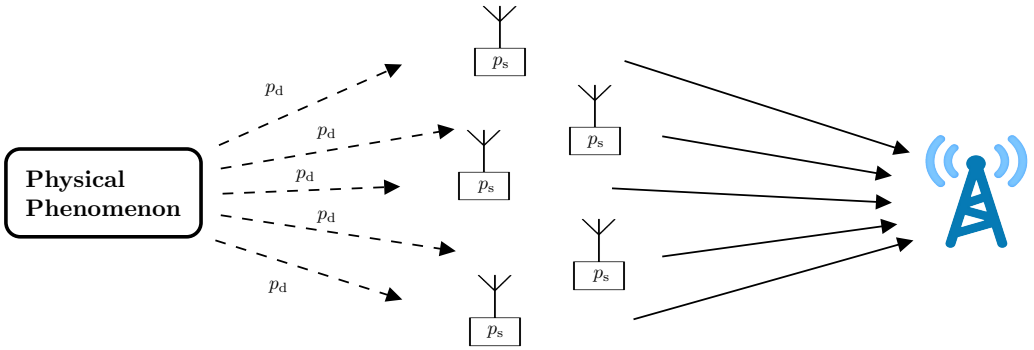
- ultra-reliable low-latency communications
- latency and general timing
- how to guarantee reliability
- URLLC in action

simultaneous massive and ultra-reliable?



possible when the information across devices is correlated (e.g. alarm)

the scenario



a device generates two message types

individual update, independent of others

alarm-type message, correlated for all sensors

modeling the tradeoff between massive and ultra-reliable

the alarm set has M_a messages

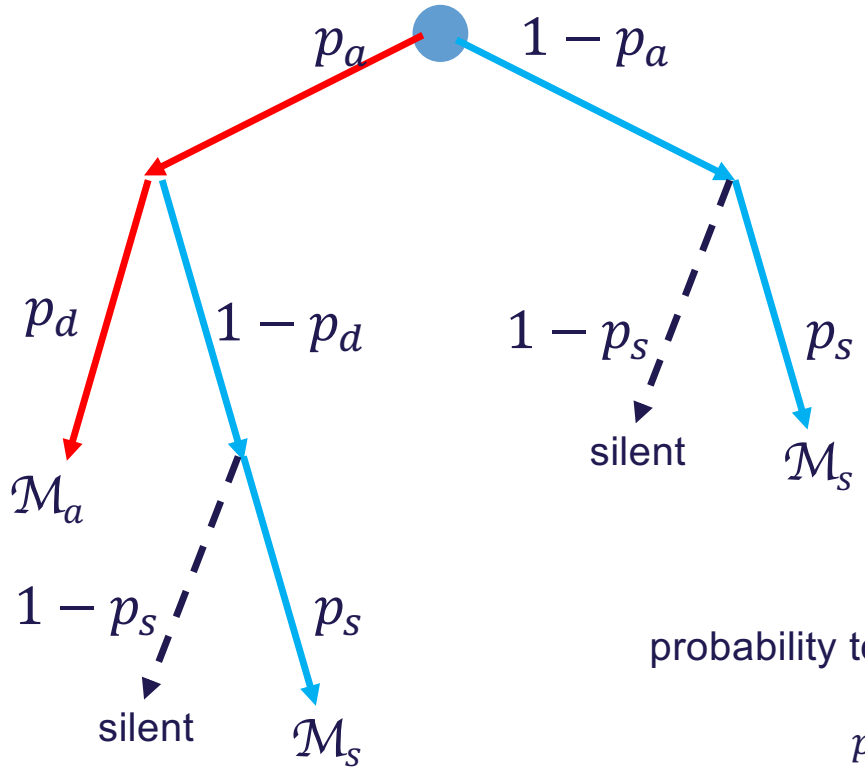
alarm occurs with probability p_a

when alarm occurs, the message to be sent
is chosen uniformly among the M_a messages

the same (common) alarm message
is sent by all nodes that detect it

a node detects alarm with probability p_d

modeling the tradeoff between massive and ultra-reliable



probability to send a standard message

$$p_s(1-p_a p_d)$$

modeling the tradeoff between massive and ultra-reliable

there are in total N devices and K of them are active

example of low spectral efficiency
low p_s, high p_a and high p_d, $N = 10$ and $K = 9$
then most likely all nodes send the same alarm message
the information generated in this transmission is low

example of high spectral efficiency
high p_s, low p_a and p_d high, large N

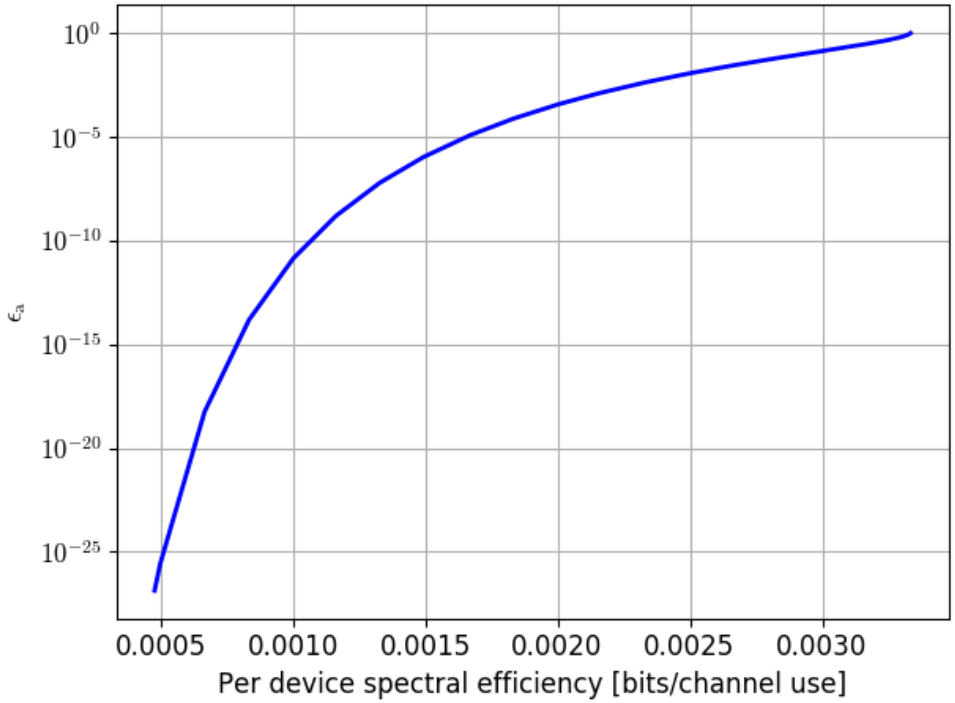
error events

the probability of error in alarm decoding is much lower

error in standard message decoding occurs
if two devices send the same standard message

errors due to false positive needs to be taken into account

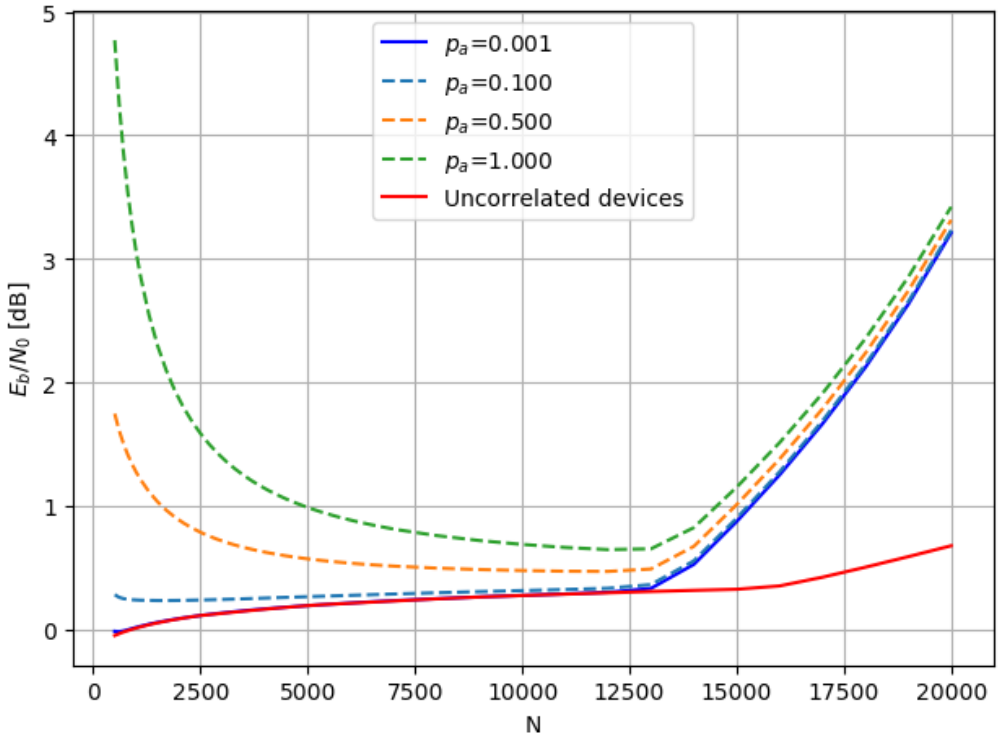
undetected alarm vs. spectral efficiency



Blocklength n = 30 000, N = 1000, target error probabilities $\varepsilon_s = 10^{-1}$, $\varepsilon_{fp} = 10^{-5}$,
message set sizes: standard message 100 bits, alarm message 3 bits. $p_s = 0.01$ and $p_a = 1$

We use p_d to control the spectral efficiency.

tradeoff EbN0 and number of devices



Blocklength n = 30 000, target error probabilities $\varepsilon_s = \varepsilon_{sa} = 10^{-1}$, $\varepsilon_{fp} = \varepsilon_a = 10^{-5}$,
message set sizes: standard message 100 bits, alarm message 3 bits. $p_s = 0.01$, p_d is optimized

outline

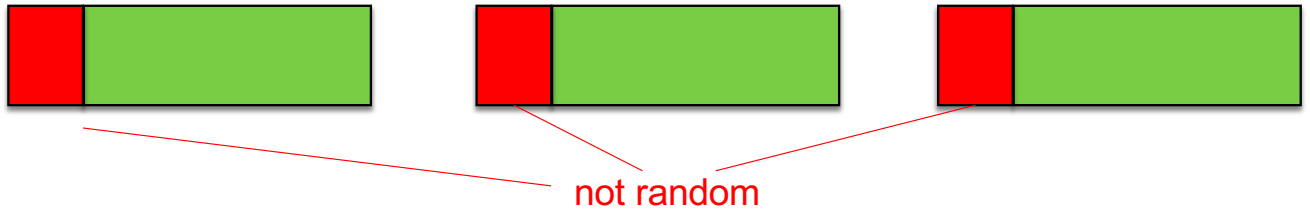
6G and IoT connectivity

- massive access
- model for correlated access
- **user identification in unsourced access**
- massive downlink ACK
- ultra-reliable low-latency communications
- latency and general timing
- how to guarantee reliability
- URLLC in action

U-RA and identification

true unsourced random access requires that
the messages are independent and identically distributed across users

- this precludes the use of traditional address added to the packet.
- however, straightforward identification is not possible



objective

key idea for providing identification

we use a message authentication codes (MAC)
which is a function of the data and
a unique secret key assigned to each device

- this is enough to preserve the capability
  to identify and authenticate the devices reliably

the resulting packet structure allows to simplify the transmitters
and use less power at the cost of increased processing at the receiver

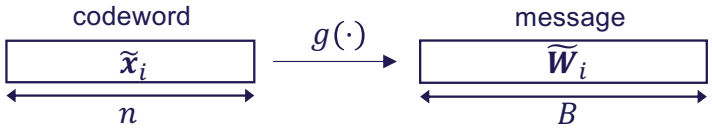
system model

at the base station the decoder $g: \mathcal{Y}^n \rightarrow [M]^K$ produces an unordered list of K messages

The K is fixed and it is assumed known to the base station

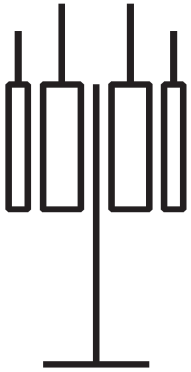
an error occurs whenever $g(\boldsymbol{y})$ does not contain a transmitted message, or if multiple users transmit the same message, i.e. $E_i = \{W_i \notin g(\boldsymbol{y})\} \cup \{W_i = W_j \; s.t. \; i \neq j\}$

- fixed K: for each error E_i, the list $g(\boldsymbol{y})$ contains a message which was not transmitted by any device
- these are false positives with probability of occurrence $P[E_i] = p_{FP}$

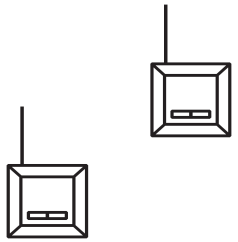


two-step protocol

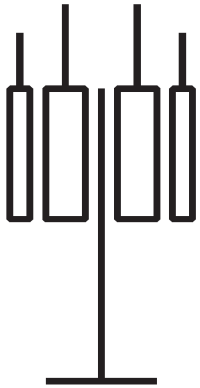
downlink phase



beacon with nonce $\boldsymbol{b}$



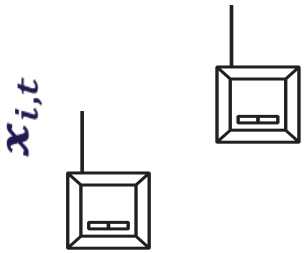
uplink phase



codeword $\boldsymbol{x}_{i,t}$

n

$\boldsymbol{x}_{i,t}$



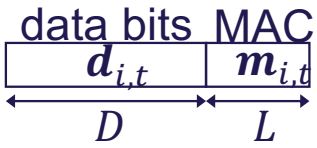
data bits $\boldsymbol{d}_{i,t}$ MAC $\boldsymbol{m}_{i,t}$

D L

operations at the user side

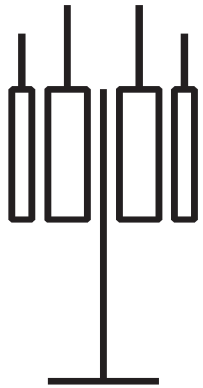
- the MAC $\boldsymbol{m}_i = \{0,1\}^L$ is generated by user i
  based on its data $\boldsymbol{d}_i \in \{0,1\}^D$ and its secret key $\boldsymbol{k}_i$, i.e. $\boldsymbol{m}_i = h(\boldsymbol{d}_i, \boldsymbol{k}_i, \boldsymbol{b})$
- the function h should be:
  computationally hard to invert and have low collision probability
- the nonce can be introduced to prevent replay attacks

MAC is appended to the data creating a complete message $W_i = [\boldsymbol{d}_i, \boldsymbol{m}_i]$



operations at the BS side

- the authenticity of the decoded message $\widetilde{W}_i = [\widetilde{\boldsymbol{d}}_i, \widetilde{\boldsymbol{m}}_i]$ is verified by recomputing the MAC and comparing it with the received one
- the BS tries different keys $\boldsymbol{k}$ to find a match
- since each key is a unique identifier of a device, finding a matching key provides an estimate of the identity of the sender



cryptographic errors

matching MAC is generated by a key
that does not belong to the actual sender

- the generated MAC might not be a unique identifier for the user
- the BS must generate many MACs with different secret keys
  to find the one that matches the one in the received packet

ideal MACs are assumed: the probability that a given (data, key) tuple produces a specific MAC of length L is $p = 2^{-L}$

- shorter MAC key leads to a higher probability of cryptographic errors

besides these errors, there are the usual transmission errors.

probability of cryptographic errors

type 1 error

- if the key of another user produces a match

$$p_{t1} = 1 - (1 - p)^{N-1}$$

where N is the number of users (keys).

type 2 error

- if the key of a user produces a valid MAC for more than one message

$$p_{t2} = 1 - (1 - p)^{K-1}$$

overall probability of successful authentication

$$p_{succ\_auth} = (1 - p_{t1})(1 - p_{t2}) = (1 - p)^{N+K-2}$$

other types of errors

authentication of a false positive:

the unsourced decoder produces a message not transmitted by anyone

- no keys match leading to a detectable false positive

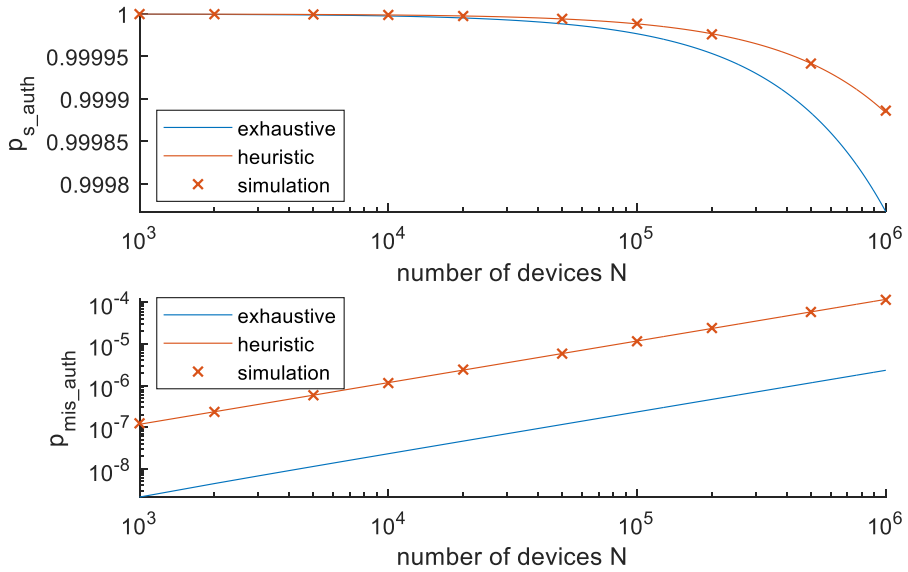
$$p_{d\_fp} = (1-p)^N$$

- exactly one key matches causing erroneous authentication

$$p_{fp\_auth} = (N - k_{TP})p(1-p)^{N+K-2}$$

heuristic search

- the authenticator tries keys only until it finds a matching key
- cannot detect type 1 and type 2 errors
- the number of operations is on average reduced by half
  but at the cost of higher mis-authentication probability



probability of successful authentication and probability of mis-authentication as a function of the total number of devices N. The number of messages is $K = 100$, $p_{FP} = 0.01$, *MAC length* $L = 32$ *bits.*

numerical illustration

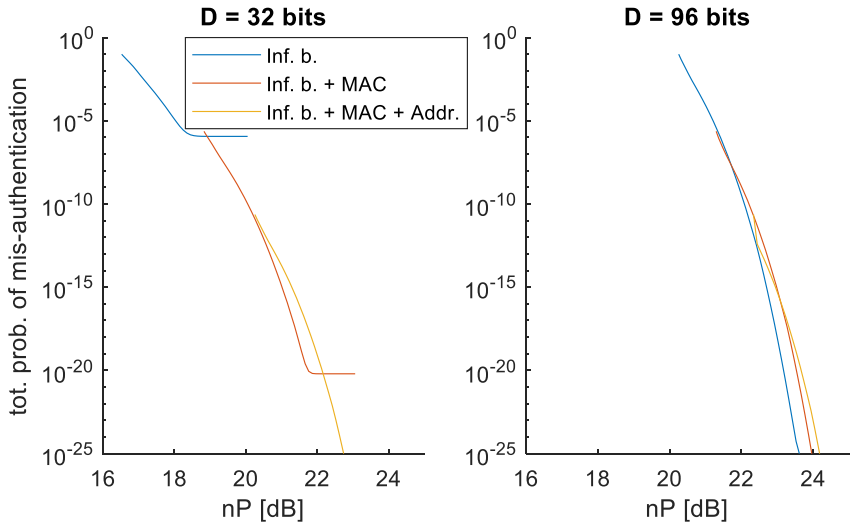
total error probability

- for a packet consisting of only data bits D
- in a proposed scheme where packet contains data and MAC, and the length is $D + L$
- for a classical packet structure that includes the address for a total of $D + L + A$ bits

$$p_{mis\_auth} = p_{FP}$$

$$p_{mis\_auth} = p_{FP} p_{fp\_auth}$$

$$p_{mis\_auth} = p_{FP} p$$



total probability of mis-authentication as a function of energy. The number of messages is $K = 100$ *and* $N = 10^5$, $L = A = 32$ *bits*

outline

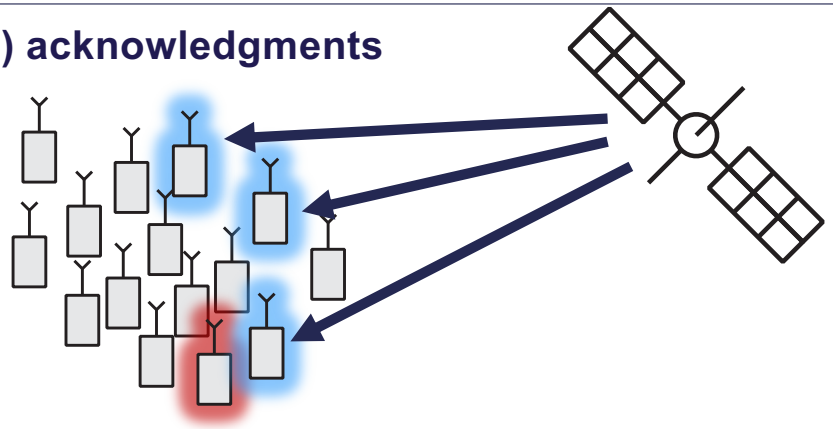
6G and IoT connectivity

- massive access
- model for correlated access
- user identification in unsourced access
- massive downlink ACK
- ultra-reliable low-latency communications
- latency and general timing
- how to guarantee reliability
- URLLC in action

common acknowledgments

1) random access

2) acknowledgments



can we do better than concatenation?

user 65	user 103	user 211

what are the limits and trade-offs?

formal problem definition

- $[N] = \{0,1, \dots, N-1\}$: set of potentially active users (e.g., $N = 2^{64}$)
- $\mathcal{A}$: set of active users ($|\mathcal{A}| \ll N$)
- $\mathcal{S} = \{s_1, s_2, \dots, s_K\}$: set of K ***recovered*** users
- K is constant ($K \ll N$, e.g., $\mathrm{K} = 100$)

encoder

$$f: \binom{[N]}{K} \to \{0,1\}^B$$

user *n*'s decoder

$$g_n: \{0,1\}^B \to \{0,1\}$$

message length

error types

false positives

$$\varepsilon_{\mathrm{fp}} = \mathbb{E}\left[\Pr\left(g_n\left(f(\mathcal{S})\right) = 1 \;\middle|\; n \in \mathcal{A} \setminus \mathcal{S}\right) \mid K\right]$$

false negatives

$$\varepsilon_{\mathrm{fn}} = \mathbb{E}\left[\Pr\left(g_n\left(f(\mathcal{S})\right) = 0 \;\middle|\; n \in \mathcal{S}\right) \mid K\right]$$

depend only on the **active** users ($s_k \in \mathcal{A}$), and not the **inactive** users

error-free encoding

i.e., $\varepsilon_{\mathrm{fp}} = \varepsilon_{\mathrm{fn}} = 0$

there are $\binom{N}{K}$ ways to pick the K recovered users, so we need

$$B^*_{\mathrm{error-free}} = \left\lceil \log_2 \binom{N}{K} \right\rceil \quad [\mathrm{bits}]$$

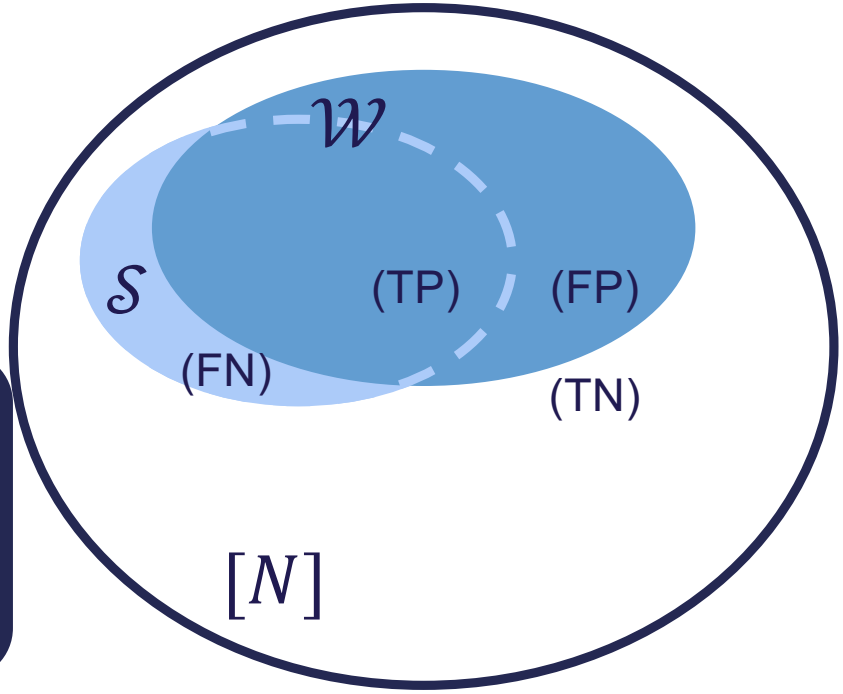
$$\geq \left\lceil \mathrm{K} \log_2 \left(\frac{N}{K} \right) \right\rceil \quad [\mathrm{bits}]$$

encoding with errors

i.e., $\varepsilon_{\text{fp}} > 0,\ \varepsilon_{\text{fn}} \geq 0$

each message $\mathcal{W}$ can be used for several sets $\mathcal{S}$

$$B^*_{\text{fp,fn}} \geq K \log_2 \left(\frac{1}{\varepsilon_{\text{fp}} + \frac{K}{N}} \right) - K \log_2 \left(\frac{e}{1 - \varepsilon_{\text{fn}}} \right)$$
$$- \varepsilon_{fn} K \log_2 \left(\frac{1 - \varepsilon_{\text{fn}}}{\varepsilon_{\text{fn}} \left(\varepsilon_{\text{fp}} + \frac{K}{N} \right)} \right) - \log_2 K \text{ [bits]}$$



R. Pagh and F. F. Rodler, "Lossy dictionaries," in *Eur. Symp. Algorithms*. Springer, 2001, pp. 300–311

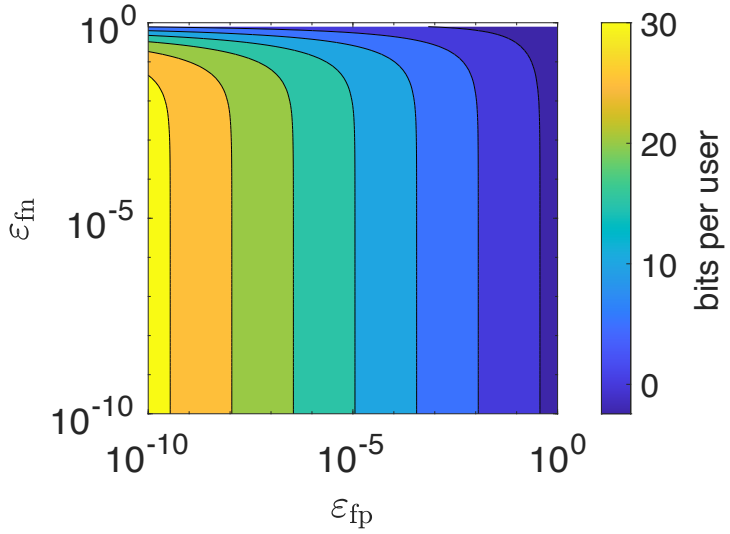
encoding with errors

does not depend on N as $N \to \infty$ for fixed K

$$B^*_{\mathrm{fp,fn}} \geq K \log_2 \left(\frac{1}{\varepsilon_{\mathrm{fp}} + \cancel{\frac{K}{N}}} \right) - K \log_2 \left(\frac{e}{1-\varepsilon_{\mathrm{fn}}} \right) - \varepsilon_{\mathrm{fn}} K \log_2 \left(\frac{1-\varepsilon_{\mathrm{fn}}}{\varepsilon_{\mathrm{fn}} \left(\varepsilon_{\mathrm{fp}} + \cancel{\frac{K}{N}} \right)} \right) - \log_2 K$$

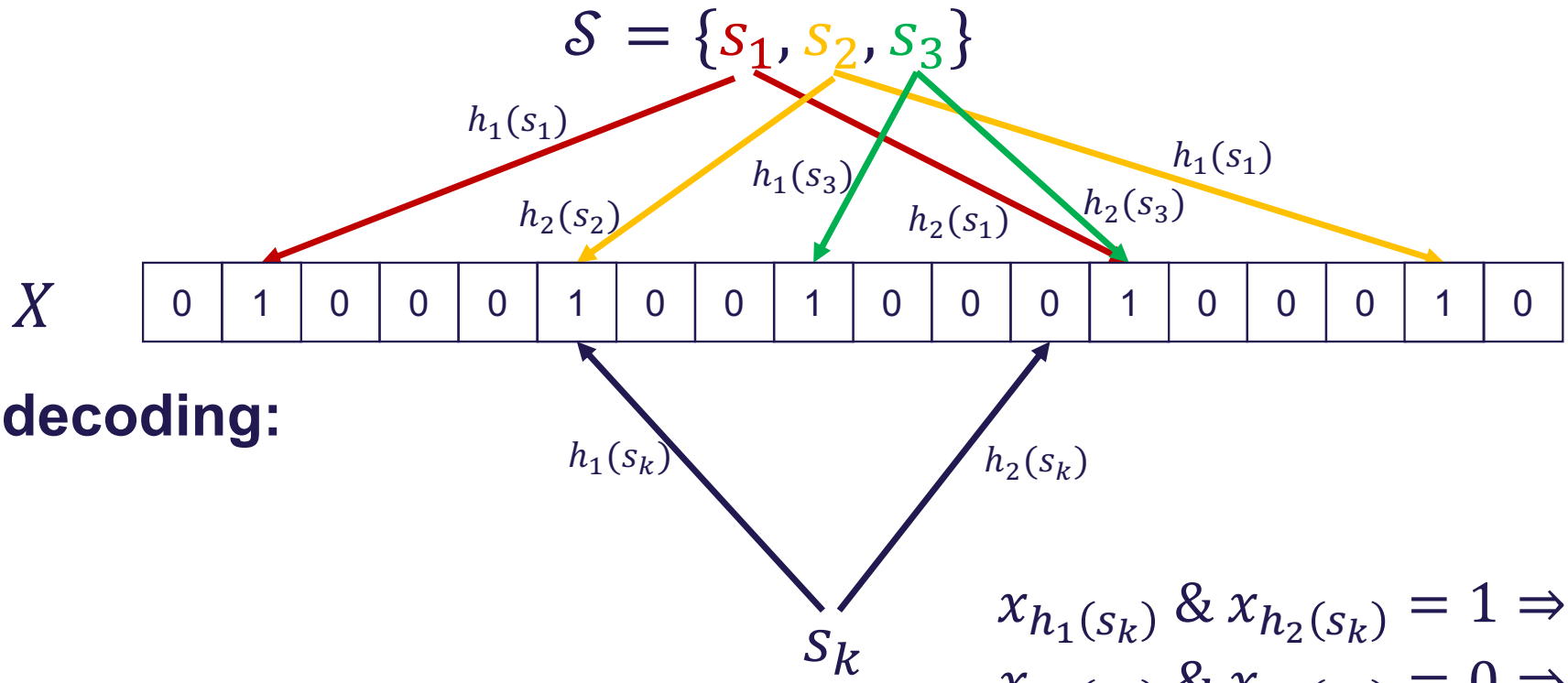
false positives give the highest gains

for large N: $B^*_{\mathrm{fp}} = K \log_2 \left(\frac{1}{\varepsilon_{\mathrm{fp}}} \right) \pm \mathcal{O}(\log \log N)$



Bloom filter

encoding:



decoding:

$$x_{h_1(s_k)} \,\&\, x_{h_2(s_k)} = 1 \Rightarrow s_k \in \mathcal{S}$$
$$x_{h_1(s_k)} \,\&\, x_{h_2(s_k)} = 0 \Rightarrow s_k \notin \mathcal{S}$$

linear equations

consider the set of K linear equations constructed using hashes of the user ids

$$\mathcal{S} = \{s_1, s_2, s_3\}$$

$$\begin{bmatrix} h_1^{(1)}(s_1) & h_1^{(2)}(s_1) & h_1^{(3)}(s_1) \\ h_1^{(1)}(s_2) & h_1^{(2)}(s_2) & h_1^{(3)}(s_2) \\ h_1^{(1)}(s_3) & h_1^{(2)}(s_3) & h_1^{(3)}(s_3) \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix} = \begin{bmatrix} h_2(s_1) \\ h_2(s_2) \\ h_2(s_3) \end{bmatrix}$$

unknown vector

All hash functions are $[N] \rightarrow \mathrm{GF}(2^p)$

M. Dietzfelbinger and R. Pagh, "Succinct data structures for retrieval and approximate membership," in *Int. Colloq. Automata, Languages, and Program.* Springer, 2008, pp. 385–396.
E. Porat, "An optimal bloom filter replacement based on matrix solving," in Int. Comput. Sci. Symp. Russia. Springer, 2009, pp. 263–273.

linear equations

$$\begin{bmatrix} h_1^{(1)}(s_1) & h_1^{(2)}(s_1) & h_1^{(3)}(s_1) \\ h_1^{(1)}(s_2) & h_1^{(2)}(s_2) & h_1^{(3)}(s_2) \\ h_1^{(1)}(s_3) & h_1^{(2)}(s_3) & h_1^{(3)}(s_3) \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix} = \begin{bmatrix} h_2(s_1) \\ h_2(s_2) \\ h_2(s_3) \end{bmatrix}$$

All hash functions are $[N] \to \mathrm{GF}(2^p)$

decoding:

$$h_1^{(1)}(s_k)z_1 + h_1^{(2)}(s_k)z_2 + h_1^{(3)}(s_k)z_3 = h_2(s_k) \Rightarrow s_k \in \mathcal{S}$$

all we need to send is $[z_1 \quad z_2 \quad z_3]^T$ ← Kp bits

$$\varepsilon_{\mathrm{fp}} = 2^{-p} \Leftrightarrow p = \left\lceil \log_2 \left(\frac{1}{\varepsilon_{\mathrm{fp}}} \right) \right\rceil$$

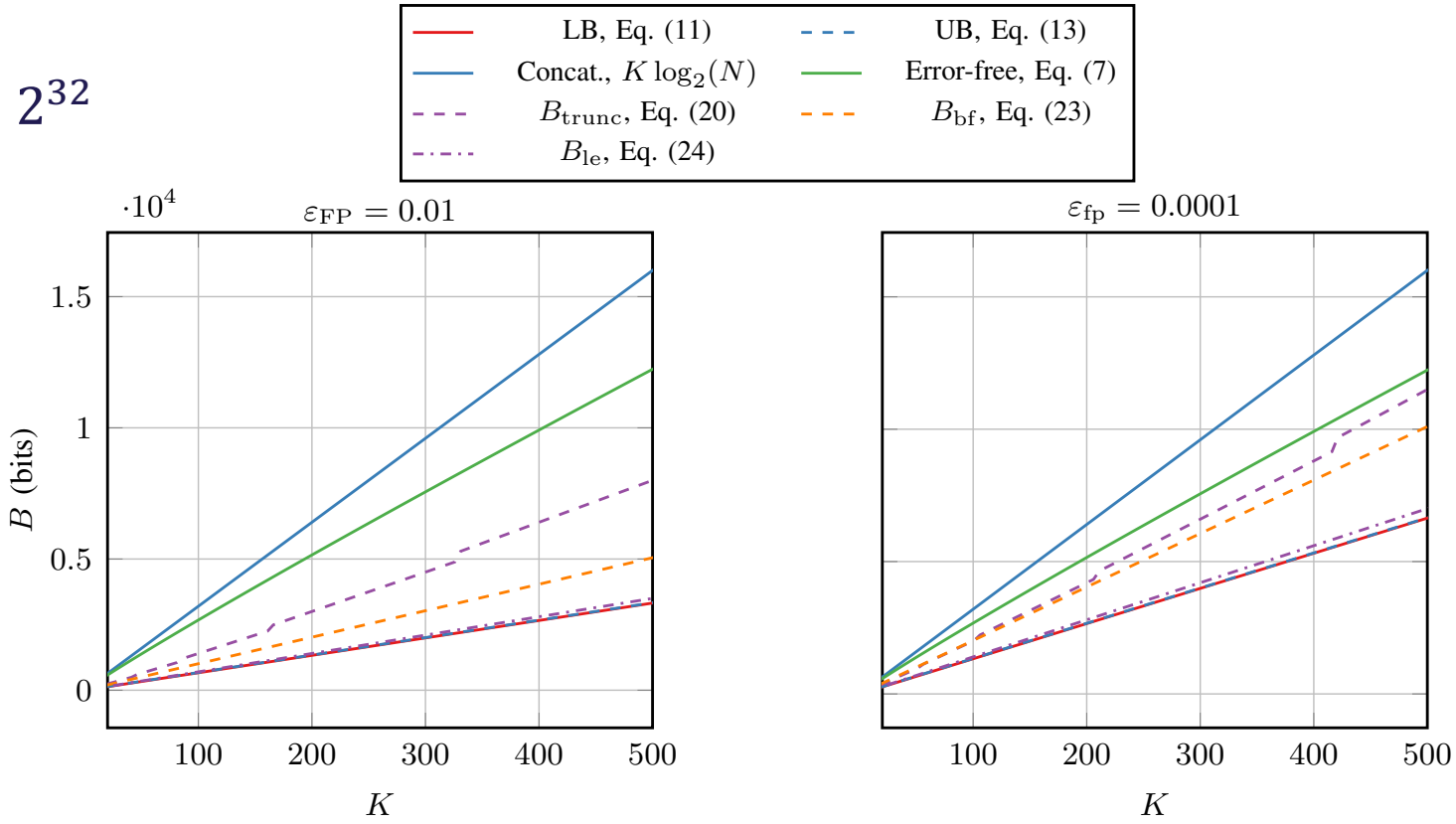
recall the bound:

$$B_{\mathrm{fp}}^* = K \log_2 \left(\frac{1}{\varepsilon_{\mathrm{fp}}} \right) \pm \mathcal{O}(\log \log N)$$

M. Dietzfelbinger and R. Pagh, “Succinct data structures for retrieval and approximate membership,” in *Int. Colloq. Automata, Languages, and Program.* Springer, 2008, pp. 385–396.
E. Porat, “An optimal bloom filter replacement based on matrix solving,” in Int. Comput. Sci. Symp. Russia. Springer, 2009, pp. 263–273.

comparison

$N = 2^{32}$



random K

we assume that K is included in the feedback packet (small additional cost)

	variable-length ACK coding	**fixed-length ACK coding**
packet length	variable	constant
$\varepsilon_{\mathrm{fp}}$	constant	variable

bounds in paper

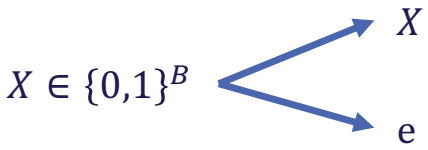
variable field order
$GF\left(2^{\left\lceil \log_2\left(\frac{1}{\varepsilon_{\mathrm{fp}}}\right)\right\rceil}\right)$

downlink erasure channel

encoder

$$f: \binom{[N]}{K} \rightarrow \{0,1\}^B$$

user *n*'s channel



user *n*'s decoder

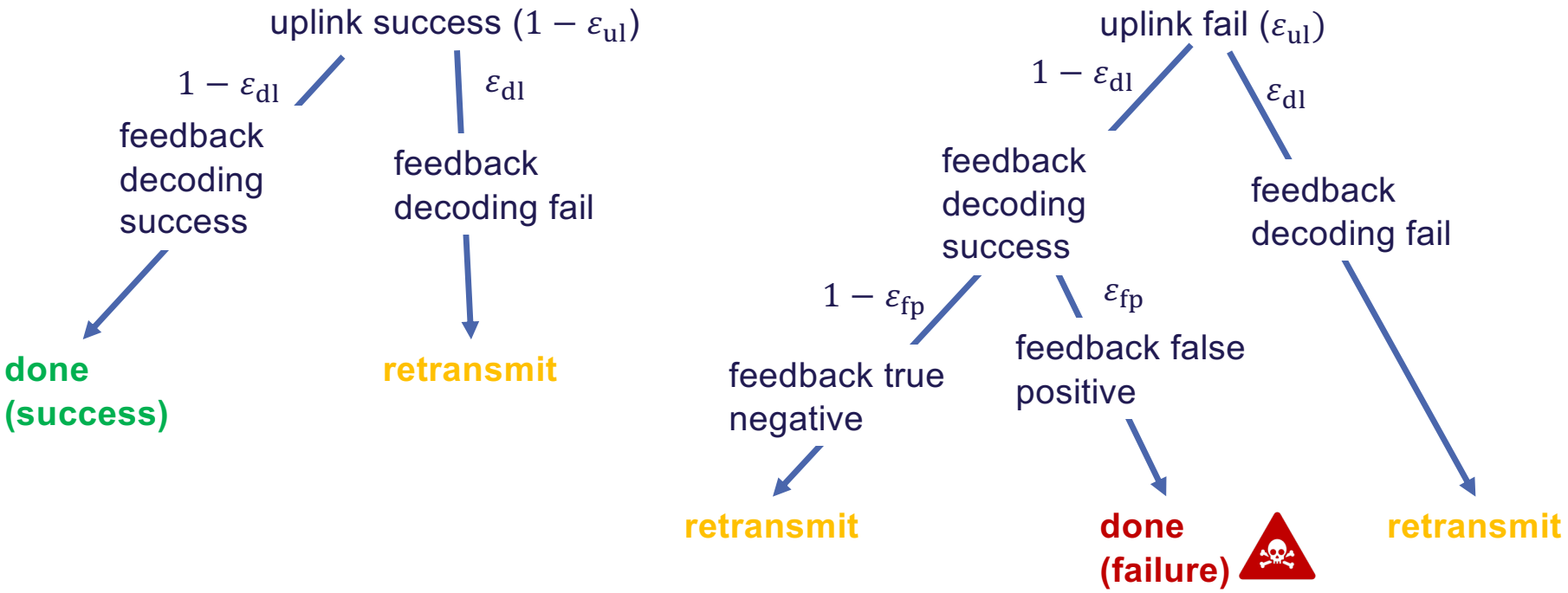
$$g_n: \{0,1\}^B \cup \mathrm{e} \rightarrow \{0,1\}$$

erasure probability assumed to be equal to the outage probability

for evaluation we will assume:

- fixed-length coding
- Poisson arrivals
- Rayleigh fading
- 2048 symbols
- 64 tx antennas

ARQ model

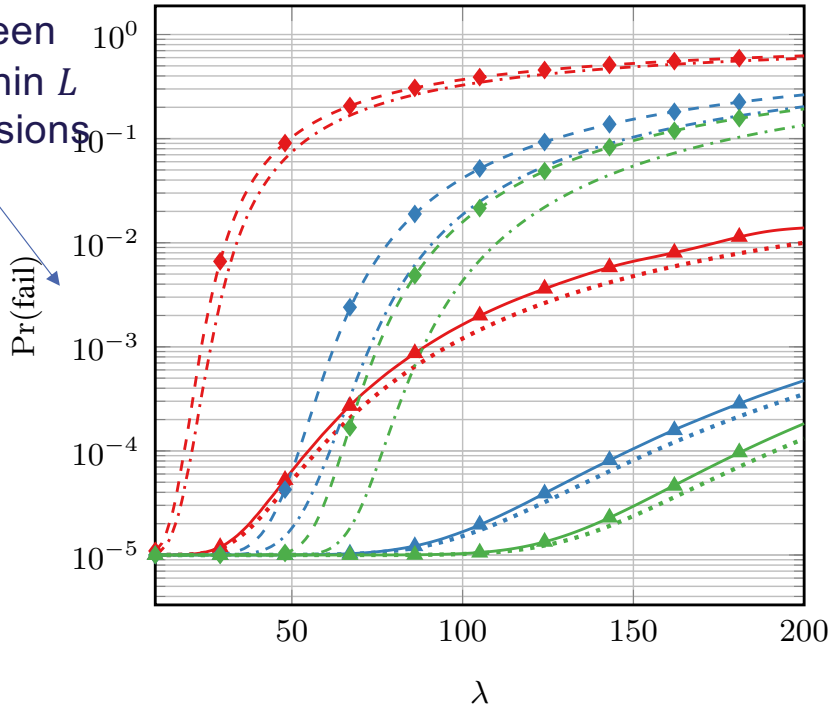


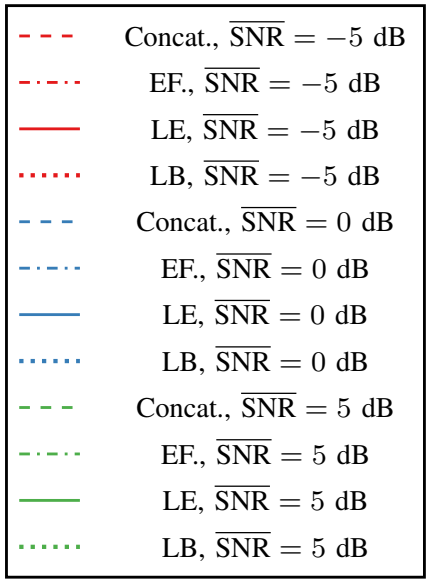
reliable feedback is a **trade-off** between

reliable transmission and **false positive** probability

fixed-length feedback with fading

not in green state within L transmissions





- $L = 5$ retransmissions
- $\varepsilon_{\text{ul}} = 0.1$
- $K \sim \text{Poisson}(\lambda)$ (iid in each retransmission)
- Rayleigh fading
- 2048 symbols
- 64 tx antennas
- markers indicate simulations

- more efficient coding allows for lower transmission rate
- significanly higher reliability despite false positives

resolving failures

- some users erroneously believe they succeeded when they fail

- false positives exist in all ARQ systems (CRC failures, etc.)
  - example: 16-bit CRC gives $\varepsilon_{\mathrm{fp}} \approx 1.5 \cdot 10^{-5}$
  - ACK messages are usually designed to have $\varepsilon_{\mathrm{fp}} \ll \varepsilon_{\mathrm{fn}}$,

    but we have assumed the opposite

- need to be resolved at higher layers, e.g., using sequence numbers

key references

- K. Stern, A. E. Kalør, B. Soret and P. Popovski, "Massive Random Access with Common Alarm Messages," 2019 IEEE International Symposium on Information Theory (ISIT), 2019

- R. Kotaba, A. E. Kalør, P. Popovski, I. Leyva-Mayorga, B. Soret, M. Guillaud, and L. Ordonez, "How to Identify and Authenticate Users in Massive Unsourced Random Access," vol. 25, no. 12, pp. 3795-3799, Dec. 2021

- A. E. Kalør, R. Kotaba and P. Popovski, "Common Message Acknowledgments: Massive ARQ Protocols for Wireless Access," in IEEE Transactions on Communications, vol. 70, no. 8, pp. 5258-5270, Aug. 2022

outline

6G and IoT connectivity

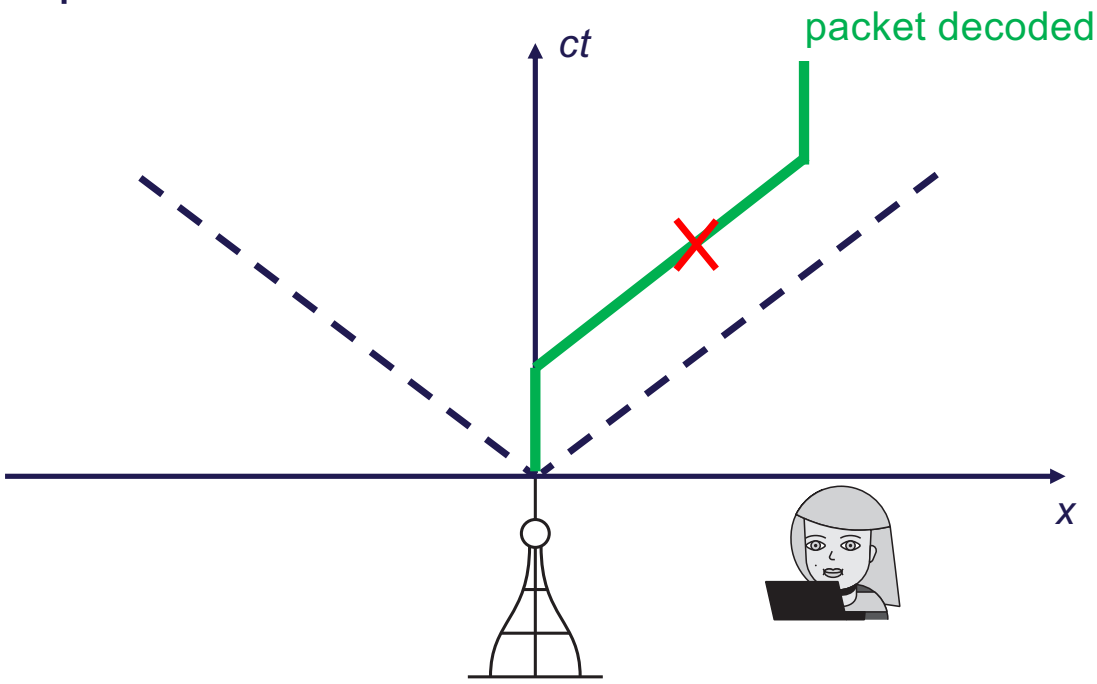
- massive access
- model for correlated access
- user identification in unsourced access
- massive downlink ACK

- **ultra-reliable low-latency communications**
- latency and general timing
- how to guarantee reliability
- URLLC in action

latency vs. reliability

defining communication as a reproduction of a data or a process at a spatially separate location

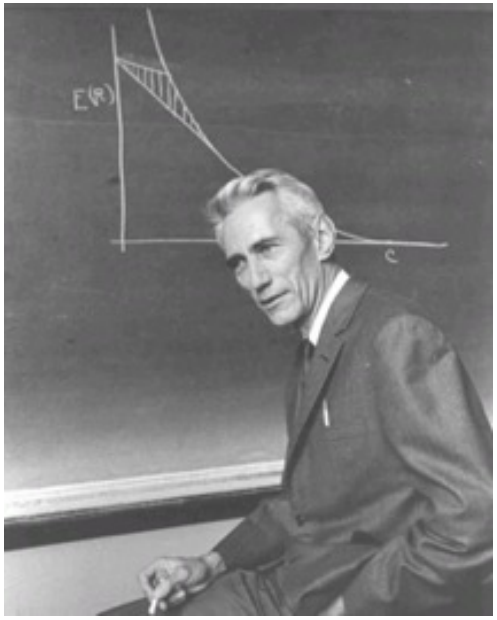
- zero latency phycially impossible: spacetime diagram
- reliability: the packet may not arrive at all



latency vs. reliability in communication models

an early choice in 5G URLLC has been

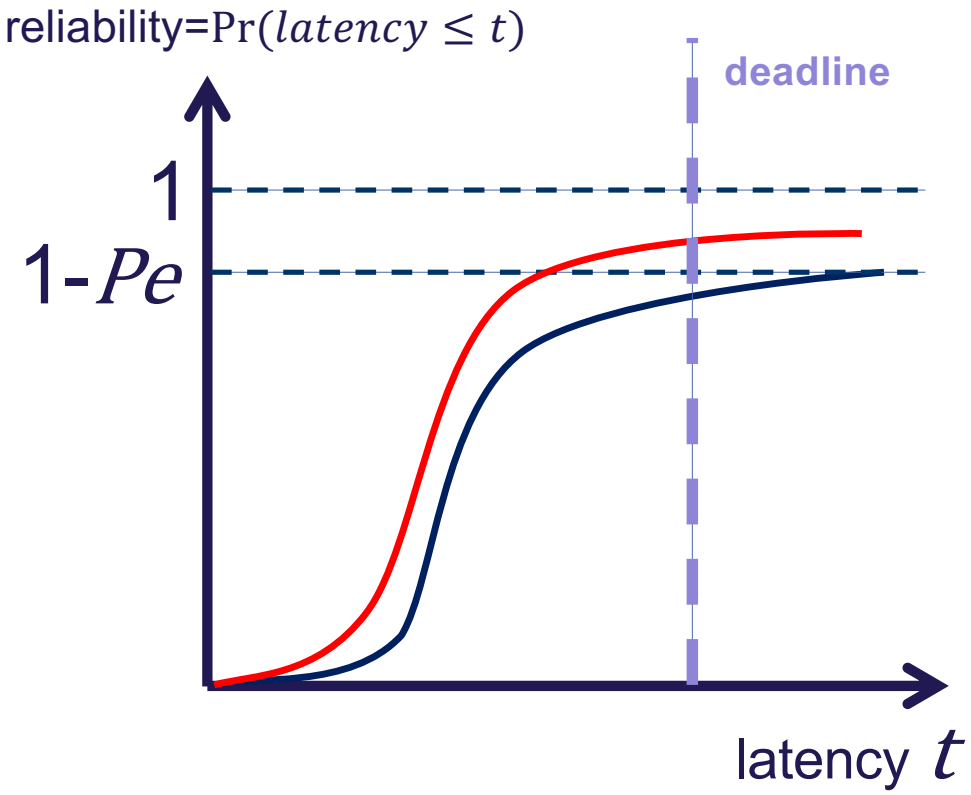
32 bytes at error rate of 10^{-5} at a latency of 1 ms



information theory guarantees perfect reliability!

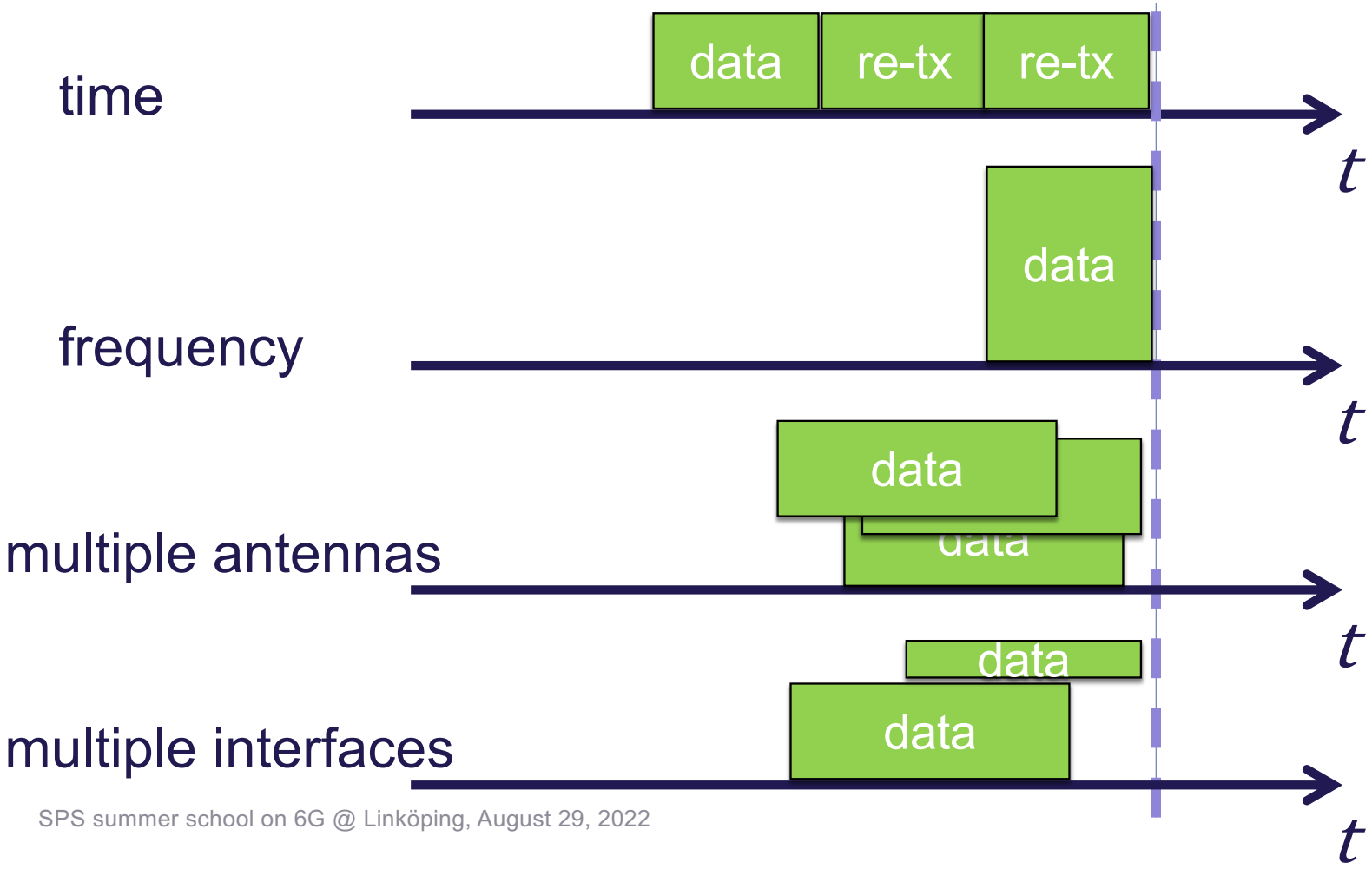
- ...provided that the sender knows the channel statistics to be able to select the data rate to be below the channel capacity
- at the price of an infinite latency

latency-reliability characterization for fixed data size



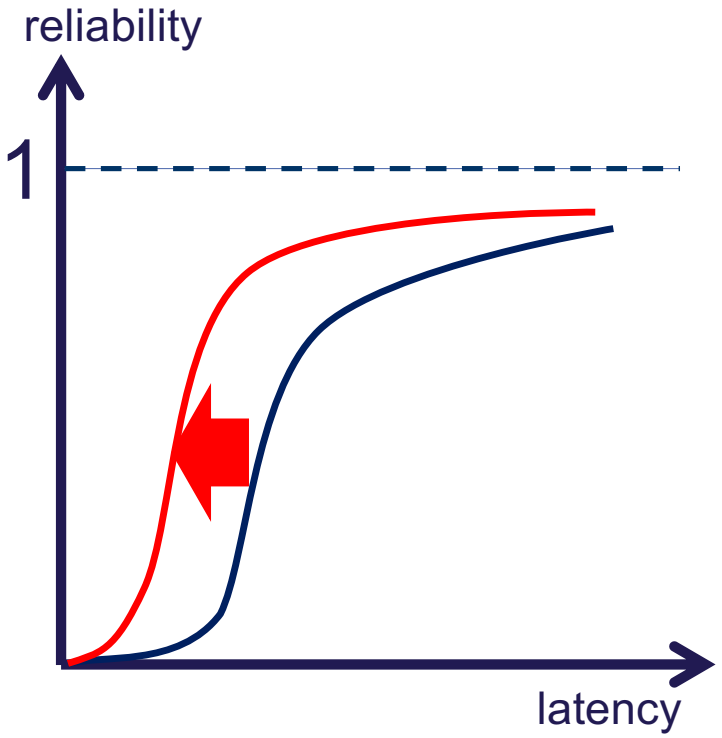
we move from the blue curve to the red curve by using diversity

diversity options

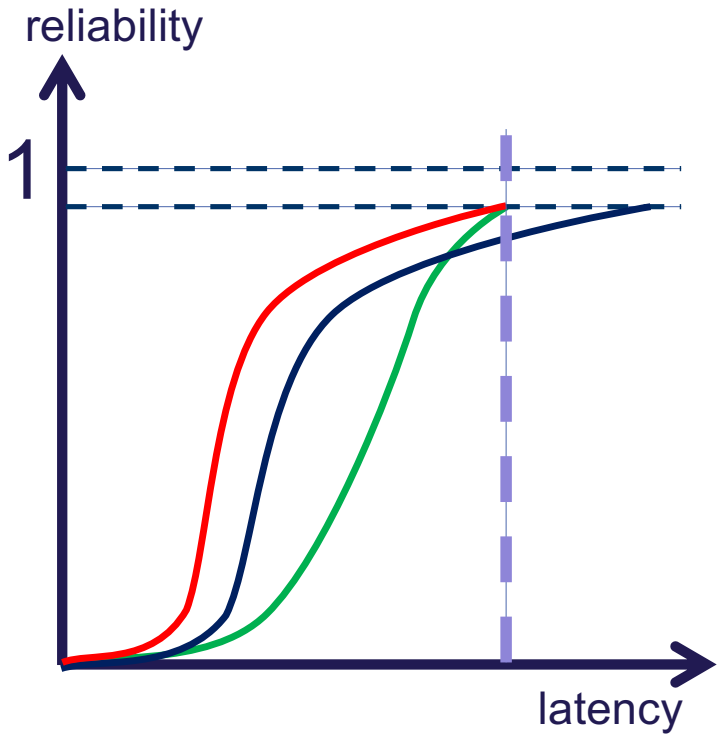


design targets

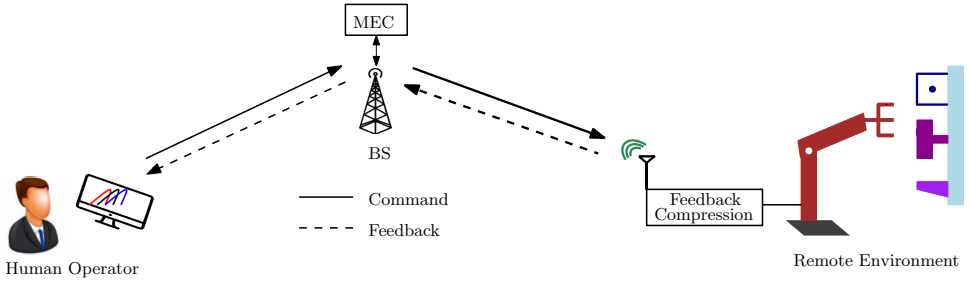
broadband rate-oriented systems



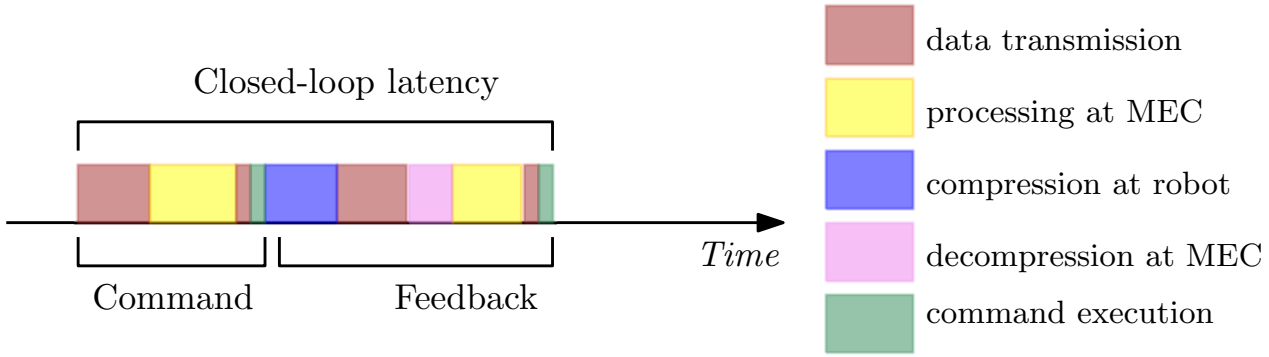
ultra-reliable low latency communication URLLC



a more comprehensive latency budget

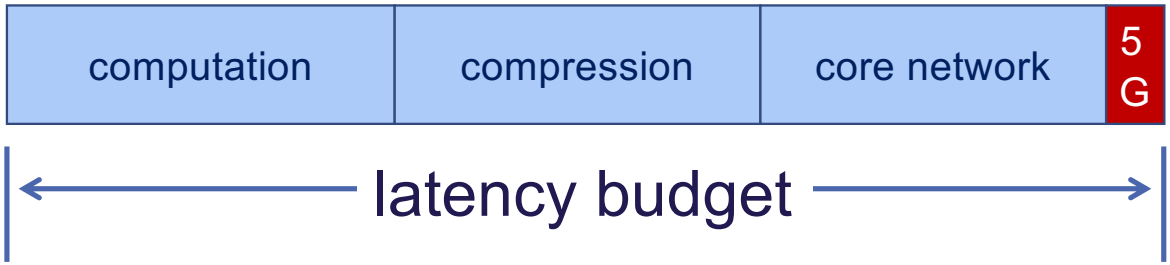


robotic hand controlled
via a Mobile Edge Computing



S. Suman, Č. Stefanović, S. Došen, and P. Popovski, "Analysis and Optimization of the Latency Budget in Wireless Systems with Mobile Edge Computing", in Proc. IEEE ICC, Seoul, Korea, May 2022.

how low the communication latency should be?



the idea with low values (~1 ms) is to
cut a low, predictable part of the latency buget

- invest latency budget into other operations to mitigate commnunication failures
- paradoxically, communication should work very well! (ultra-reliable)

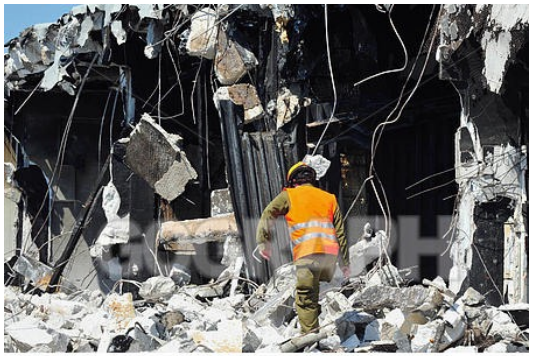
early METIS proposal [1]:

- ultra-reliable communications over a **long** term: latency >10 ms
- ultra-reliable communications over a **short** term: latency ≤ 10 ms

P. Popovski, "Ultra-reliable communication in 5G wireless systems," in 1st International Conference on 5G for Ubiquitous Connectivity, Nov. 2014, pp. 146–151.

ultra-reliable services with relaxed latency

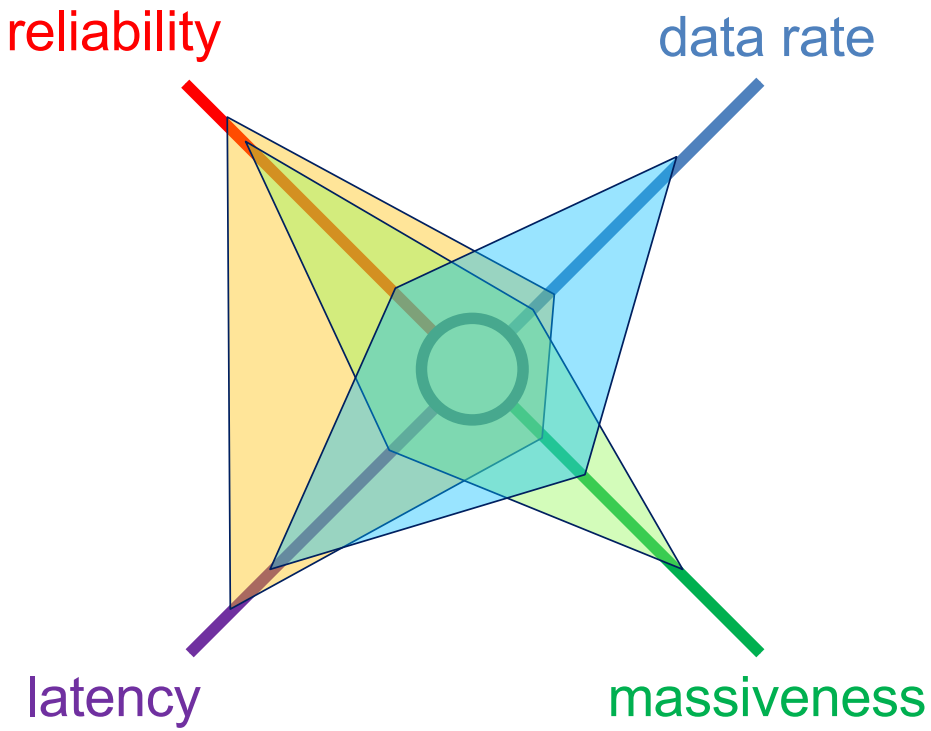
mobile health,
remote monitoring



disaster and rescue

smart grid

expanding the 5G triangle to 4 dimensions



tactile robotics

health monitors

augmented reality

outline

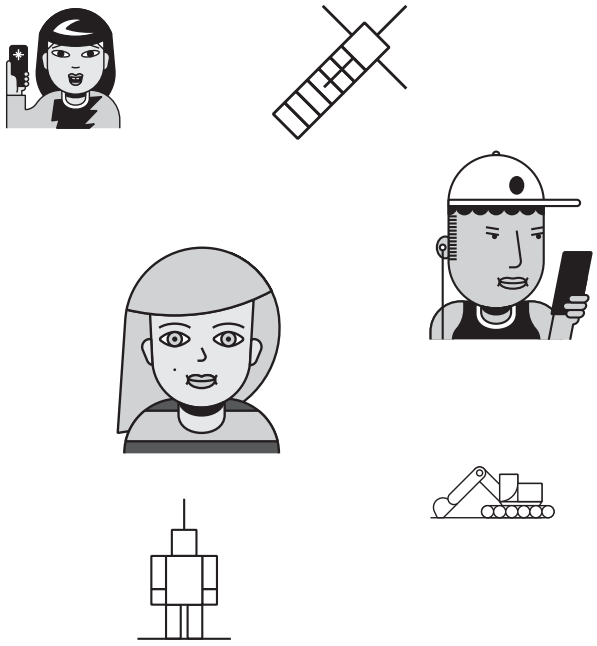
6G and IoT connectivity

- massive access
- model for correlated access
- user identification in unsourced access
- massive downlink ACK

- ultra-reliable low-latency communications
- latency and general timing
- how to guarantee reliability
- URLLC in action

understanding the more general timing requirements

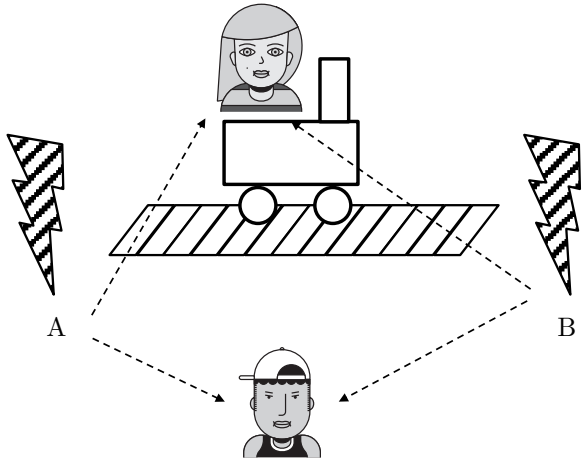
- perception of time by humans and machines
  - Tactile Internet or Internet of Senses
- wireless connectivity augments the natural time-space context
- digital time gets intertwined with physical time
  - revisiting simultaneity, presence, causality
- increased interest in various timing measures
  - latency, Age of Information and its derivatives



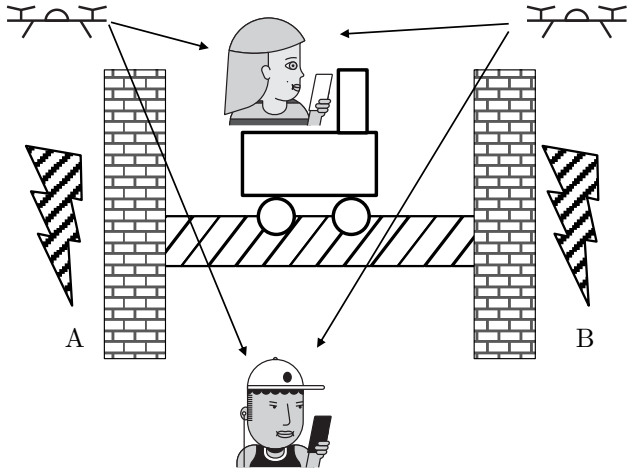
P. Popovski, F. Chiariotti, K. Huang, A. E. Kalør, M. Kountouris, N. Pappas, and B. Soret, "A Perspective on Time towards Wireless 6G," arXiv:2106.04314, June 2021.

time, real-time, and simultaneity

- real-time not universally defined
  - simultaneity and causality key in defining timing
  - horizon of simultaneity



Einstein's special relativity



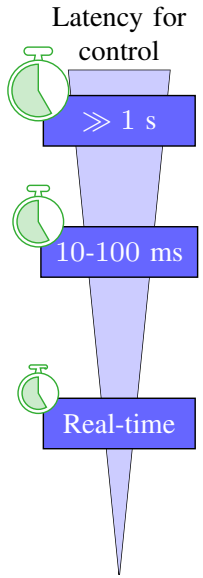
simultaneity in digital time

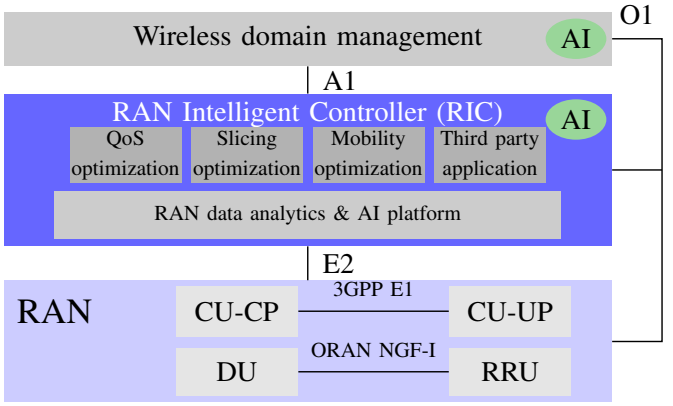
timing scales and requirements

general tendency in 6G towards smaller timing scales

O-RAN (Open Radio Access Network) alliance defines 3 categories

- real-time
- near-real time
- non-real time





similarly, 5G-ACIA defines hard real-time, soft real-time, and non real-time

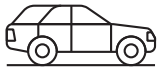
timing and communication actors

human-human

- inter-sensory synchronicity
- image processed by brain
  if eyes can see it for at least 13 ms*

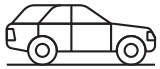
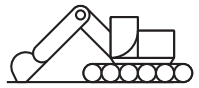

human-machine

- humans perceive a response time
  of 100 ms as instantaneous



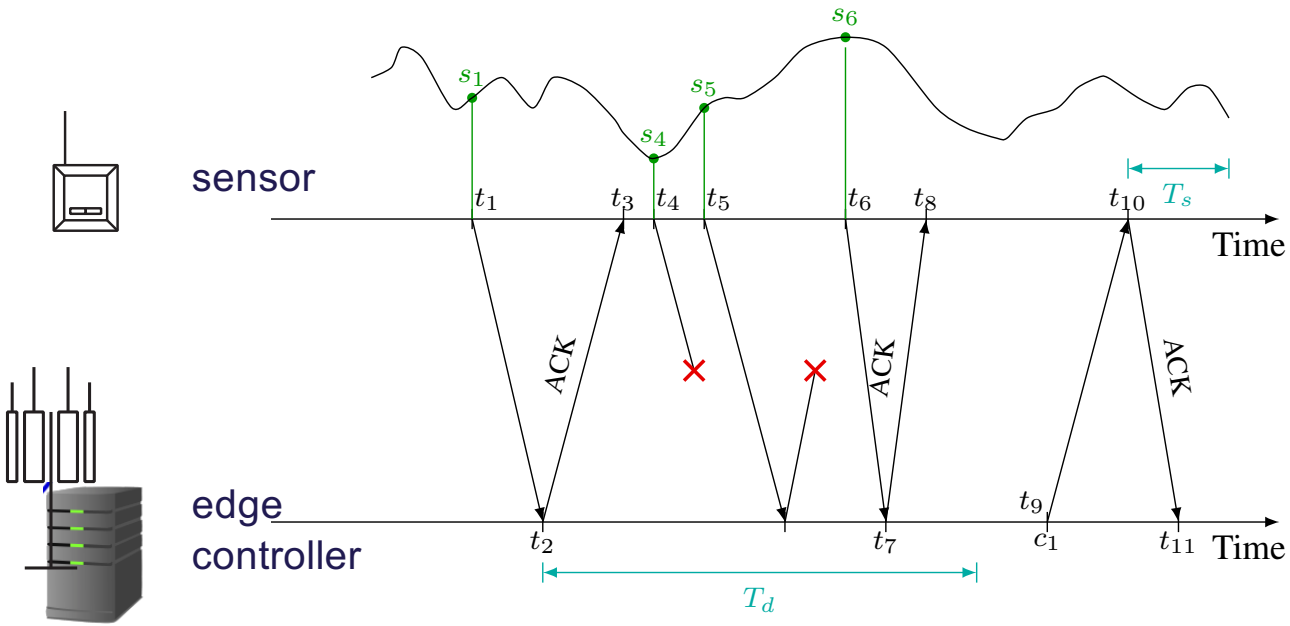

machine-machine

- required timings adjustable by design

A. Boccolini, A. Fedrizzi, and D. Faccio, “Ghost imaging with the human eye,” Optics Express, vol. 27, no. 6, pp. 9258–9265, Mar. 2019.

towards more general timing: references



time delay in learning each other's state

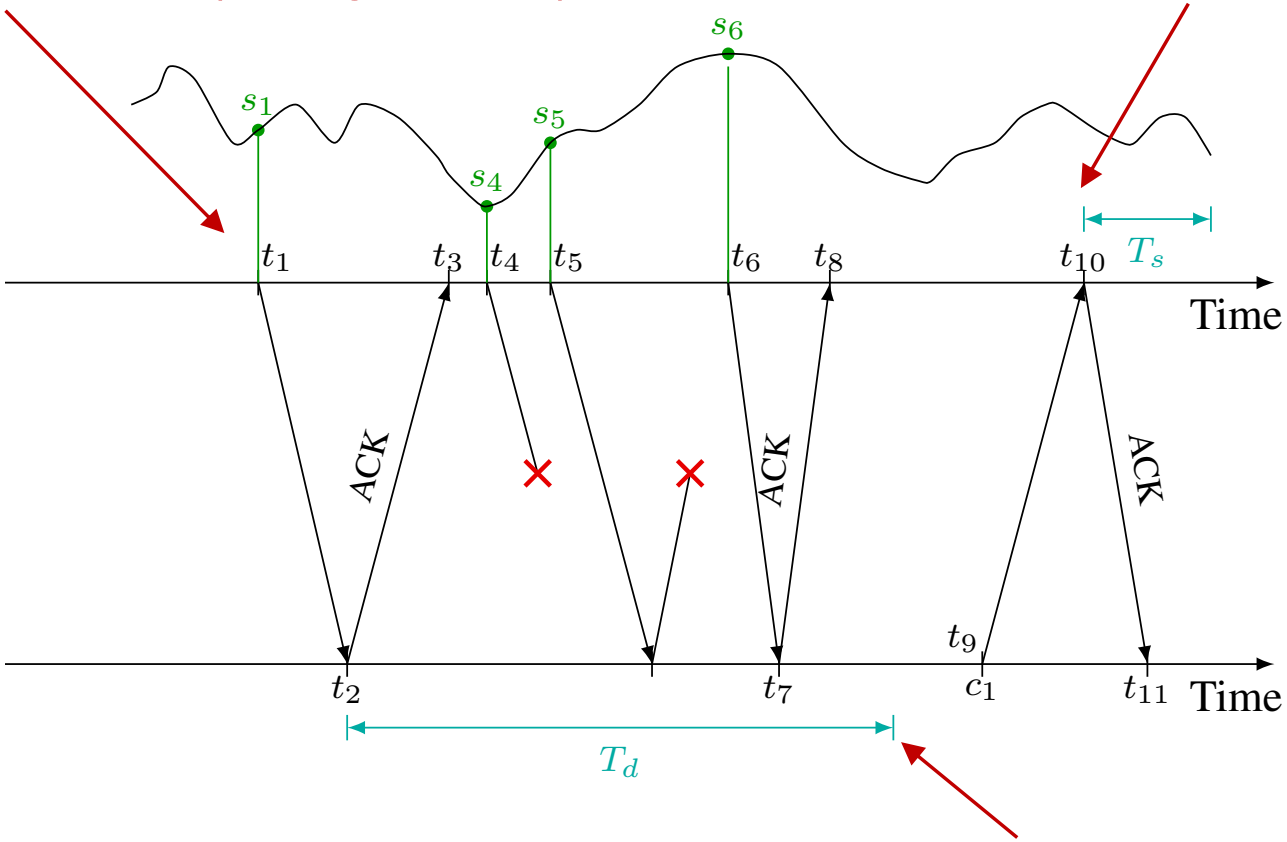
if the controller does not get any update within T_d, it shuts down the system

the sensor goes to sleep T_s seconds after receiving a command

timing references

past timing reference (timing anchor)

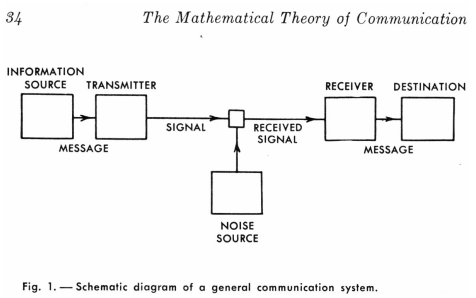
relative timing reference



future timing reference (deadline)

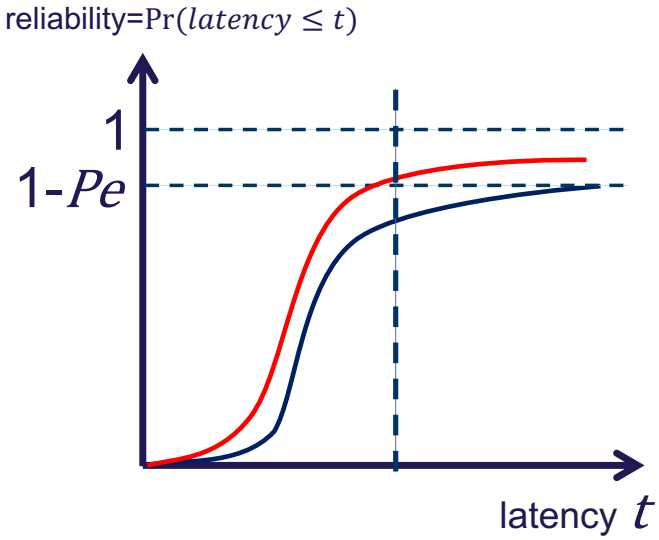
push-based communication

Shannon's model:
the transmitter determines what to send
and the receiver is always ready to receive



suitable to represent

- latency:
  packet already at the transmitter
- information freshness:
  past anchor in the physical world



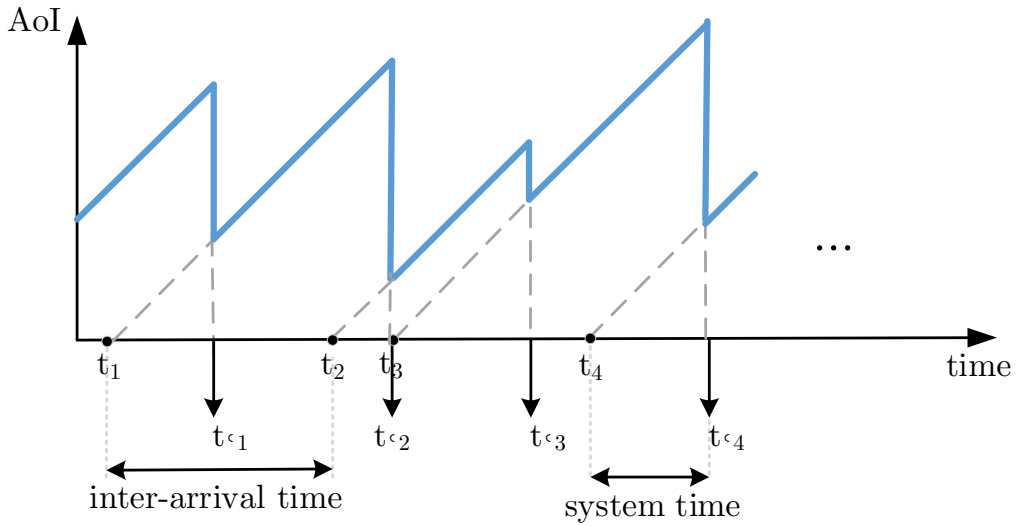
latency vs. age

- latency performance historically characterized with packet delays
- tracking applications and sense-compute-actuate cycles
  are not sensitive to packet delay,
  but to the freshness of the information at the receiver

example:
satellite-based tracking

Age of Information (AoI)

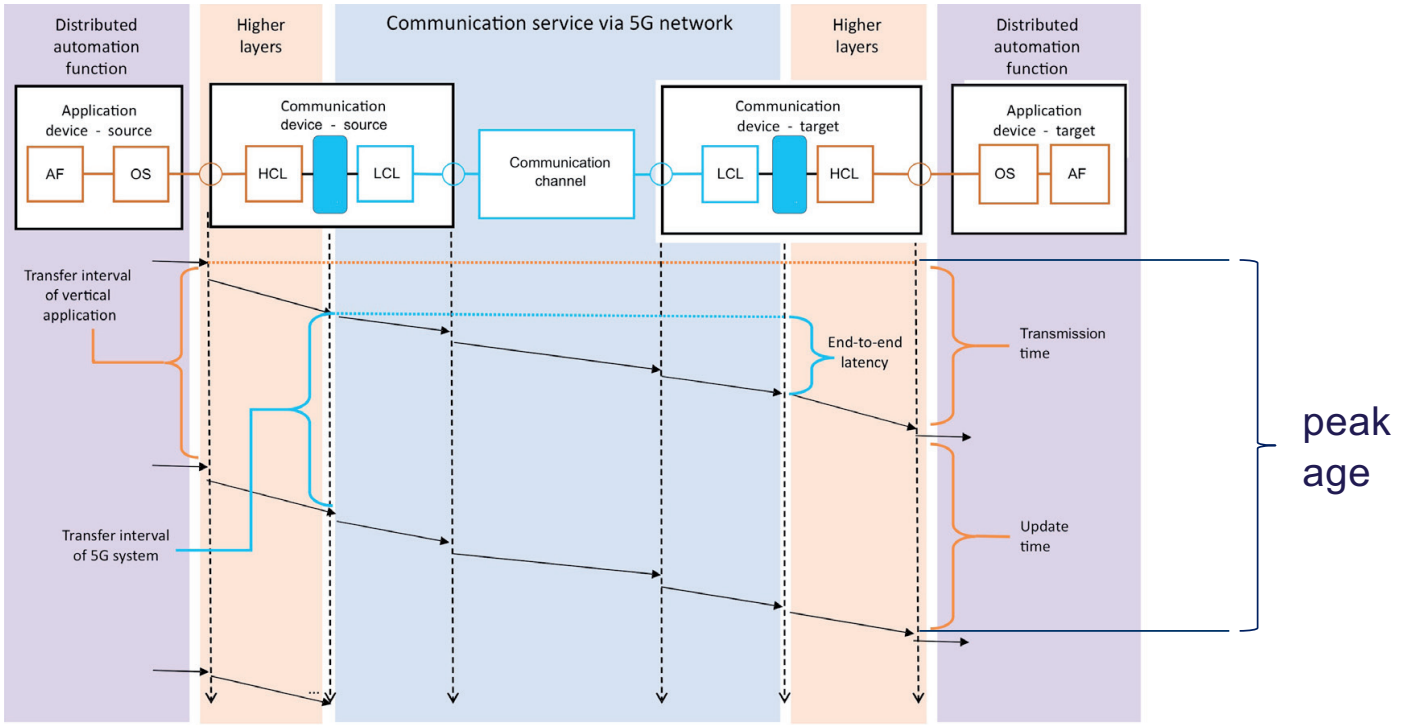
- Age of Information (AoI) and its byproducts
  are better metrics to capture the freshness of the information
- exogeneous vs. controlled sampling (generate-at-will)



[a] S. Kaul, R. Yates, and M. Gruteser, “Real-time status: How often should one update?” in *International Conference on Computer Communications*. IEEE, Mar. 2012, pp. 2731–2735.

[b] A. Kosta, N. Pappas, and V. Angelakis. "Age of information: A new concept, metric, and tool." Foundations and Trends in Networking 12.3 (2017): 162-259.

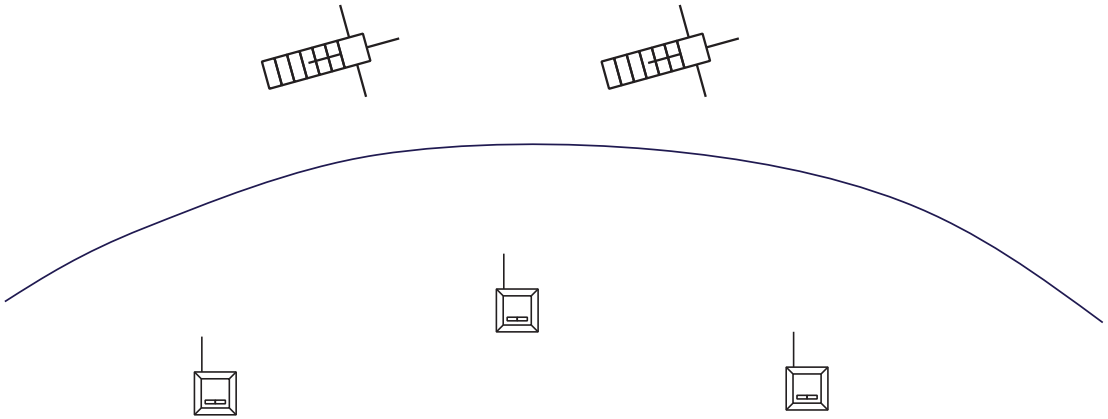
related considerations within 3GPP



3GPP Technical Specification 22.104. “Service requirements for cyber-physical control applications in vertical domains” https://www.3gpp.org/ftp/Specs/archive/22_series/22.104/

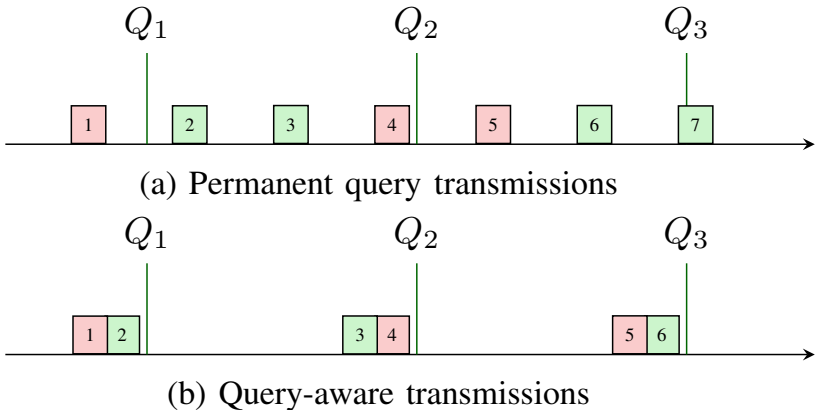
pull-based communication

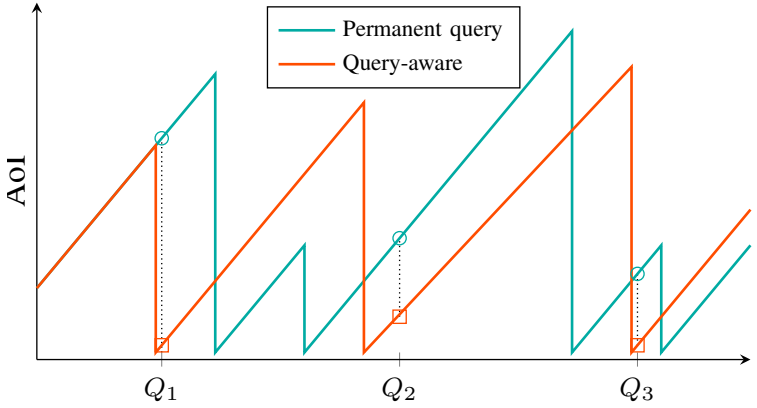
- in push-based models the receiver is *permanently subscribed* to the sender
- pull-based communication
  - reading process initiated by query
  - satellites, cloud-based queries to the edge devices, data fetching in a control loop



J. Holm, A. E. Kalør, F. Chiariotti, B. Soret, S. K. Jensen, T. B. Pedersen, and P. Popovski, "Freshness on Demand: Optimizing Age of Information for the Query Process", in Proc. IEEE ICC, Montreal, Canada (Virtual), June 2021.

Age of Information in pull-based model

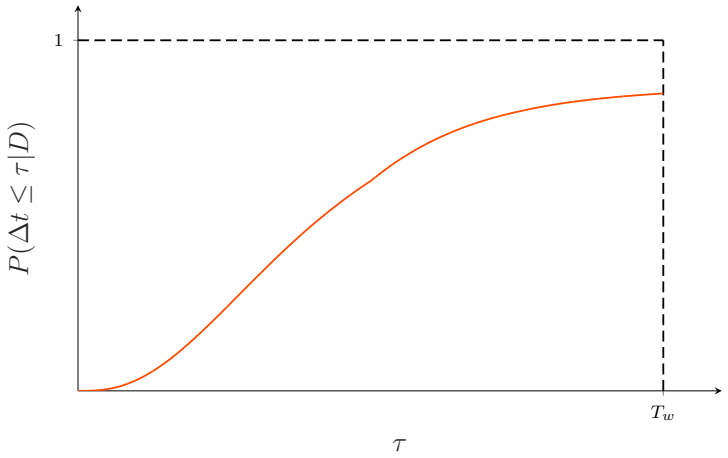




generalized view on timing: statistical operations in a time frame

- statistical **decision** in time
  - arrival of a packet

$$F_D(\tau) = \Pr(\Delta t \leq \tau | D)$$



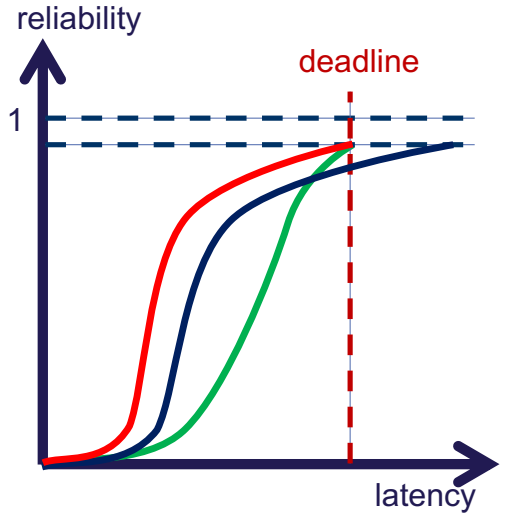
- statistical **estimation** in time
  - empirical risk of a machine learning algorithm

$$G_t(\tau) = \|\hat{x}_\tau(t) - x(t)\|$$

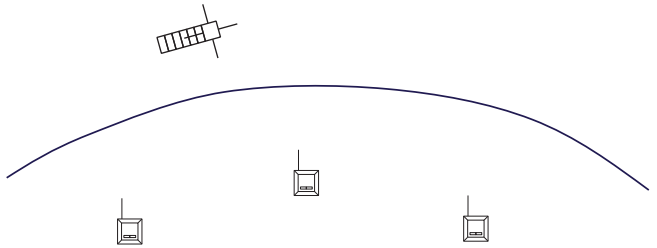
examples: deadline

statistical decision

- URLLC, judging whether the packet has arrived



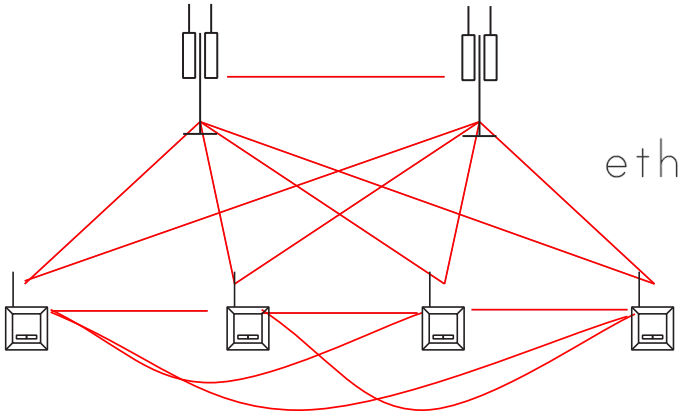
- statistical estimation
  satellite pass until link is unavailable



examples: relative timing reference

statistical decision

- consensus in distributed ledger



ethereum

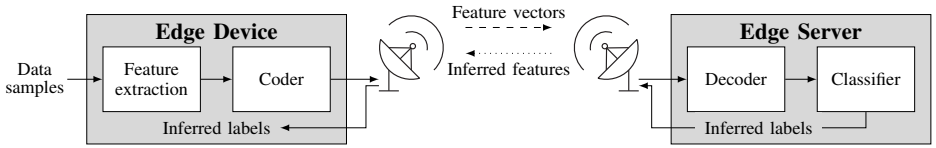
statistical estimation

- distributed learning and minimization of empirical risk

timing in remote/edge inference

- feature extraction and remote classification

- timing anchor
  how fresh is the conclusion about
  a physical phenomenon
  example: satellites detecting a pirate ship



- deadline
  make a decision before a deadline, for example at a vehicle crossing

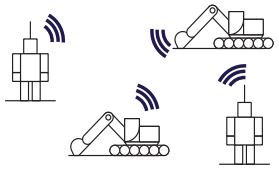
Q. Lan, Y. Du, P. Popovski, and K. Huang, "Capacity of Remote Classification Over Wireless Channels," in IEEE Transactions on Communications, accepted, 2021.

outline

6G and IoT connectivity

- massive access
- model for correlated access
- user identification in unsourced access
- massive downlink ACK
- **ultra-reliable low-latency communications**
- latency and general timing
- **how to guarantee reliability**
- URLLC in action

data-driven performance guarantees


Alice sells ultra-reliability to Bob.

many questions:

- how does Alice measure the reliability performance?
- under what conditions is ultra-reliability guaranteed?
- …

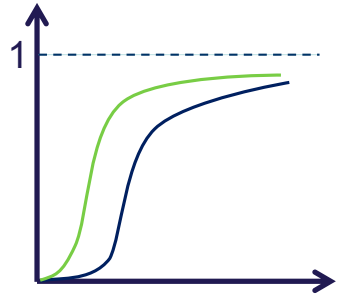
M. Angjelichinoski, K. F. Trillingsgaard and P. Popovski, "A Statistical Learning Approach to Ultra-Reliable Low Latency Communication," in IEEE Transactions on Communications, vol. 67, no. 7, pp. 5153-5166, July 2019

the inconvenient questions asked by Bob

Bob poses inconvenient questions
and wants to be convinced.



- what is the true distribution F used to calculate the reliability?
- what if the true channel distribution differs from F?
- what if we have no knowledge of F at all?



parameter selection for ultra-reliability

assume that the interference is absent.

we (somehow) know that the channel is Rayleigh.

the target error rate is ε_U, average SNR is $\bar{\gamma}_S$

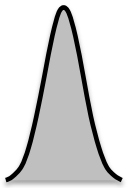
how do we choose the rate R?

$$\Pr(E) = \Pr(\log_2(1+\gamma_s) < R)$$

$$R = \log_2\left(1 + \bar{\gamma}_S \ln\left(\frac{1}{1-\varepsilon_U}\right)\right)$$

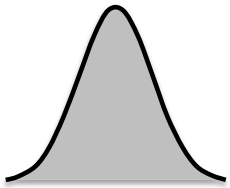
uncertainty in ultra-reliable wireless

inherent randomness of wireless



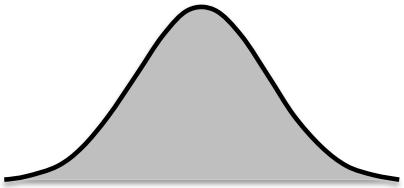
+

ignorance about channel statistics



+

changes of the channel statistics



ML is must-have for ultra-reliability

three key elements

- model selection
  - parametric models for F
  - non-parametric estimation of F
- learning
  - generate an estimate $\hat{F}$ using a training sample X^n
  - bad training data leads to a bad estimate $\hat{F}$
- performance evaluation
  - specification of the packet error rate only is insufficient

performance evaluation done right

learning can affect various decisions the system

- rate selection is the simplest example
- other examples:
  choice of resources in slicing, user admission, etc.

we define two types of reliability

Probably Correct Reliability (PCR)

Averaged Reliability (AR)

Probably Correct Reliability (PCR)

- inspired by PAC-learning
- calculated as a reliability that obtainable for a specific training sample X^n

$$\mathrm{P}(PER > \epsilon) \leq \xi$$

- suitable for a static environment where training precedes operation

Averaged Reliability (AR)

- calculated as a reliability that is averaged across all possible training samples $\{X^n\}$

$$\sup_{F \in \mathcal{F}} \overline{PER} \leq \epsilon$$

- suitable for dynamic environments, where separate training is not feasible

parametric vs. non-parametric

- if parametric estimation is used
  then we need to have a good guess about the channel model

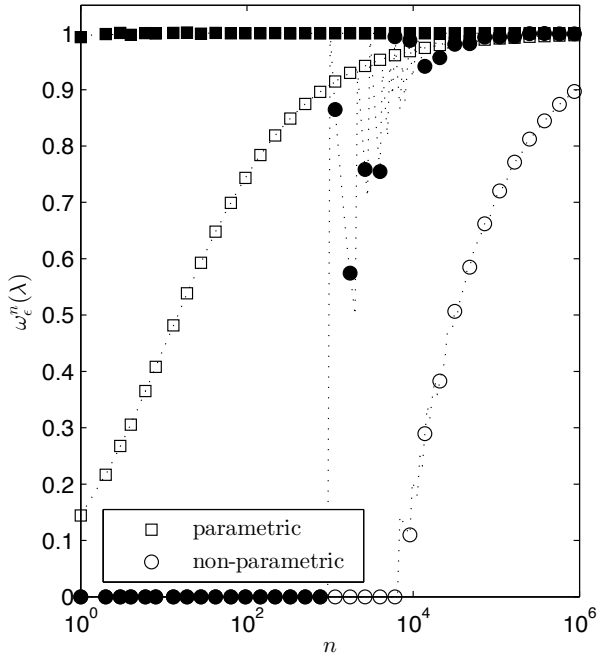
- non-parametric estimation does not assume any distribution
  - number of required training samples overwhelming,
    in the order of $1/\epsilon$, which can go beyond 10^9!

P. C. F. Eggers, M. Angjelichinoski and P. Popovski, "Wireless Channel Modeling Perspectives for Ultra-Reliable Communications," in IEEE Transactions on Wireless Communications, vol. 18, no. 4, pp. 2229-2243, April 2019

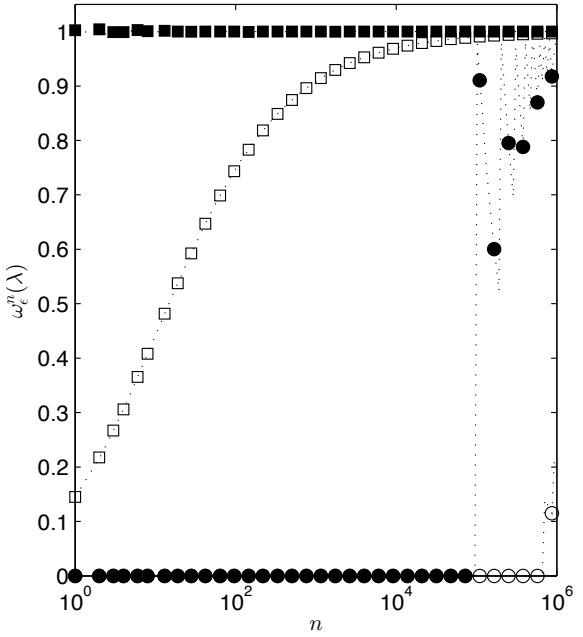
example performance evaluation

- exploring the correct 10-3 rate vs. exploiting for data transmission

conservative performance while learning

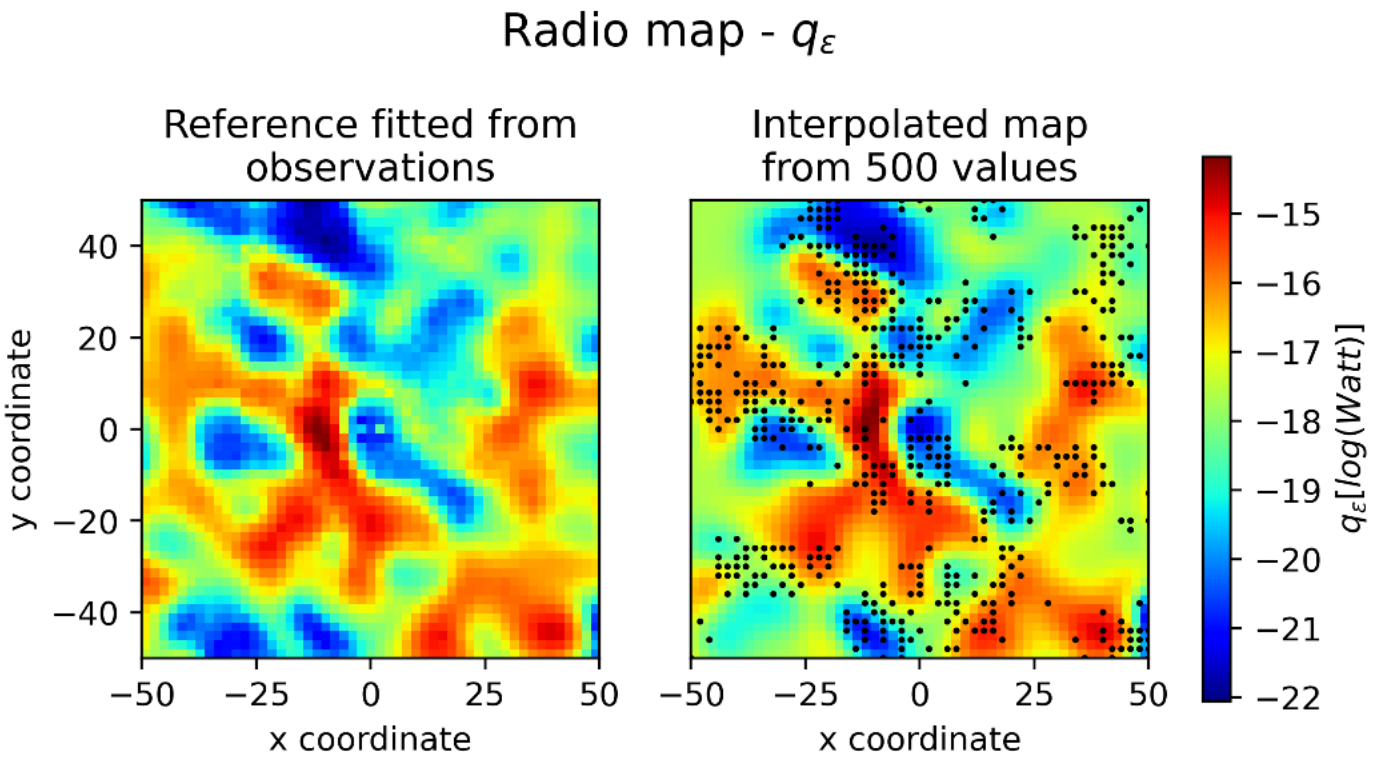


(a) $\epsilon = 10^{-3}, \xi = 10^{-3}$



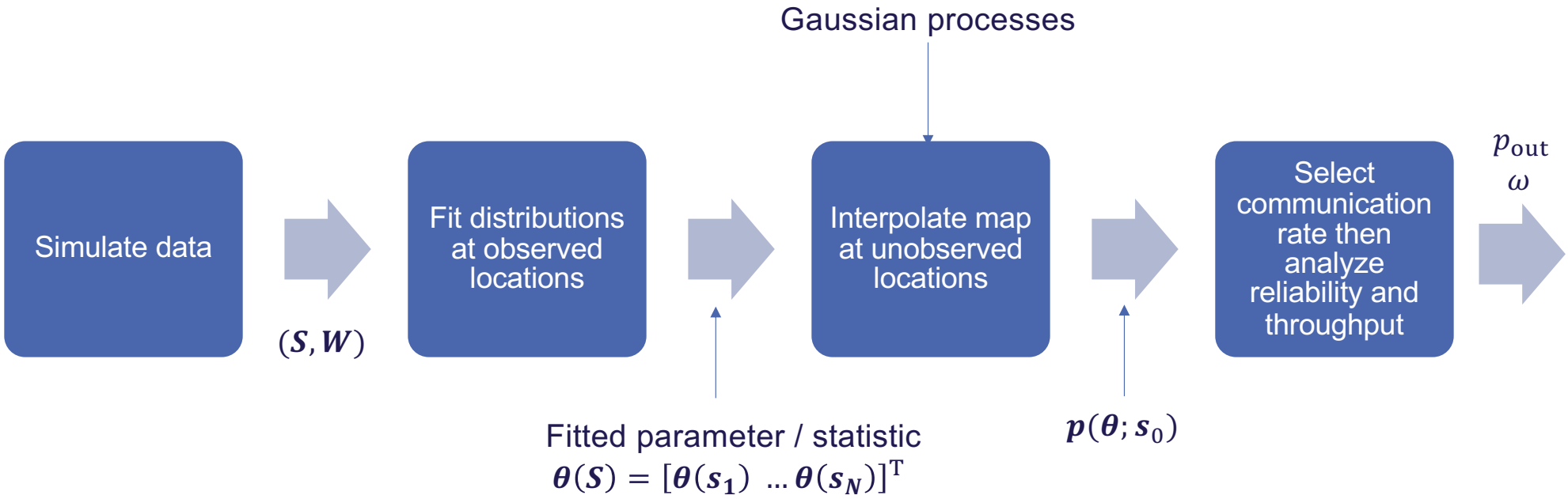
(c) $\epsilon = 10^{-5}, \xi = 10^{-3}$

radio maps for ultra-reliability guarantees

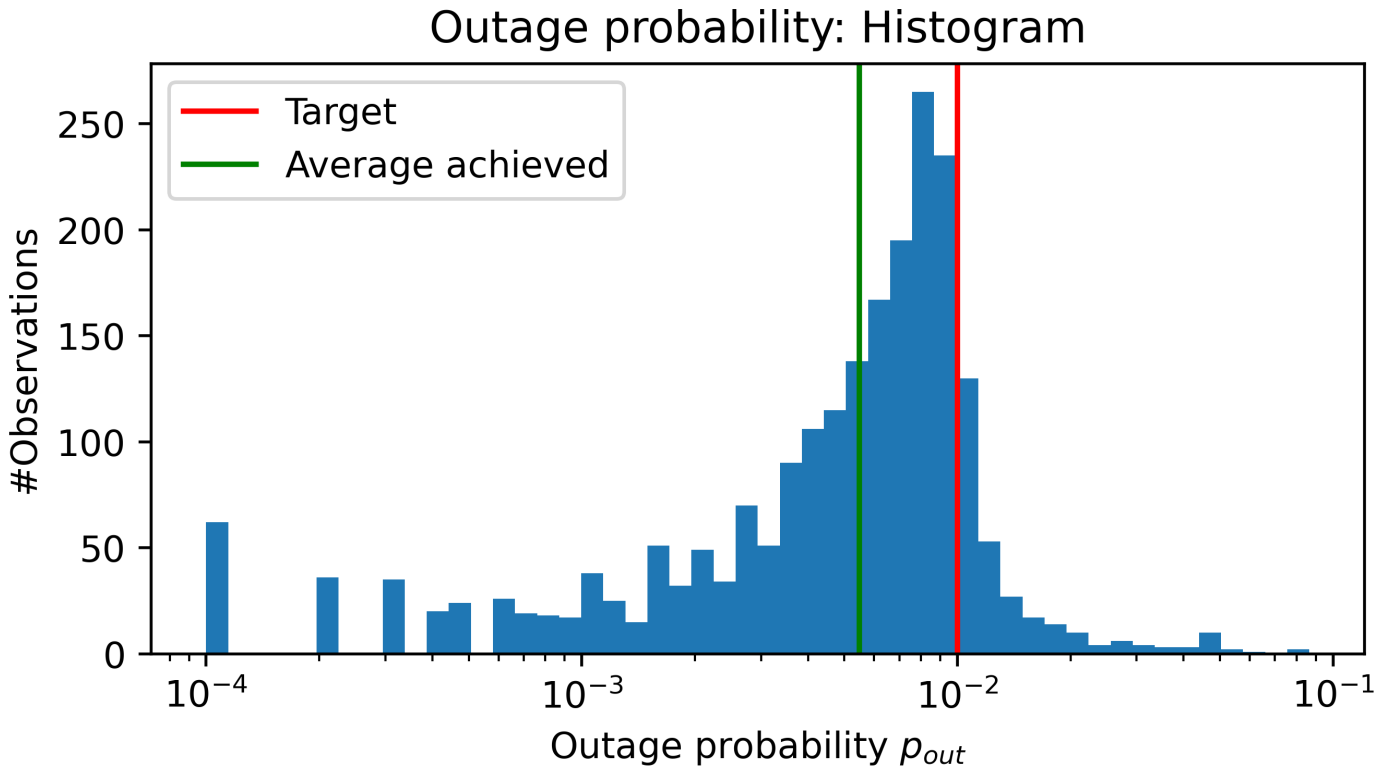


how to predict the channel statistics anywhere
based on a set of discrete spatial points where measurements have been taken

radio maps for ultra-reliability guarantees



example result averaged across space



outline

6G and IoT connectivity

massive access

model for correlated access

user identification
in unsourced access

massive downlink ACK

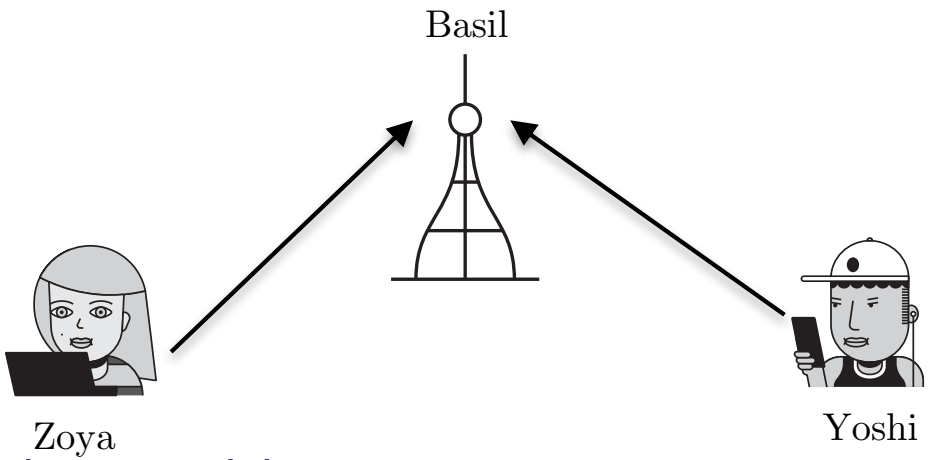
ultra-reliable low-latency communications

latency and general timing

how to guarantee reliability

URLLC in action

slicing the spectrum



broadband connectivity

continuous transmission

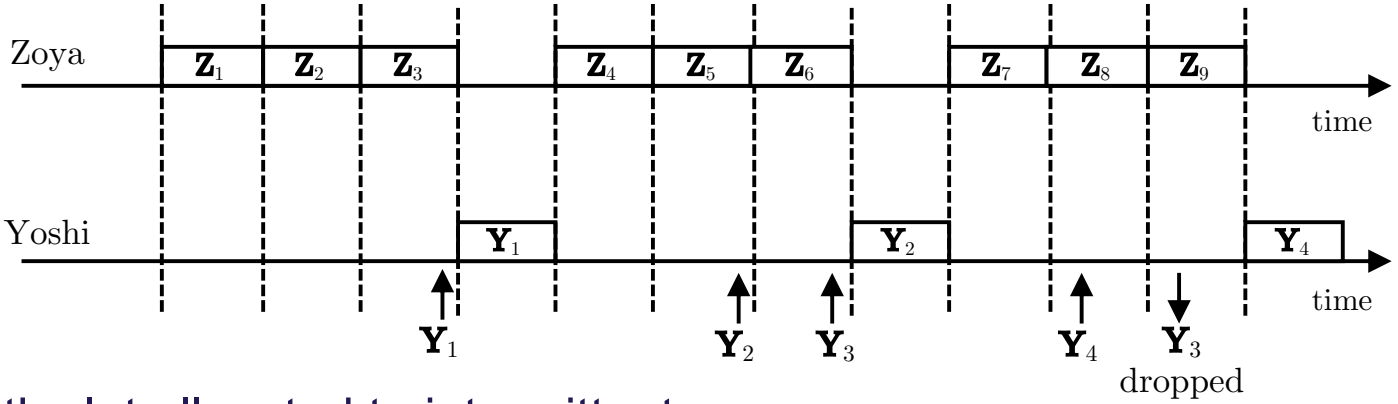
example: video stream

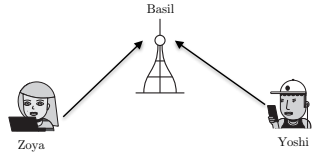
target rate

- low-latency connectivity
- intermittent transmission
- example: reliable control
- target latency/reliability

P. Popovski, *Wireless Connectivity: An Intuitive and Fundamental Guide,* Chapter 4, Wiley, 2020.

toy example: orthogonal slicing





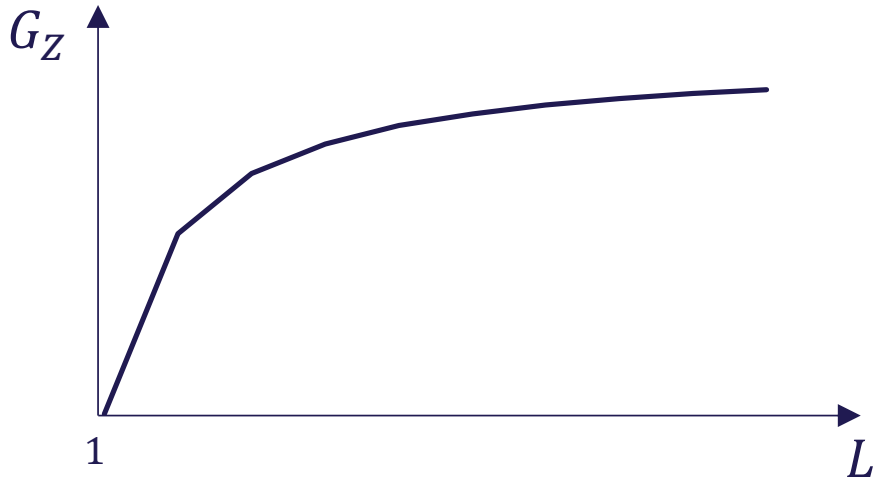
L-th slot allocated to intermittent

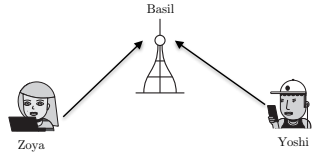
- he waits at most L slots to deliver the packet

the broadband goodput is $G_Z = \frac{L-1}{L} R$

intermittency does not affect the broadband goodput

toy example: orthogonal slicing

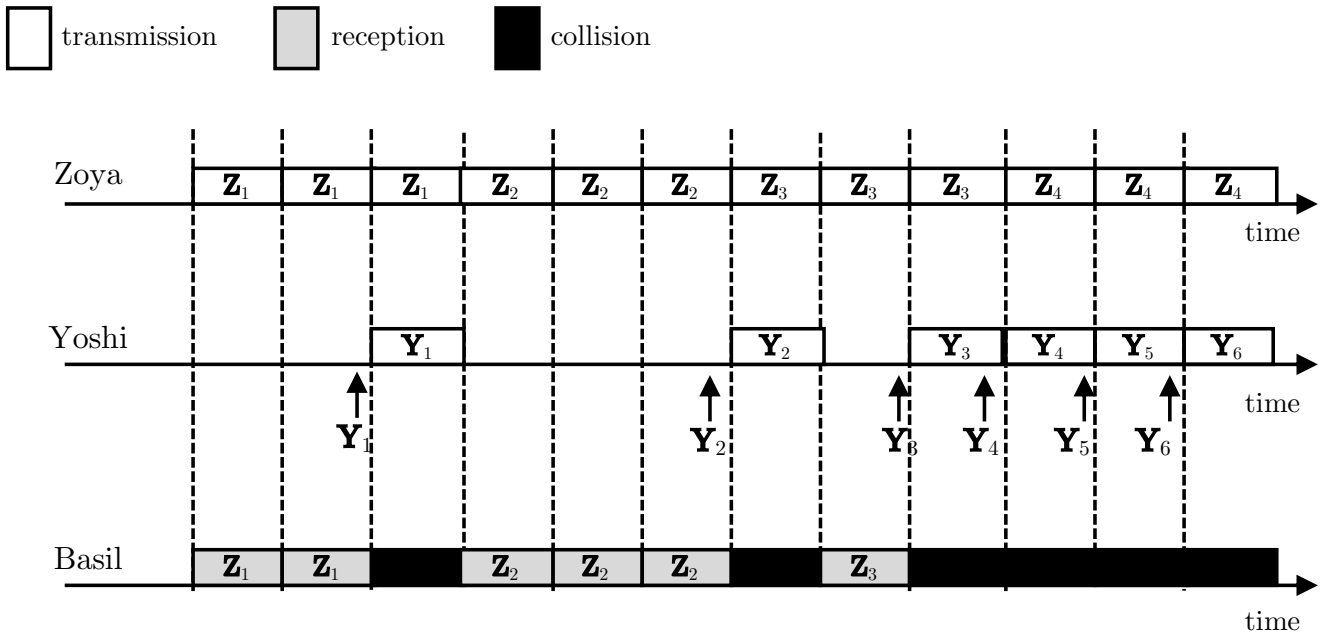


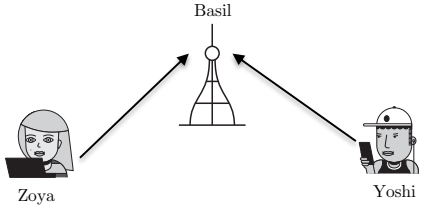


if Yoshi wants instantaneous transmission, then the goodput of Zoya is strictly 0.

what if the broadband connectivity can live with some error probability?

toy example: non-orthogonal slicing

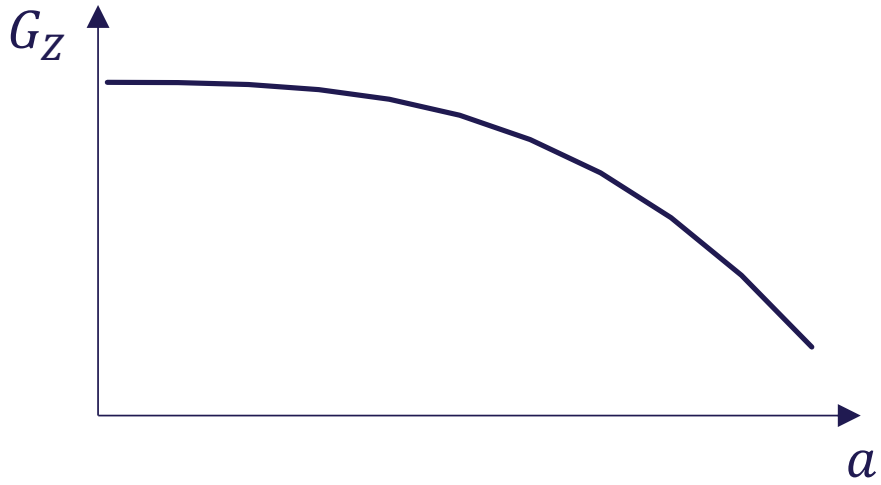


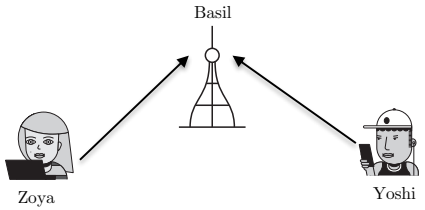


broadband repeats each packet 3 times

Successive Interference Cancellation

toy example: non-orthogonal slicing

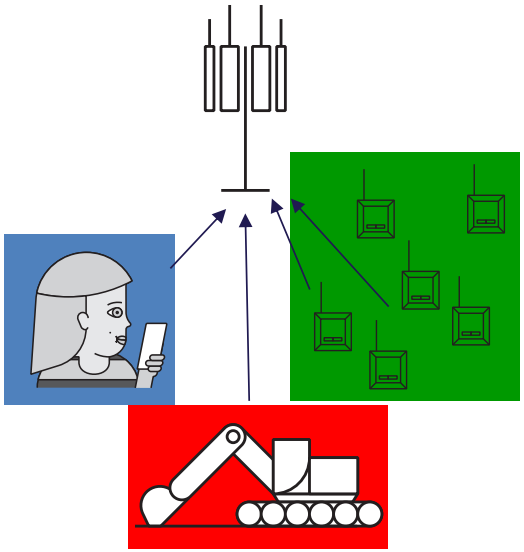


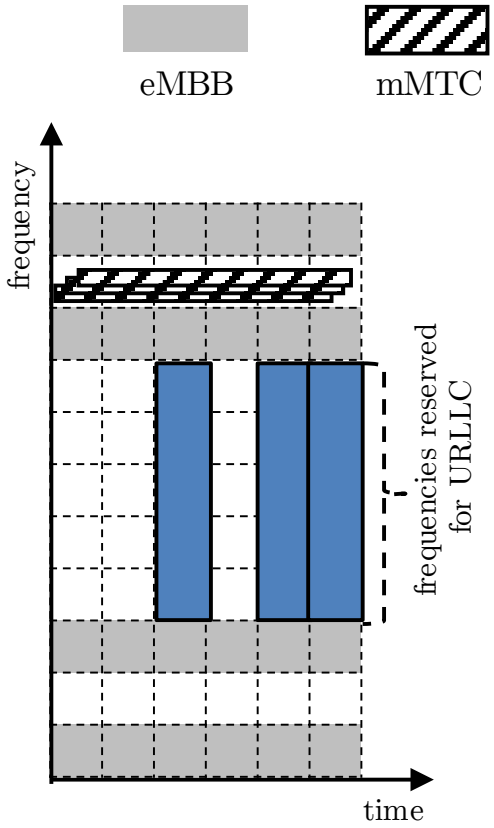


error only if three repetitions are lost

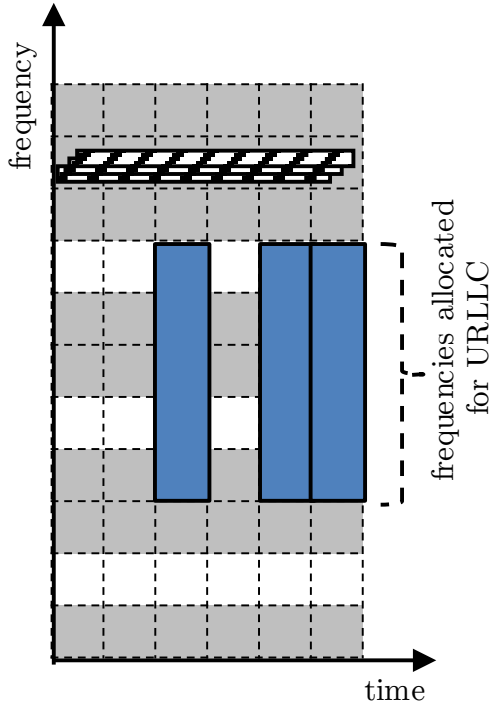
goodput of Zoya: $G_Z = \frac{R}{L}(1 - a^L)$

two types of spectrum slicing





orthogonal



non-orthogonal

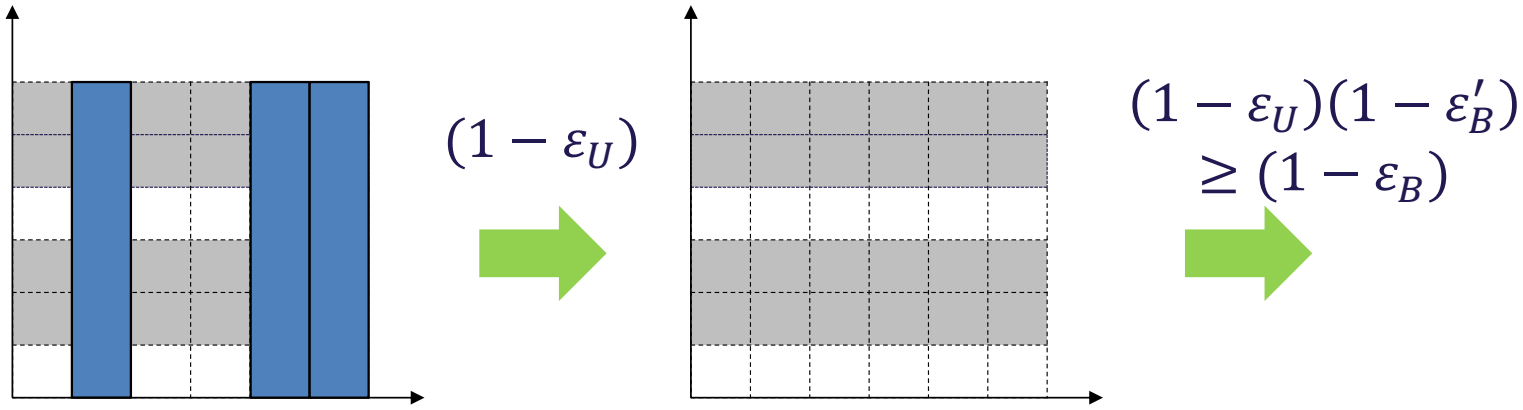
P. Popovski, K. F. Trillingsgaard, O. Simeone and G. Durisi, "5G Wireless Network Slicing for eMBB, URLLC, and mMTC: A Communication-Theoretic View," in IEEE Access, vol. 6, pp. 55765-55779, 2018

the reliability diversity

$$\varepsilon_U \ll \varepsilon_B \ll \varepsilon_M$$

design that benefit from heterogeneous reliability requirements

example of interfering eMBB and URLLC:

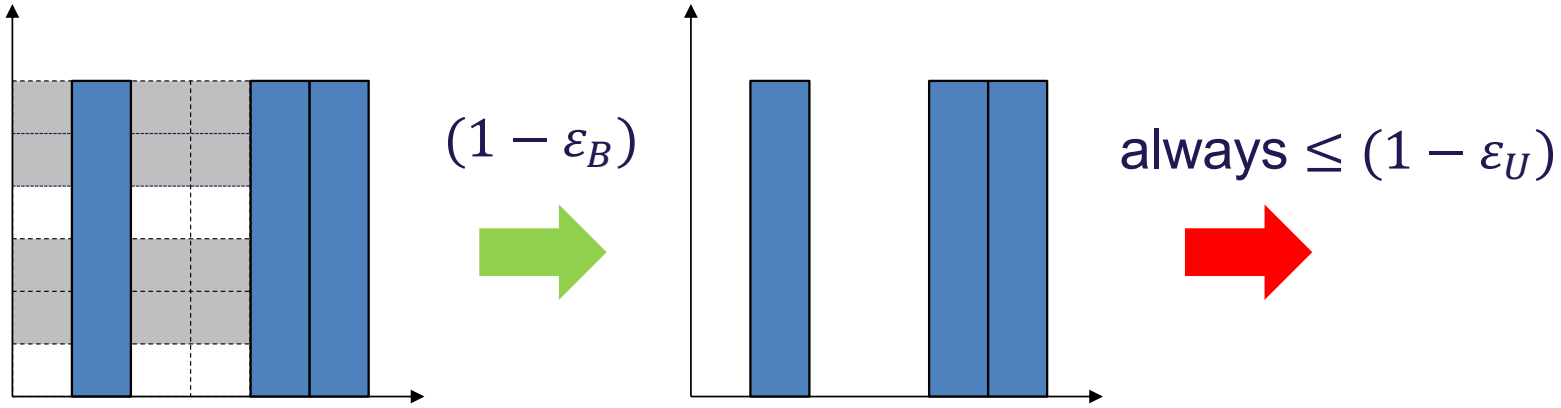


the reliability diversity

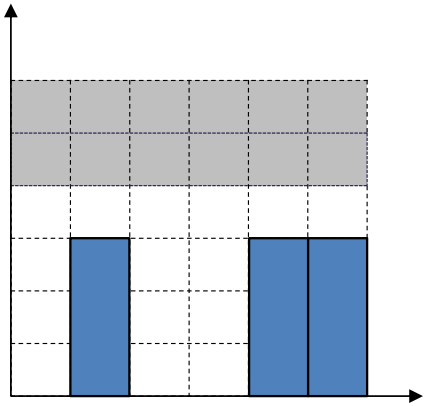
$$\varepsilon_U \ll \varepsilon_B \ll \varepsilon_M$$

design that benefit from heterogeneous reliability requirements

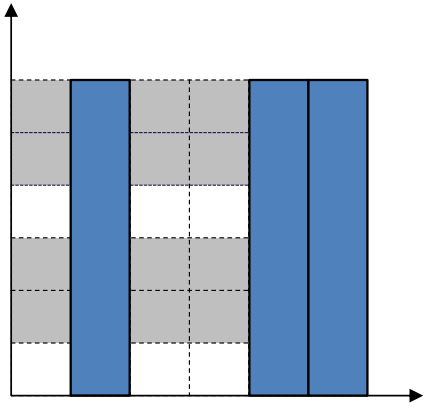
example of interfering eMBB and URLLC:



slicing for eMBB and URLLC



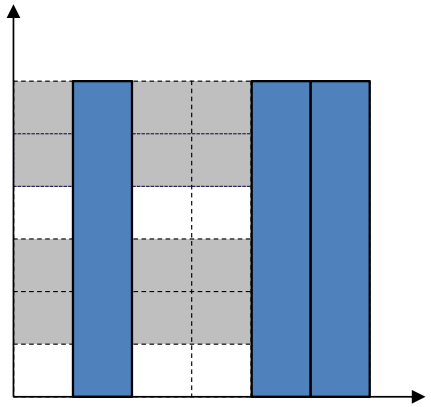
orthogonal



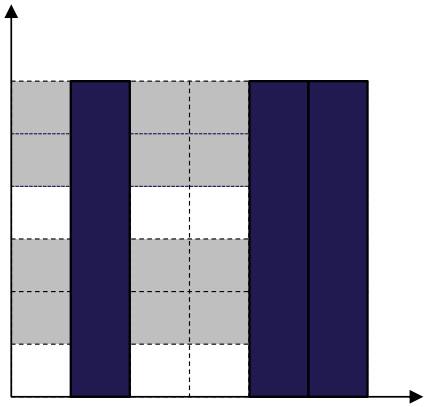
non-orthogonal
with SIC

$$\Pr\left(\frac{1}{F_U}\sum_{f=1}^{F_U}\log_2\left(1+\frac{G_{U,f}}{1+G_{B,f}^{\mathrm{tar}}}\right)<r_U\right)\leq\varepsilon_U$$

slicing for eMBB and URLLC



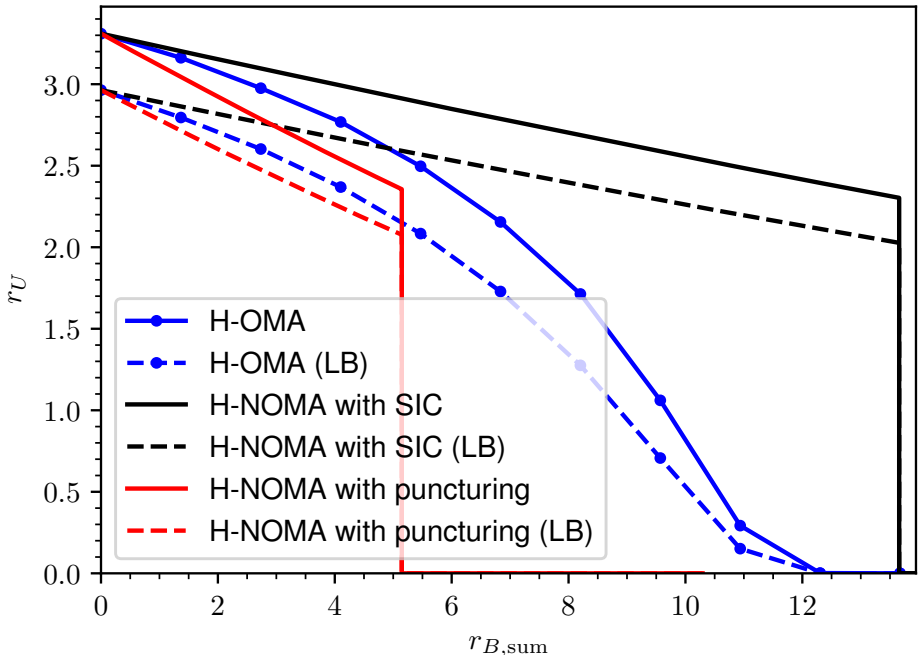
non-orthogonal
with puncturing



perspective of the
eMBB

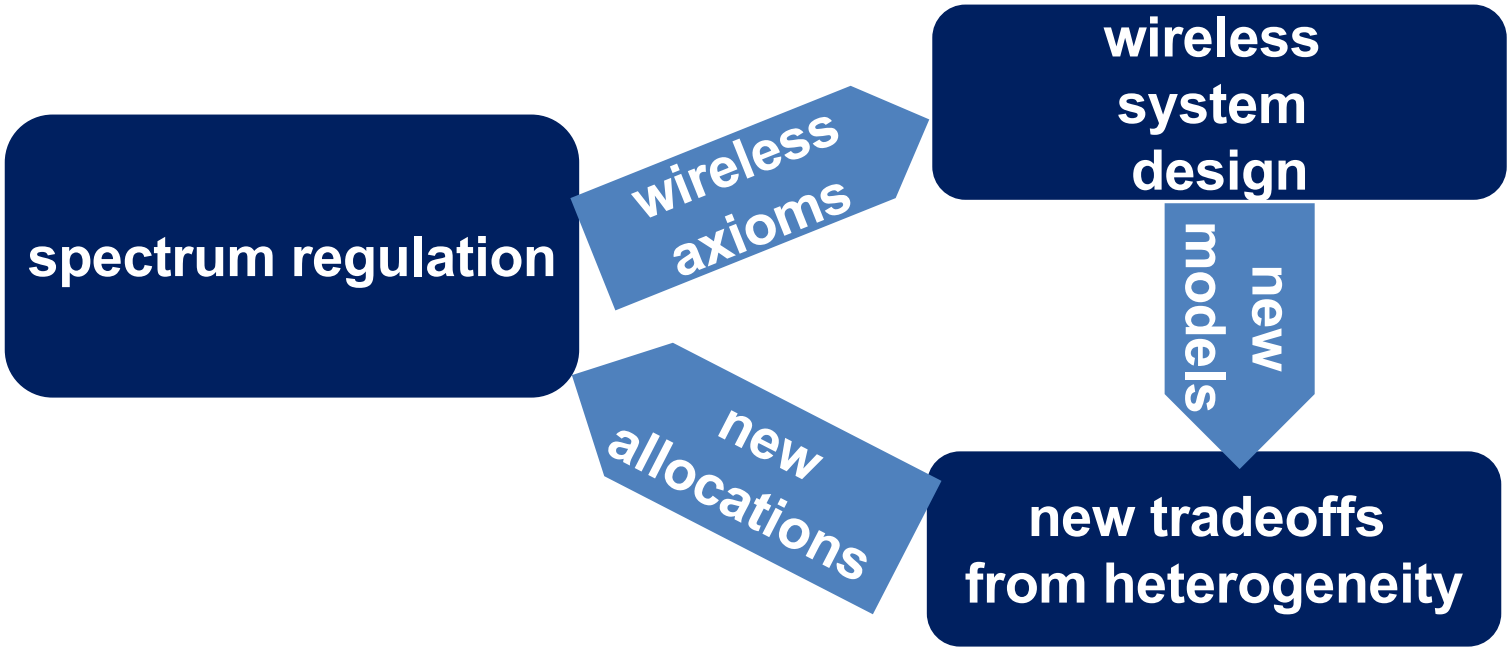
eMBB uses erasure code of rate $1 - \frac{k}{S}$
and thus has a decreased rate

example results



H-NOMA: heterogeneous NOMA

long-term spectrum considerations



generic access mechanisms

scheduled access

- no or low uncertainty about the activity

multiple access: long-term resource allocation

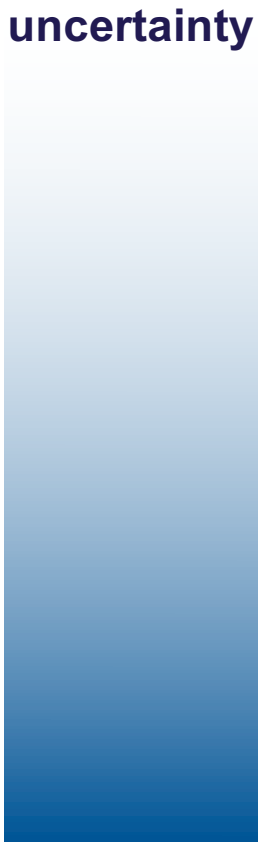
grant-based access: Access request followed by allocation

grant-free access

- transmit data without prior signaling
- no need for access request but uncertainty of activity is high

random access

- no prior information about the user activity
- the BS wants to determine the set of active users
- objective is to find appropriate activation patterns



error sources in access protocols

Orthogonal multiple access

Mostly deterministic activation and no per-packet scheduling is needed

Reliability depends on synchronization and channel error probabilities

$$\epsilon_{\mathrm{OMA}} = 1 - (1 - \epsilon_{\mathrm{sync}})(1 - \epsilon_{\mathrm{data}})(1 - \epsilon_{\mathrm{ack}})$$

Grant-based access

The activity of the users is not known a priori; requires reservation

$$\epsilon_{\mathrm{GB}} = 1 - (1 - \epsilon_{\mathrm{sync}})(1 - \epsilon_{\mathrm{req}})(1 - \epsilon_{\mathrm{grant}})(1 - \epsilon_{\mathrm{data}})(1 - \epsilon_{\mathrm{ack}})$$

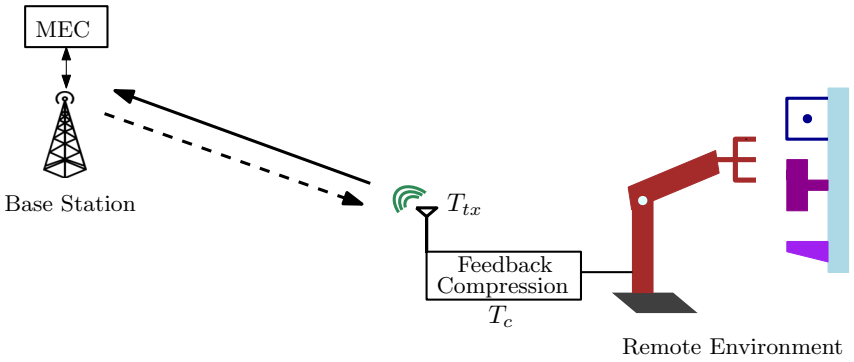
Grant-free access

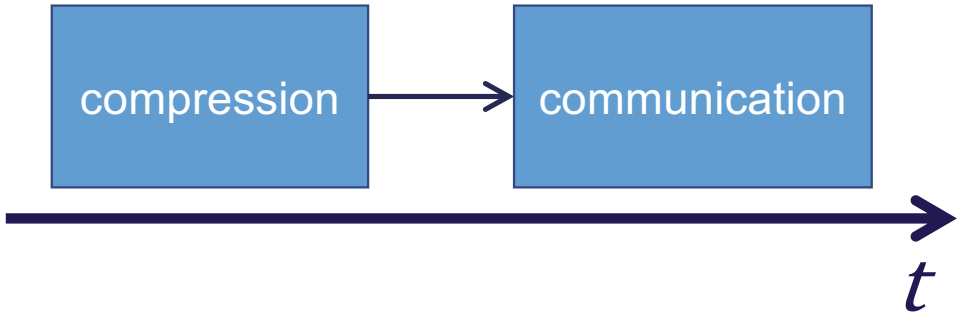
No reservation required, just transmit

$$\epsilon_{\mathrm{GF}} = 1 - (1 - \epsilon_{\mathrm{sync}})(1 - \epsilon_{\mathrm{data}})(1 - \epsilon_{\mathrm{ack}})$$

P. Popovski et al., "Wireless Access in Ultra-Reliable Low-Latency Communication (URLLC)," in IEEE Transactions on Communications, vol. 67, no. 8, pp. 5783-5801, Aug. 2019

tradeoff compression vs. communication





latency budget can be spent on

- compression, to lower the data size
- transmission, to make a more reliable transmission

the latency of compression depends on the

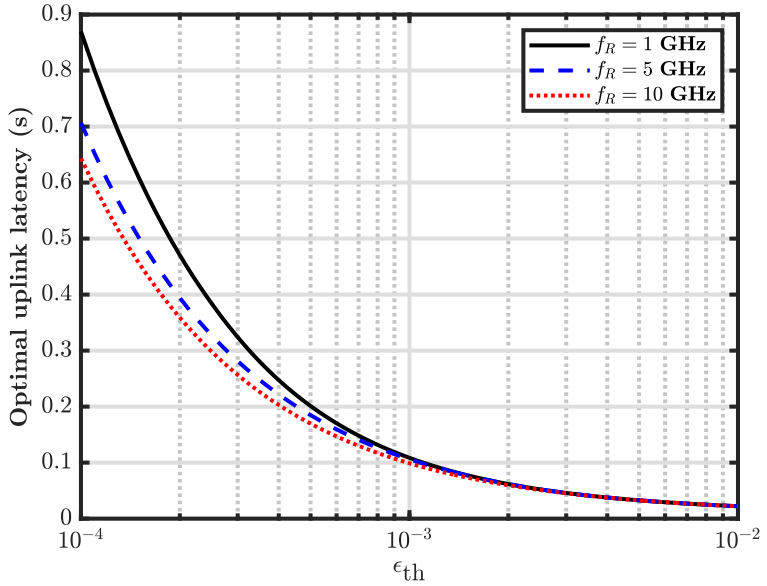
- compression ratio
- processing capability of the device

S. Suman, Č. Stefanović, S. Došen, and P. Popovski, "Analysis and Optimization of the Latency Budget in Wireless Systems with Mobile Edge Computing", in Proc. IEEE ICC, Seoul, Korea, May 2022.

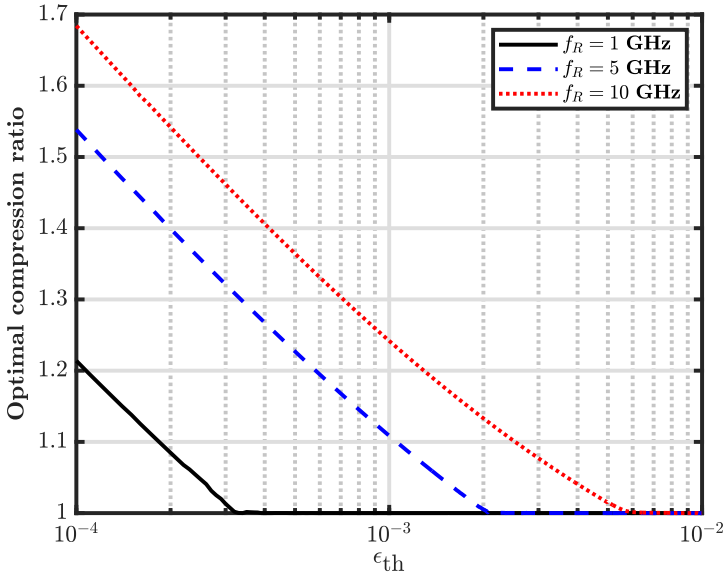
tradeoff compression vs. communication

two types of optimization problems

- constraint the outage and minimize the time
- set the deadline and minimize the outage



minimized latency



optimal compression ratio

conclusions and outlook

we need to take a critical look on the role of IoT in 6G
and type of problems that will need to be addressed

massive access problem to be enriched by:

user correlation, alternative packet structure, downlink

there is a tradeoff between reliability and latency,
the rigit requirements on the wireless link can be softened by
system-level considerations

a number of exciting problems awaiting at the intersection of
IoT, time, space, intelligence, and value