REDISCOVERING ALOHA FOR LATENCY-CRITICAL SERVICES:

THE BLIND AND THE FAR-SIGHTED

Salah El Ayoubi

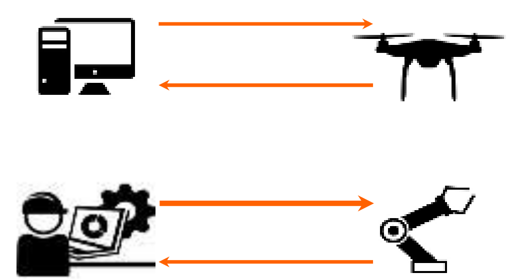
Professor at CentraleSupelec – Université Paris Saclay

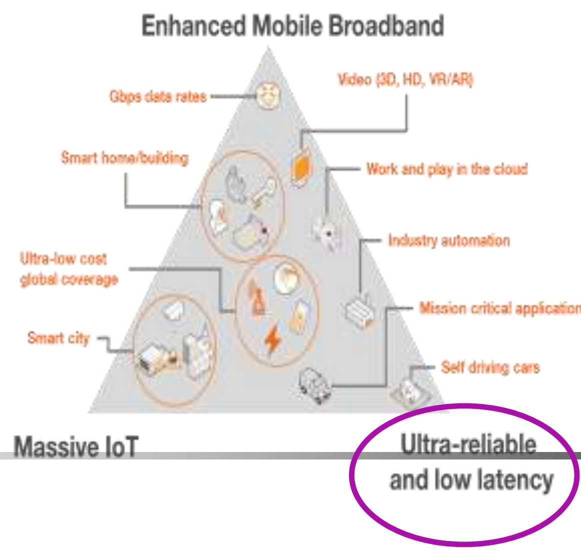
Researcher at CNRS, Laboratory of Signals and Systems

salaheddine.elayoubi@centralesupelec.fr

Industrial IoT and URLLC

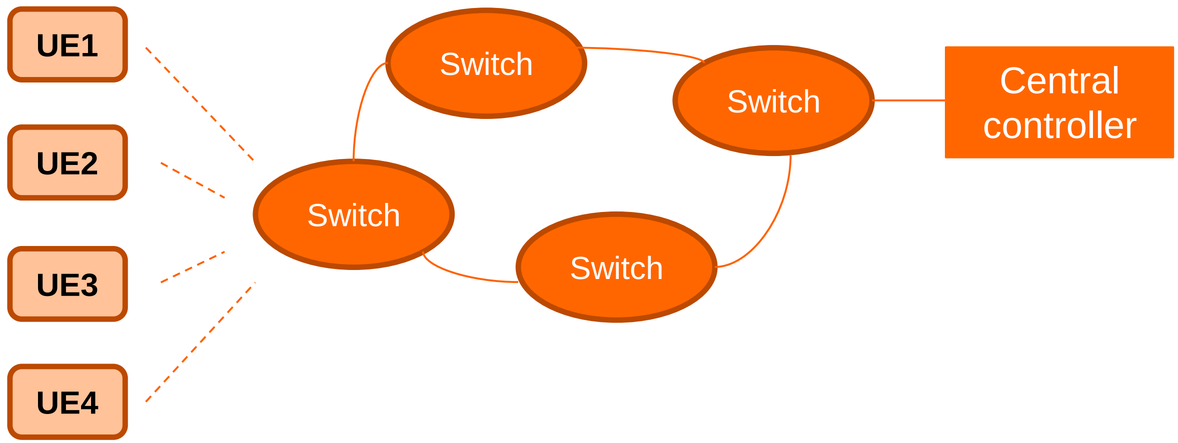
- **IoT networks allow more than low rate sensor connectivity**
- **Any application requiring reliability and resilience belongs to the IIoT (Industrial IoT):**
  - Communications between machines in a factory
  - Tele-operation of drones and machines
  - Aeronautical applications
- **5G networks intend to serve IIoT:**
  - Ultra Reliable Low Latency Communications (URLLC) service





Scenario: Removing wires between machines in a factory

- **Machines (UEs) communicate wirelessly with a central controller via a set of switches**



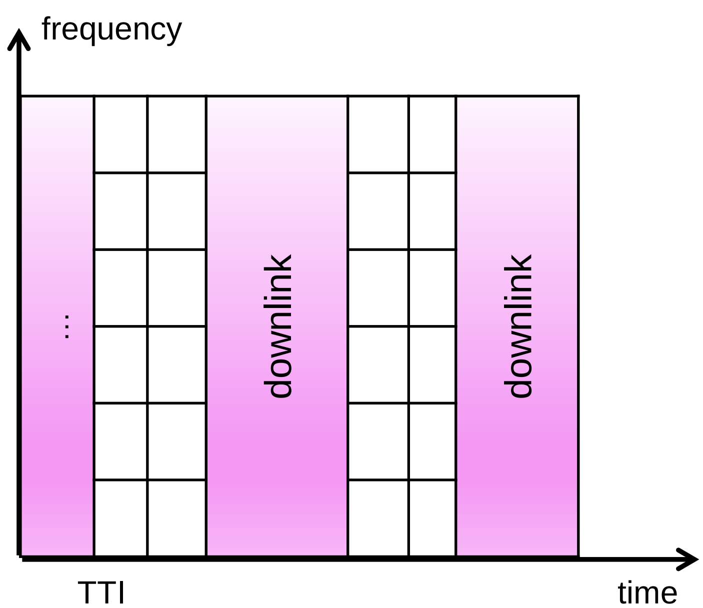
- **Each machine generates sporadically packets of fixed size**
- **Objective is to ensure that**
  - the **proportion** of packets,
  - **correctly received** by the controller
  - within the **delay budget** (e.g. 1 ms)
  - is larger than a **reliabilty target** (e.g. loss probability < 10^{-5})

Outline

- **Introduction to resource allocation for critical IoT**
- **The blind**
- **The far-sighted**
- **Perspectives**

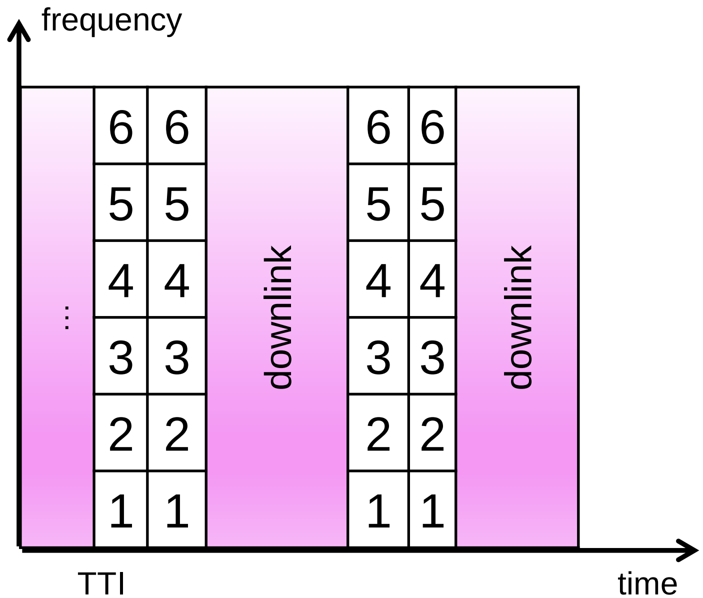
Deterministic resource reservation for URLLC traffic

- **How reliability targets are classically achieved?**
  - **reserve resources** for each UE.
  - called in 5G: semi-persistent scheduling
- **Illustration for the 5G frame in the 3500 MHz band**
  - DDUU configuration: 2 slots for uplink and 2 slots for downlink. 1 slot=0.144ms
  - for a 1ms delay target, the packet has to be received within 4 slots (as there is 1 slot for alignment and 1 slot for processing)
  - Packets of 32 bytes
  - QPSK ½ modulation (1 bit/symbol)
  - 1 packet occupies 8 subcarriers=240 KHz



Deterministic resource reservation for URLLC traffic

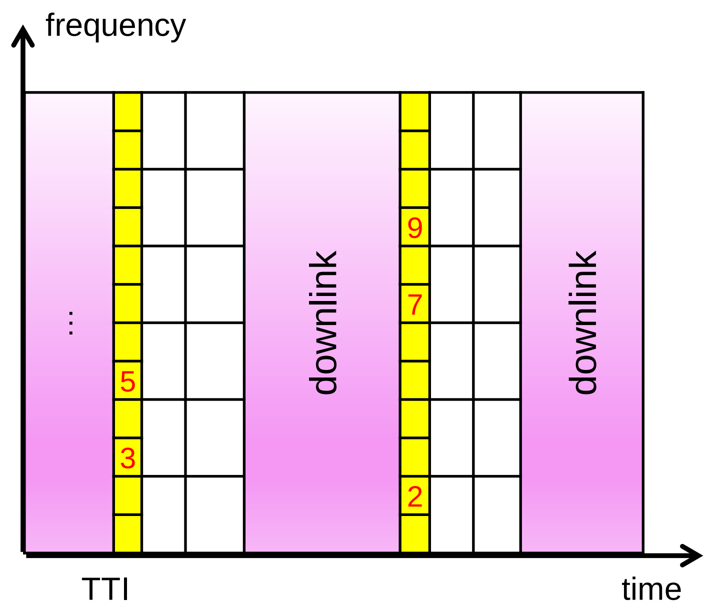
- **How reliability targets are classically achieved?**
  - **reserve resources** for each UE.
  - called in 5G: semi-persistent scheduling
- **Illustration for the 5G frame in the 3500 MHz band**
  - DDUU configuration: 2 slots for uplink and 2 slots for downlink. 1 slot=0.144ms
  - for a 1ms delay target, the packet has to be received within 4 slots (as there is 1 slot for alignment and 1 slot for processing)
  - Packets of 32 bytes
  - QPSK ½ modulation (1 bit/symbol)
  - 1 packet occupies 8 subcarriers=240 KHz



- **Resources to be reserved each slot**
  - packets may be generated at any slot
- **Individual reservation is suboptimal:**
  - large number of users and sporadic traffic
  - 6 users, deterministic traffic: need 1.44 MHz
  - 60 users, each generates a packet per slot with probability *q=0.1*: need 14.4 MHz

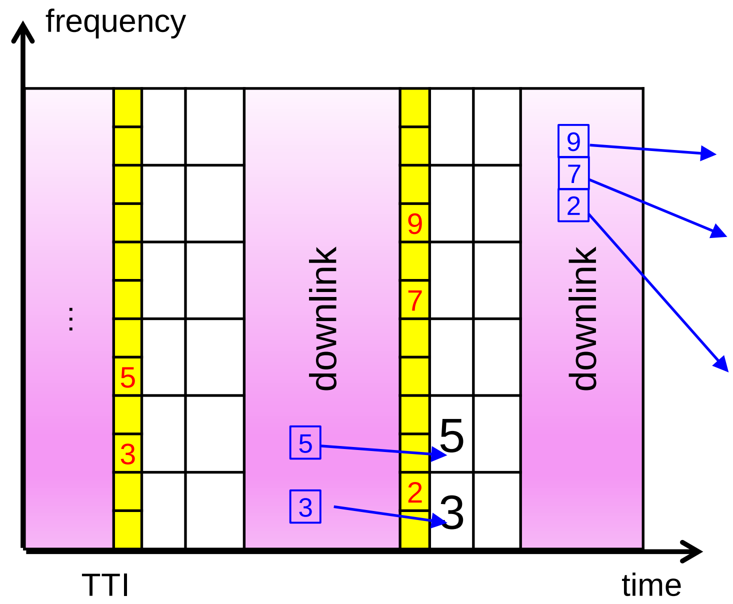
Classical solution: grant-based communication

- **How this problem of sporadicity is classically solved?**
  - when a packet is generated, the user issues a **scheduling request** in the next slot.
  - requests are small and sent on dedicated resources
  - the base station decodes the request, and sends back a **scheduling grant**
  - the user uses the reserved resource for sending its packet
  - called in 5G: grant-based scheduling



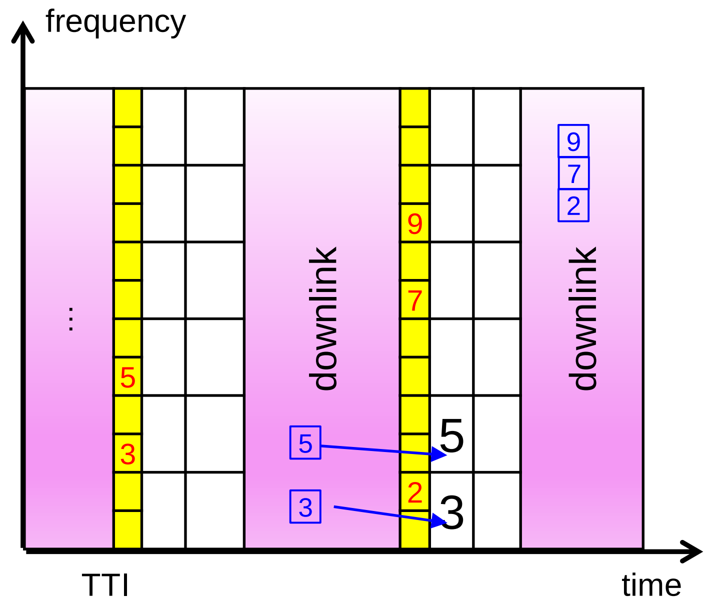
Classical solution: grant-based communication

- **How this problem of sporadicity is classically solved?**
  - when a packet is generated, the user issues a **scheduling request** in the next slot.
  - requests are small and sent on dedicated resources
  - the base station decodes the request, and sends back a **scheduling grant**
  - the user uses the reserved resource for sending its packet
  - called in 5G: grant-based scheduling



Classical solution: grant-based communication

- **How this problem of sporadicity is classically solved?**
  - when a packet is generated, the user issues a **scheduling request** in the next slot.
  - requests are small and sent on dedicated resources
  - the base station decodes the request, and sends back a **scheduling grant**
  - the user uses the reserved resource for sending its packet
  - called in 5G: grant-based scheduling
- **Grant-based scheduling is not adequate for URLLC:**
  - 1 slot for alignment
  - 1 slot for transmission of the request
  - 1 slot for receiving the grant
  - 1 slot for transmission
  - and the budget of 4 slots expires
  - no time for processing…

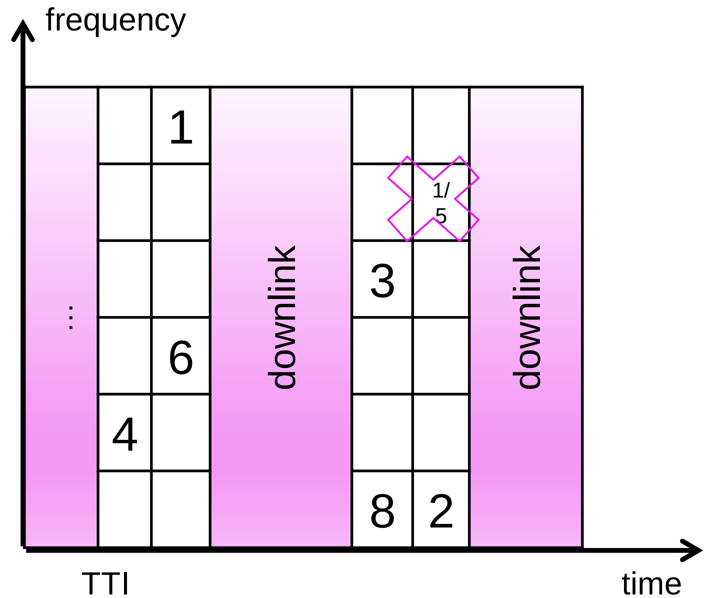


Back to the old contention-based access: Aloha

- **Scenario: large number of users, sporadic traffic**
  - probability of having a packet in a slot: q<1
- **No resource reservation per user, but a pool of reserved resources**
  - each active UE selects a resource at random (ALOHA-like)
- **Issue: low reliability:**
  - collisions between packets
- **condition of success:**
  - no one chooses the same resource
- **probability of loss:**

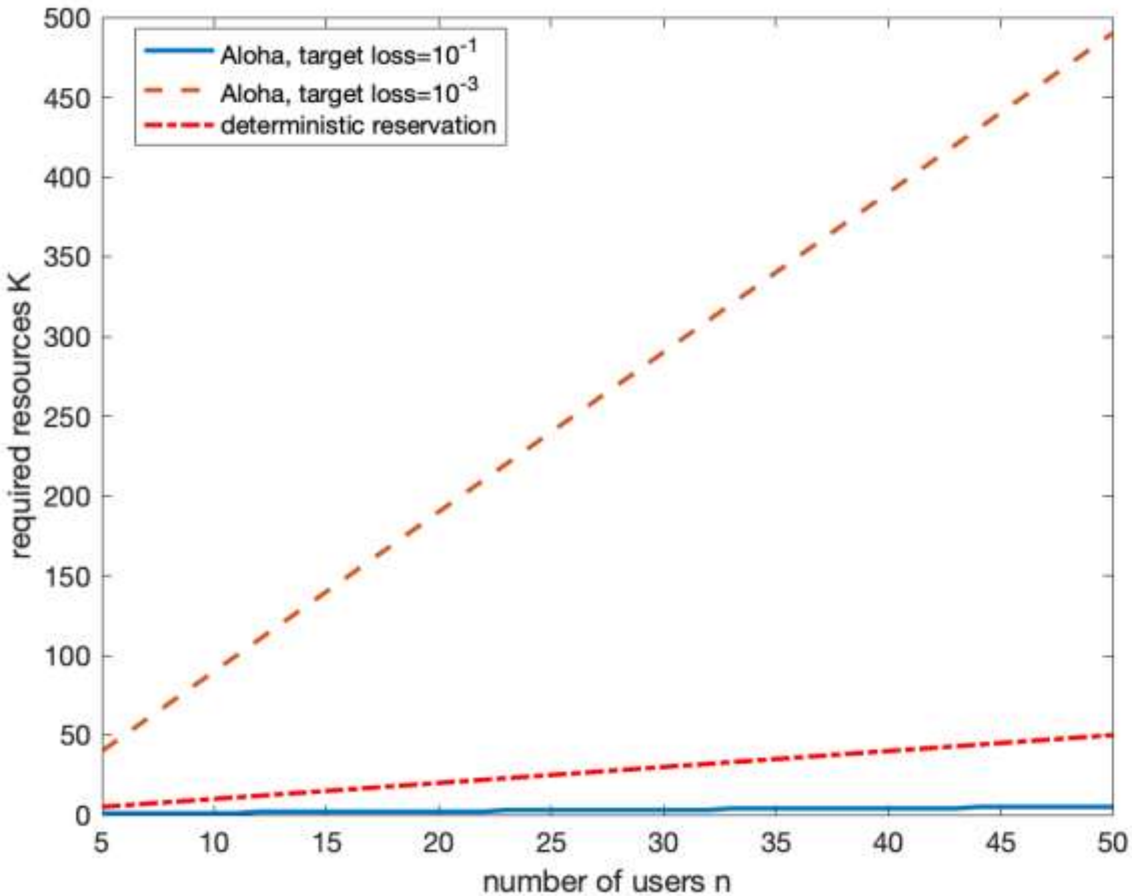
$$loss = 1 - \left(1 - \frac{q}{K}\right)^{n-1}$$

  - n users in total
  - each user active with probability q<1
  - K resources in total



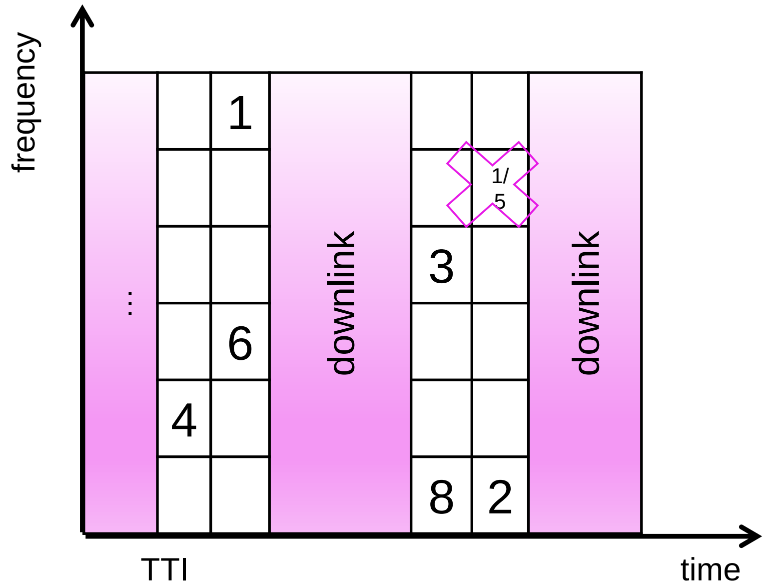
Performance of Aloha: unacceptable

- **Aloha brings a gain when the service is tolerant to loss**
- **High reliability is not achievable with simple Aloha**

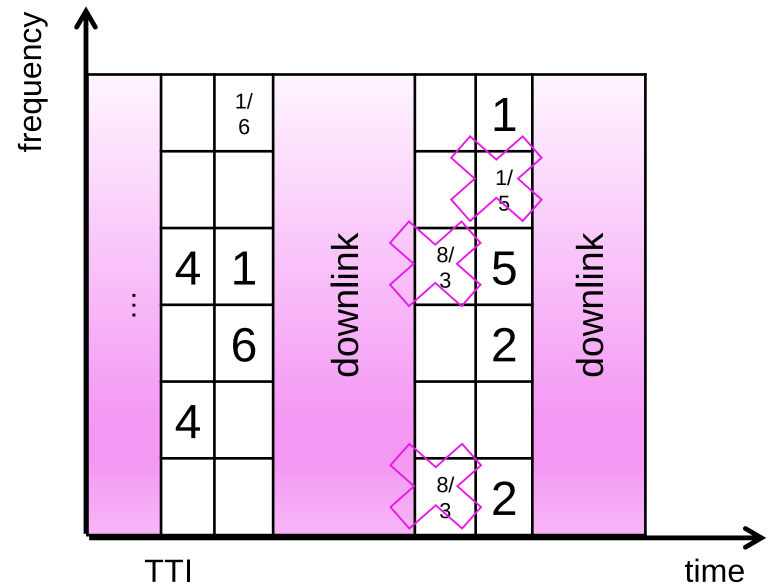


Scheme 1: Packet replication for increased reliability

- **For increasing reliability, replicate packets and send them on different resources on the reserved pool**
  - each active UE selects ***p*** resources at random
- **Creates more collisions**
  - but the chance that at least one replica is collision free may be larger
  - an optimal balance is to be found



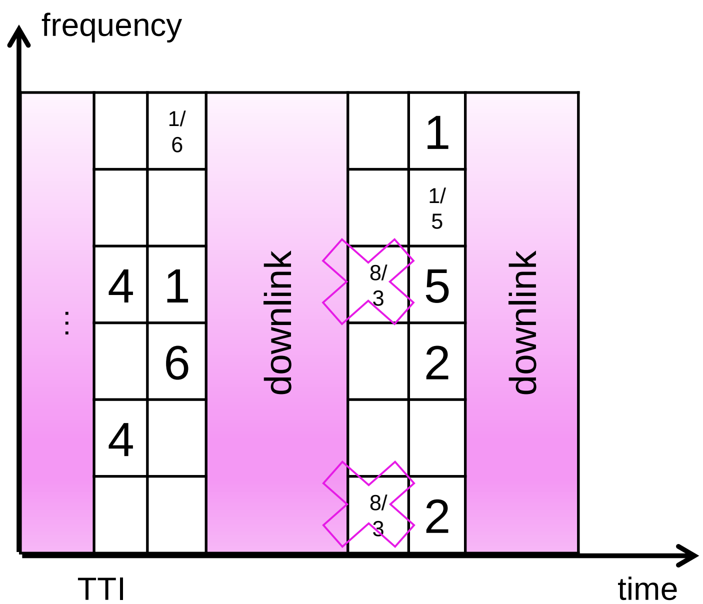
Basic ALOHA



ALOHA with replication p=2

Scheme 1: Packet replication for increased reliability

- **Proposition:** When each packet is replicated $p \in \mathbb{N}^*$ times on resources chosen at random from the reserved pool, the probability of loss is:



$$l(p) = 1 - \sum_{j=1}^{p} (-1)^{j+1} C_p^j \left((1-q) + q \frac{C_{K-j}^p}{C_K^p} \right)^{n-1}$$

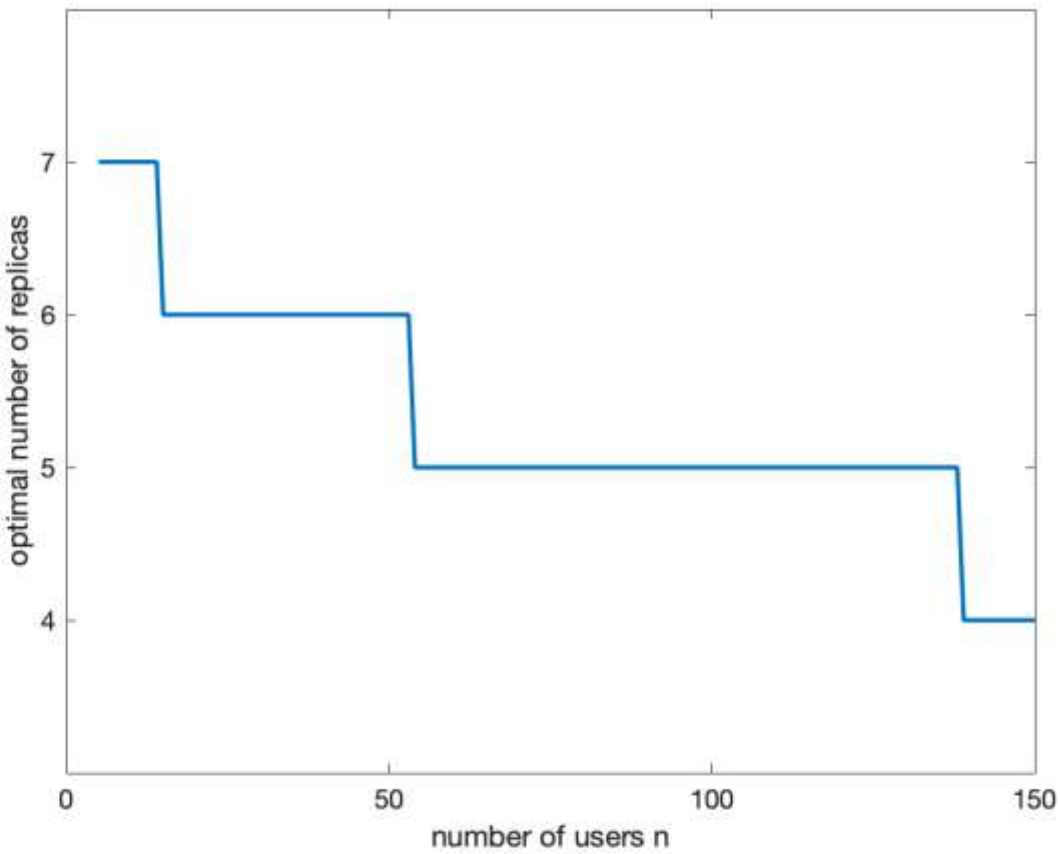
  - n users in total
  - each user active with probability *q<1*
  - K resources in total
  - p replicas for each packet
  - C_n^m : combinations of m among n

- **Hint about the proof:**
  - A_i is the event that the i-th resource is free
  - probability that a subset of size j is free $\mathbb{P}\{A_1 \cap \ldots \cap A_j\} = \left(1 - q + q \frac{C_{K-j}^p}{C_K^p}\right)^{n-1}$
  - probability of success:

$$\mathbb{P}\{A_1 \cup \ldots \cup A_p\} = \sum_{j=1}^{p} (-1)^{j+1} C_p^j \, \mathbb{P}\{A_1 \cap \ldots \cap A_j\}$$

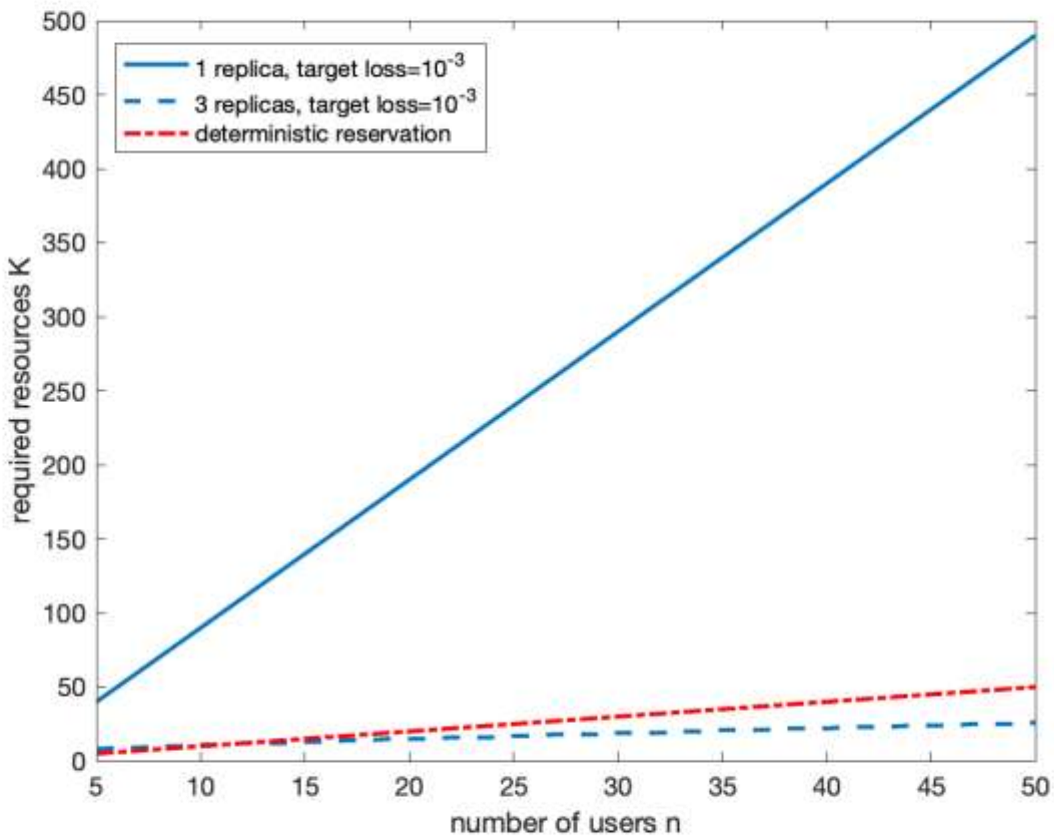
Optimizing the number of replicas

- The number of replicas that minimizes loss can be found
- Example for q=0.01, K=30

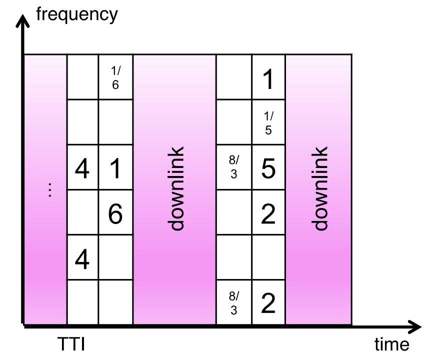


Performance of Aloha with frequency replication

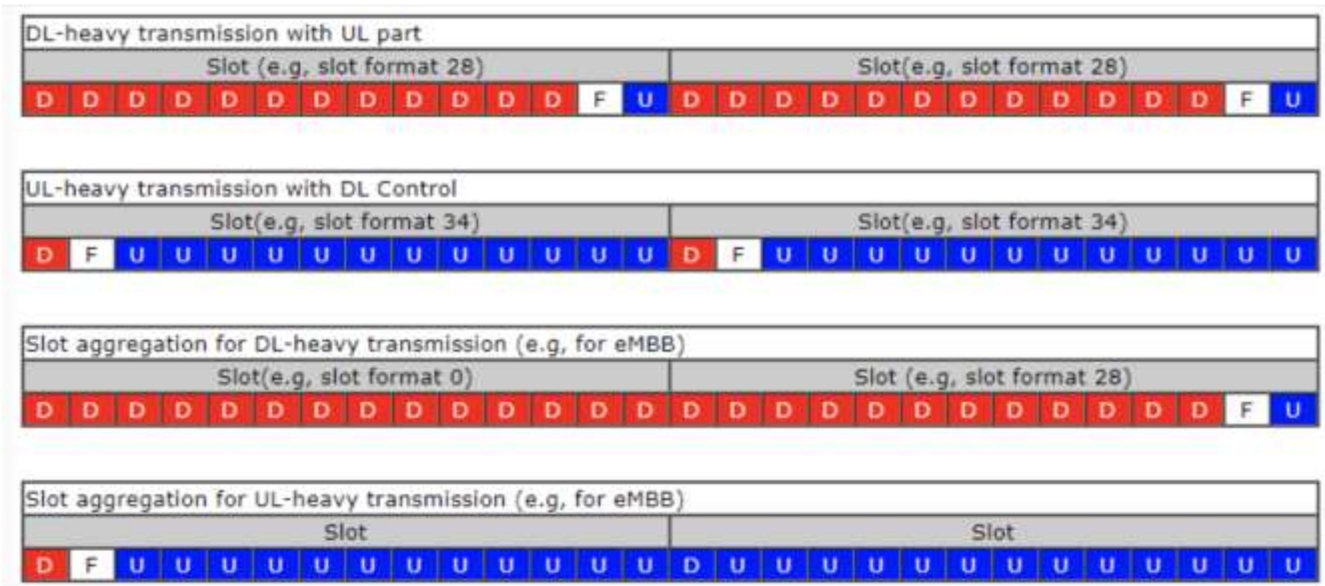
- **High reliability can be achieved by replication**



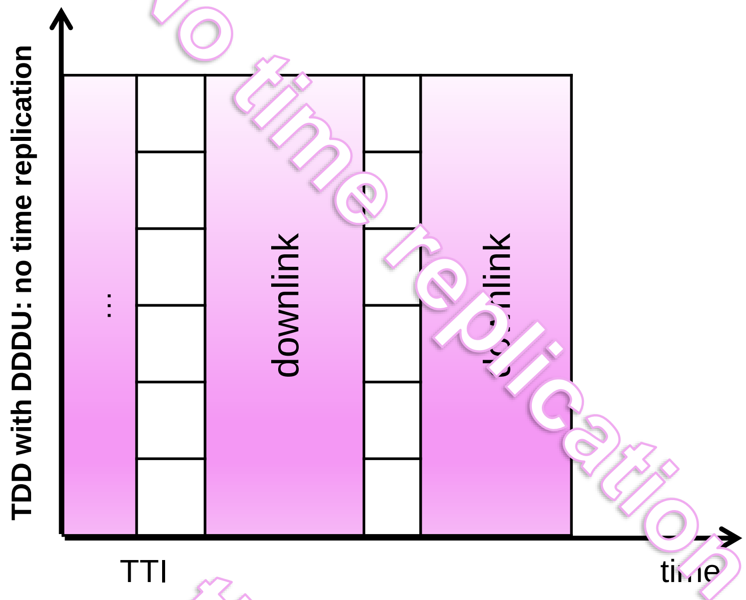
What if? temporal dimension, feedback…

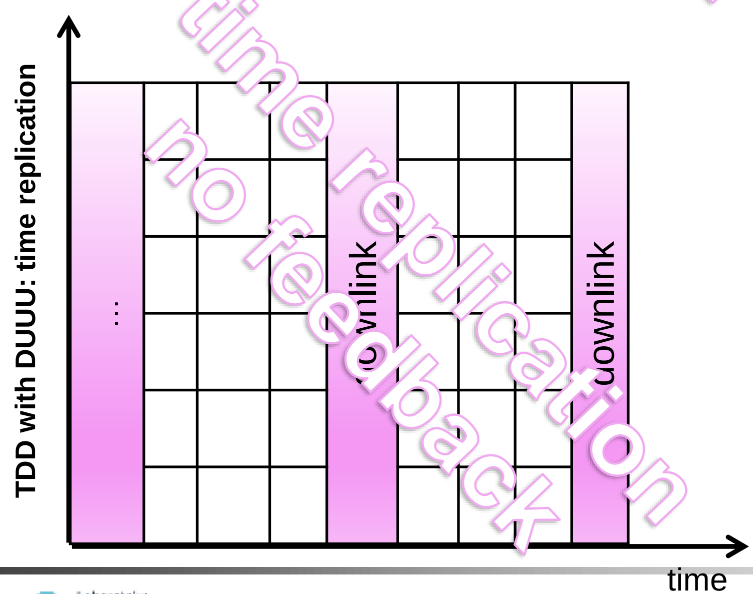


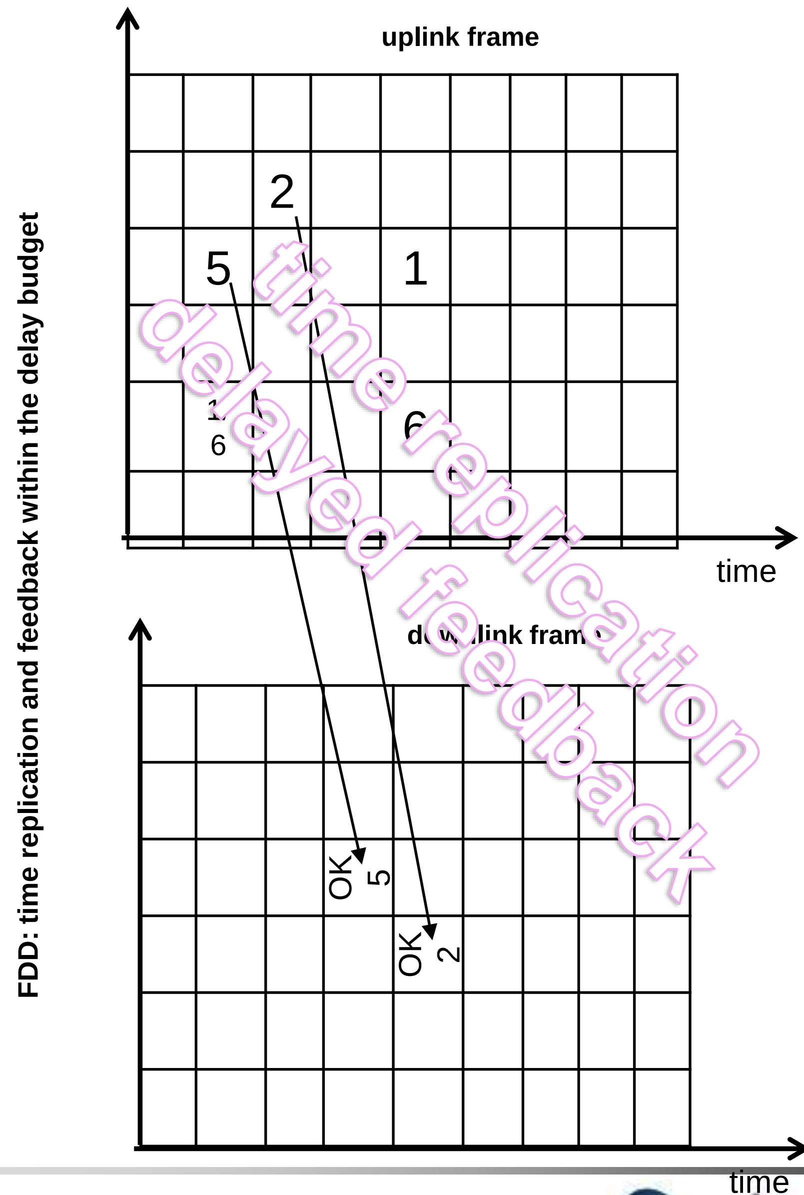
- **We proposedreplication in the frequency dimension**
- **Why not replicate in the temporal dimension as well?**
- **Why not wait for feedback before retransmission?**
- **Response: it depends:**
  - if the delay budget, combined with 5G interface, allows for temporal replication
  - if, in addition, there might be feedbacks within the delay budget, exploit them
- **Radio and service requirements are very diverse in 5G**



When to exploit time domain replication and feedback?

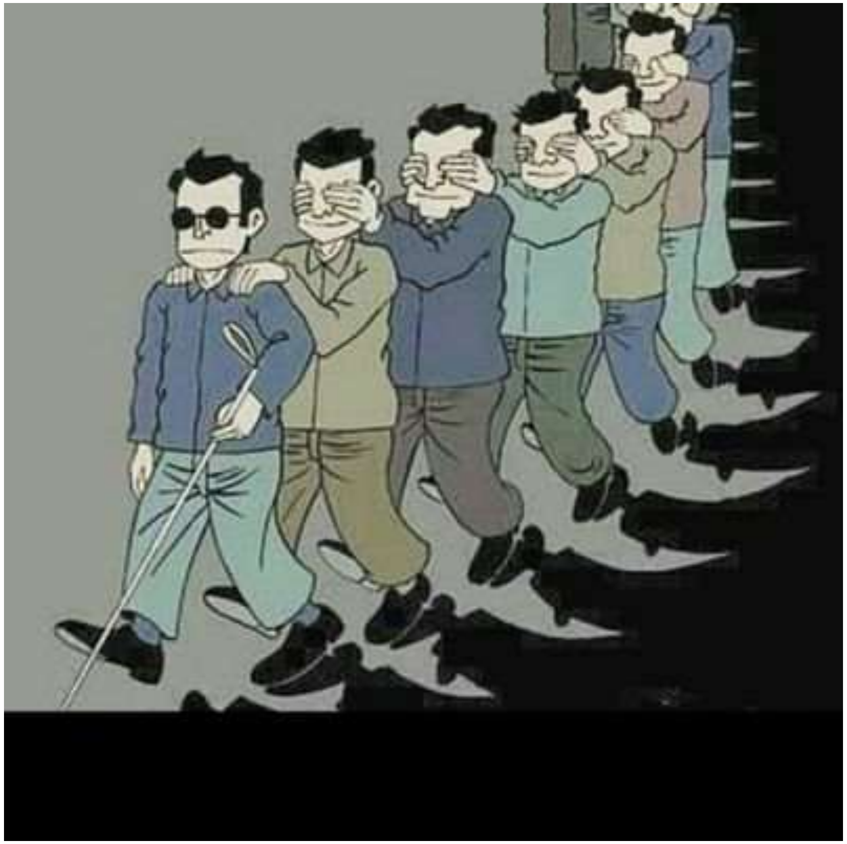






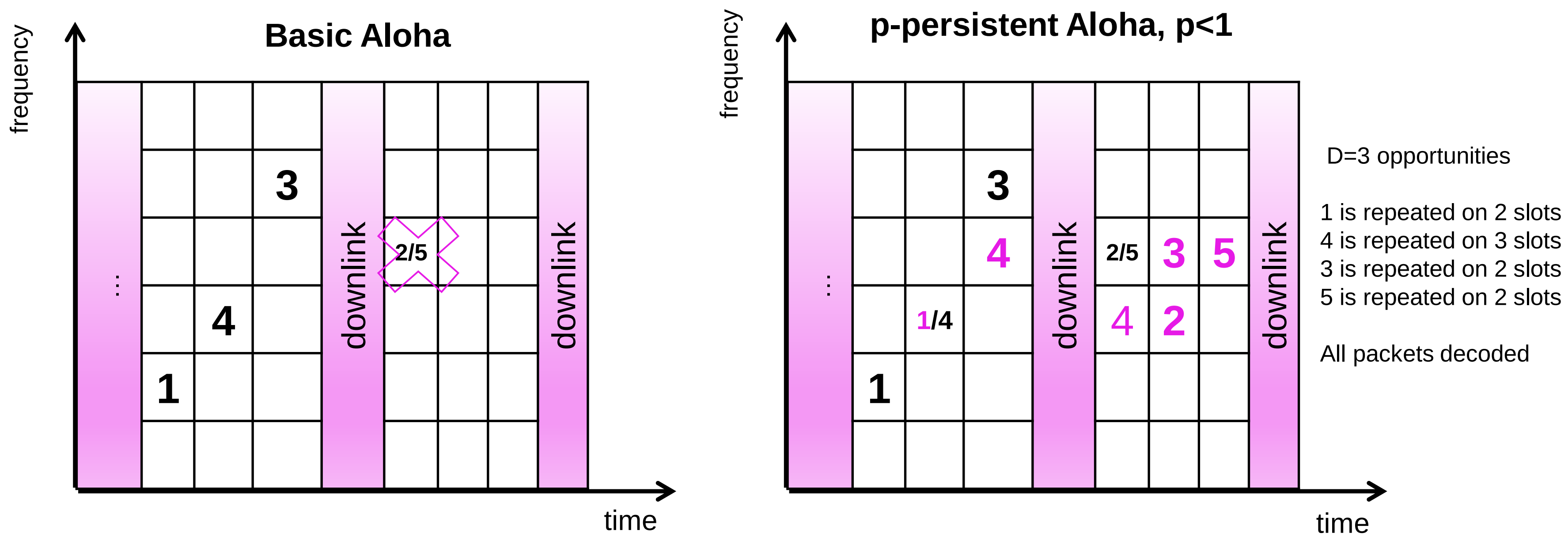
Outline

- Introduction
- **The blind**
- The far-sighted
- Perspectives



Solution 2: Blind probabilistic temporal replication

- **Hypothesis: D>1 slots are available within the delay budget**
  - each active UE transmits a replica per slot with probability ***p≤1***
  - similar to **p-persistent Aloha**, with multiple parallel channels



- **Advantage: time diversity increases reliability**
- **Drawback: the system is no more memoryless, a user remains active for several slots (*q* increases).**

Solution 2: Optimal blind probabilistic temporal replication

- **Proposition:** The optimal p-persistent scheme when there are *D* slots within the delay budget is:

$$p^* = \min[\frac{K}{\bar{q}n}, 1]$$ with the activity factor: $$\bar{q} = 1 - (1-q)^D$$

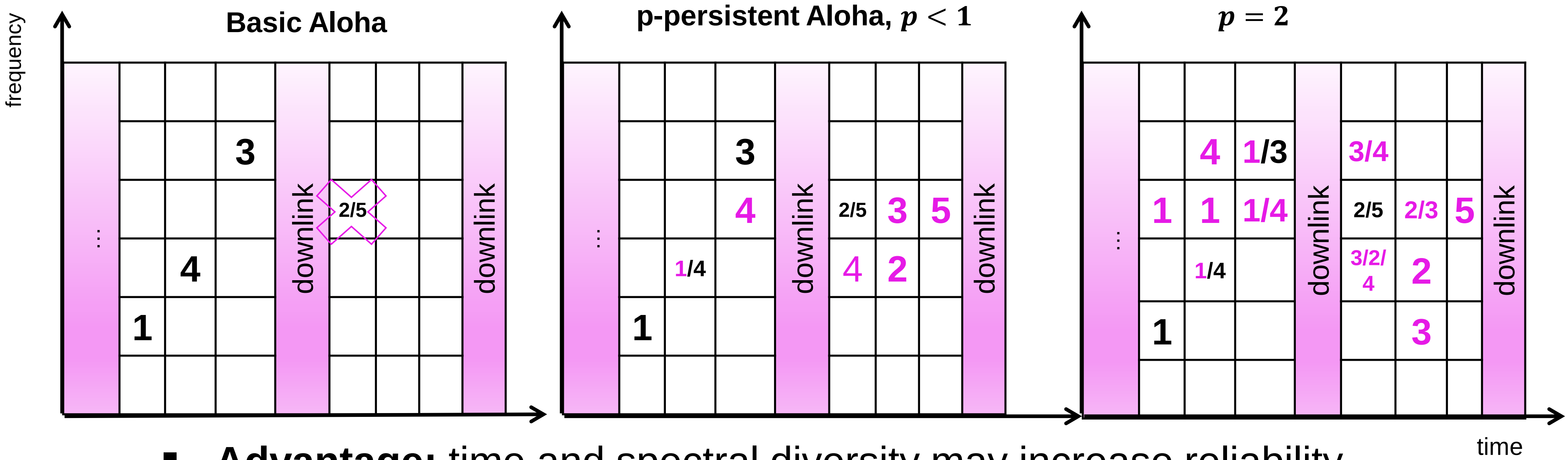
and the packet loss is:

$$l = \begin{cases} (1 - \frac{Ka_n}{\bar{q}n})^D, & \mathit{if}\ n > \frac{K}{\bar{q}} \\ (1 - (1 - \frac{\bar{q}}{K})^{n-1})^D, & \mathit{otherwise} \end{cases}$$ with $$a_n = (1 - \frac{1}{n})^{n-1}$$

- **Hints about the proof:**
  - a user is active (willing to transmit) on *D* consecutive slots, leading to $\bar{q}$
  - The probability that all replicas are lost is: $l(p) = [1 - p(1 - p\bar{q}/K)^{n-1}]^D$
  - The loss starts by decreasing and reaches its minimum for $p = K/n\bar{q}$. If however, $K/n\bar{q}$ is larger than 1, the best policy corresponds to *p = 1*.
- **Remark:** if the optimal p-persistent policy is p=1, this means that it is better to **send more than one replica per slot….** Solution 3

Solution 3: Blind temporal/spectral replication

- **Hypothesis: several slots available within the delay budget**
  - each active UE transmits a number of replicas per slot $p \in \mathbb{N}^*$

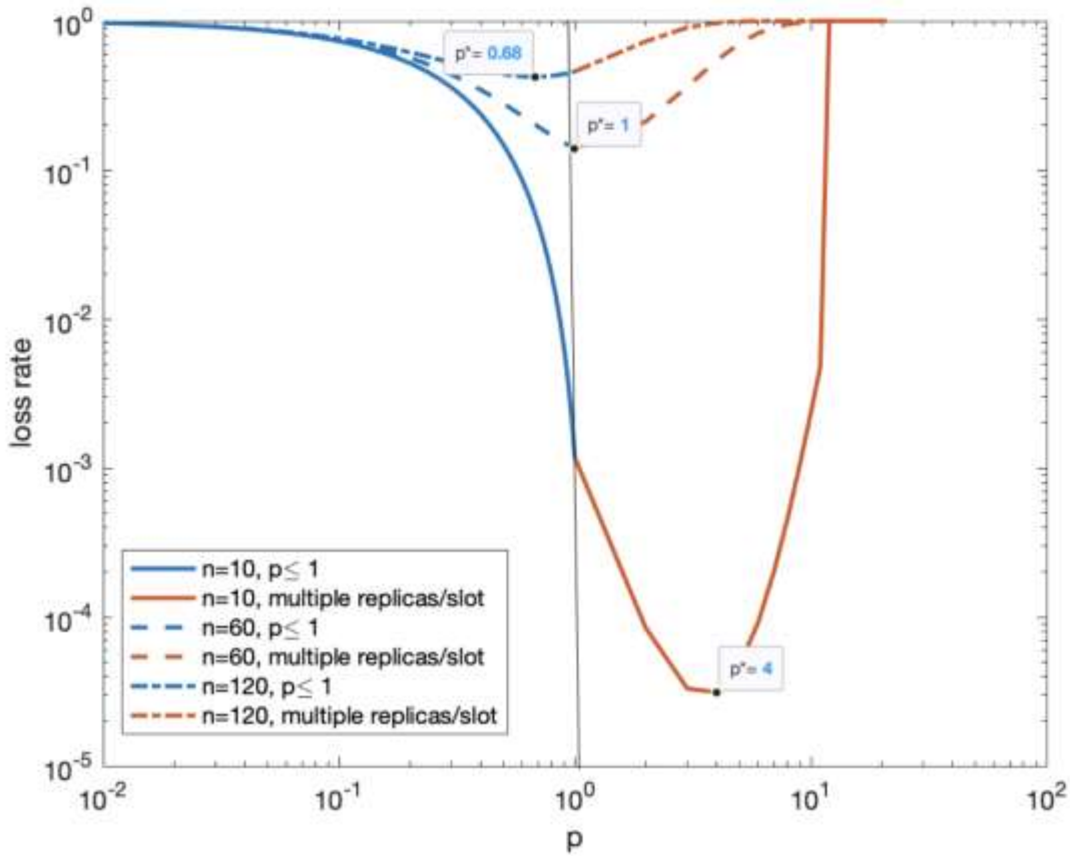


- **Advantage:** time and spectral diversity may increase reliability
- **Drawback:** the load is increased**, to be used in low traffic regimes**
- **Proposition:** For the blind repeated case with $p \in \mathbb{N}^*$, the loss is :

$$l(p) = [1 - \sum_{j=1}^{p} (-1)^{j+1} C_p^j \left((1 - \bar{q}) + \bar{q} \frac{C_{K-j}^p}{C_K^p} \right)^{n-1}]^D$$

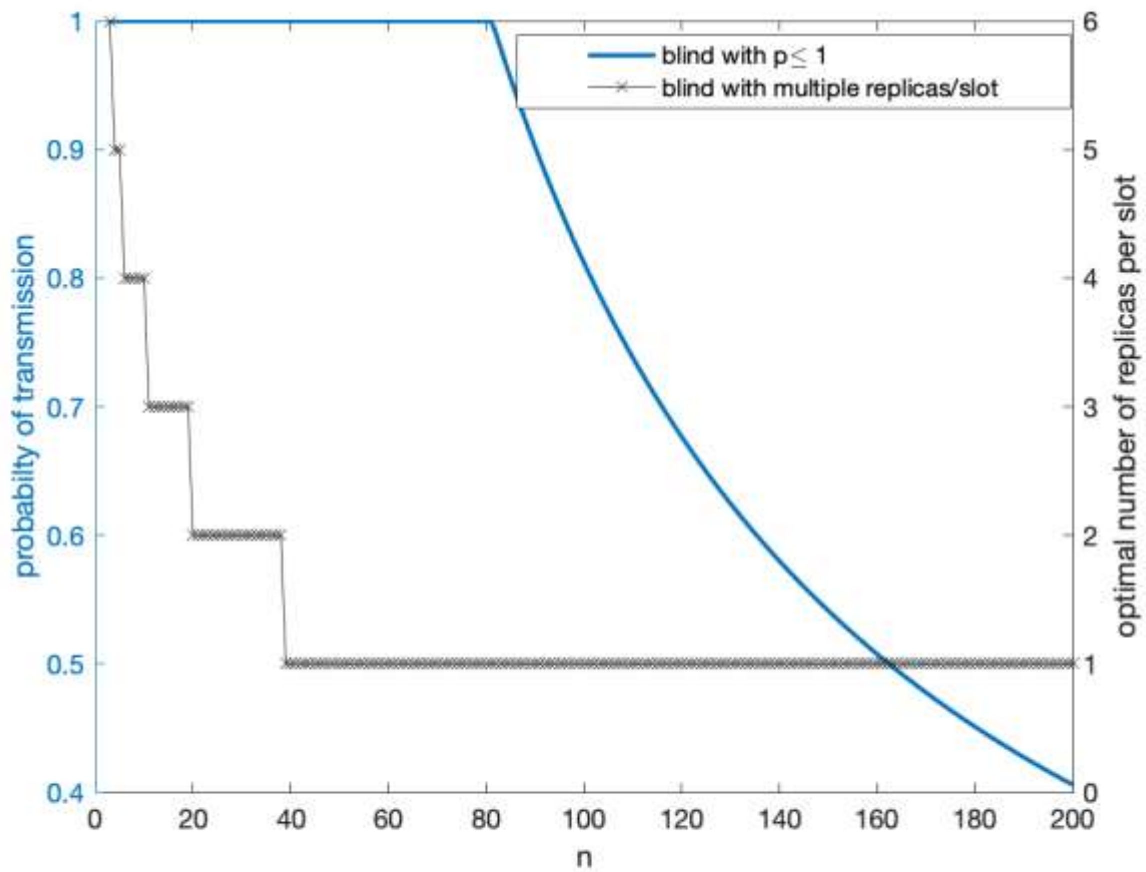
Blind repeated replication: loss versus p

- **The loss rate is minimized for some *p* that depends on the load**
  - for a low load (small n), the optimal policy corresponds to $p \in \mathbb{N}^*$ replicas per slot
  - for a high load (large n), it is optimal to send less than one replica per slot.
  - p=1 for intermediate loads



Blind repeated replication: optimal p illustration

- **For low loads, it is optimal to send more than one replica**
- **For large loads, p-persistent Aloha is better**



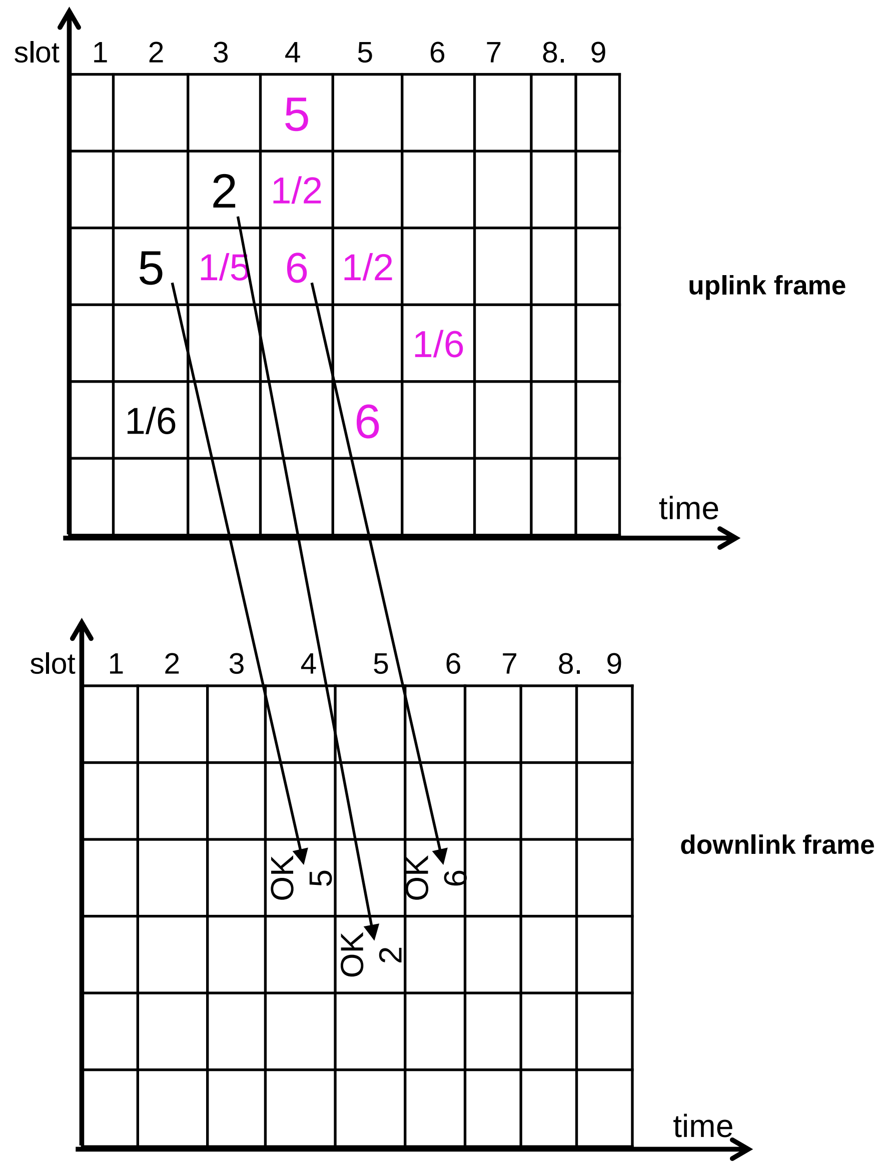
Outline

- Introduction
- The blind
- **The far-sighted**
- Perspectives

"Honey, I think my arms are getting too short!"*

Feedback impact on the transmission

- Transmitters are **far-sighted but not blind**, and when a delayed feedback is received, they stop.
- Example for an FDD system with 5 slots within the delay budget, and a feedback that arrives after 2 slots.
- 5 starts sending in slot 2, receives ACK slot 4, stops sending slot 5
- 2 starts sending in slot 3, receives ACK slot 5, sops sending slot 6
- 6 starts sending slot 2, receives ACK slot 6, never stops sending
- 1 is lost
- **1 could have been decoded if 2, 5 and 6 were aware that their replicas are received**



Optimal far-sighted temporal replication

- **Proposition:** In the far-sighted case with *D* slots for replication and a delayed feedback of $\Delta < D$ slots, the optimal replication policy is computed as for the blind case, with the activity factor computed by:

$$\bar{q} = 1 - (1-q)^{\Delta+1} \prod_{i=1}^{D-\Delta-1} (1 - q + q(1-(1-s)^i))$$

- **Why this new activity factor?**
  - once a user generates a packet, he remains active on *D* consecutive slots, unless he receives an ACK.
  - The user cannot receive an ACK before $\Delta < D$ slots, so a user that generated a packet in the previous $\Delta+1$ slots is still active for sure.
  - For a packet generated in a slot older than Δ, it might have received an ACK.
  - A user does not bring a packet from a slot *i* older than Δ if:
    - either he did not generate a packet on i
    - or the packet generated has received an ACK

$$\left.\begin{array}{l}\text{either he did not generate a packet on i}\\ \text{or the packet generated has received an ACK}\end{array}\right\} 1 - q + q(1-(1-s)^i)$$

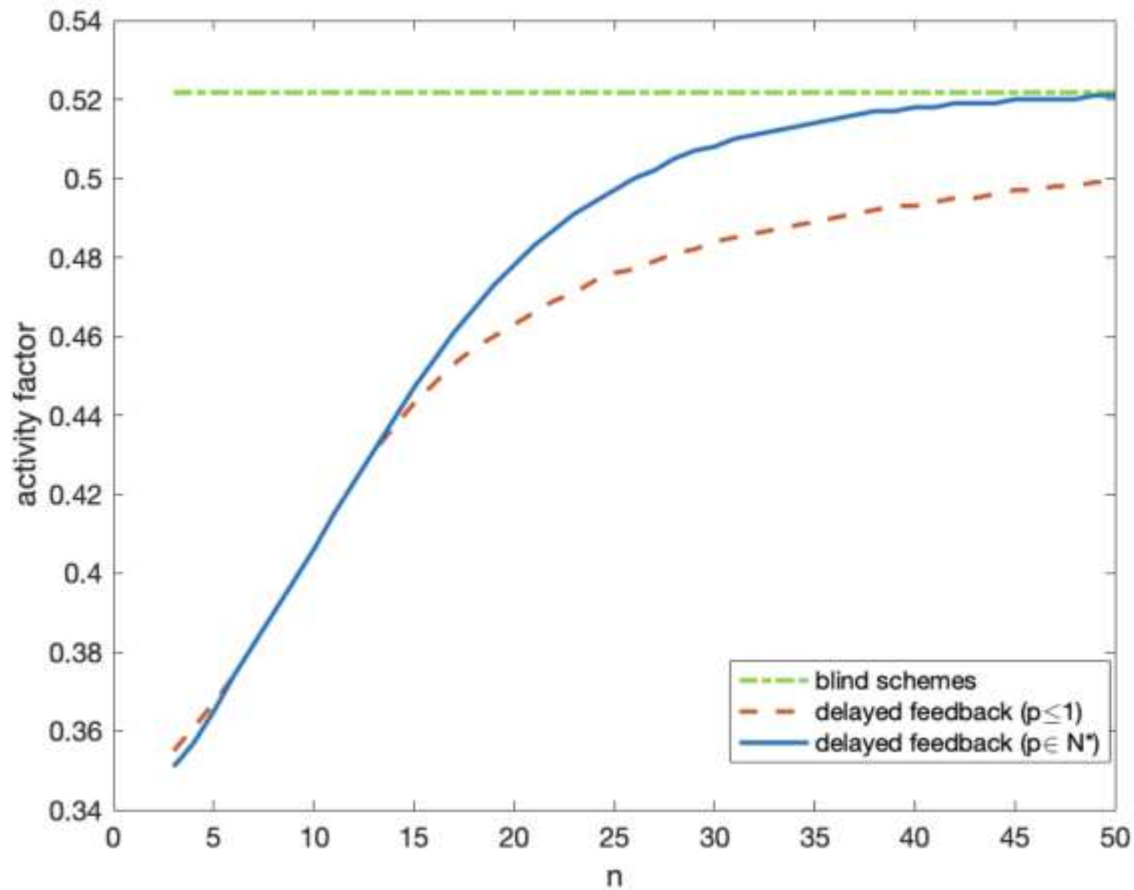
How activity varies when transmitters are no more blind

- **Activity of users, and thus load on the radio interface, decreases when there is feedback, even delayed**

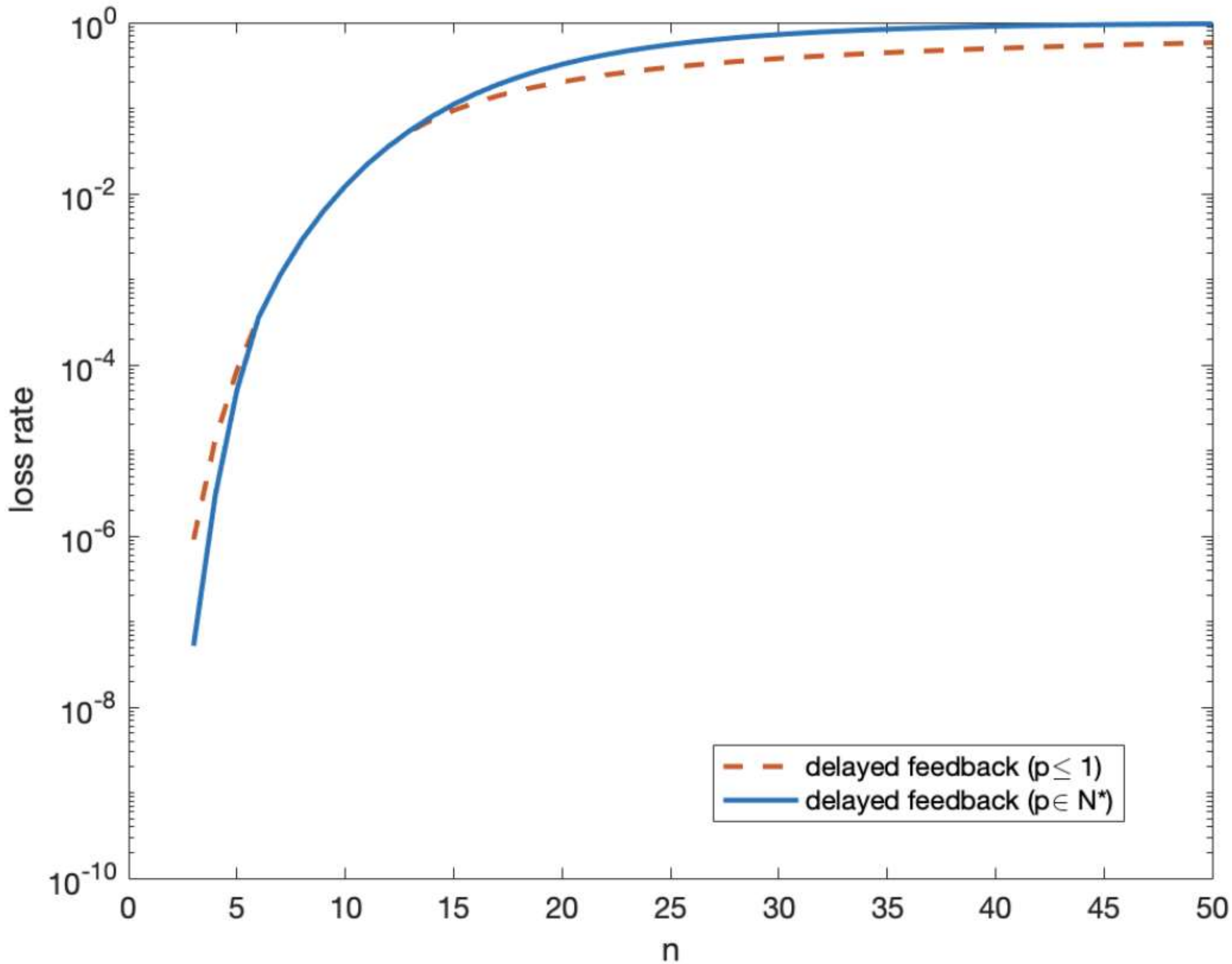
- **Blind:**
  - $\bar{q}$ large and constant

- **Far-signted**
  - $\bar{q}$ increases with *n*

- **Low number of users**
  - p>1 is better
- **Large number of users**
  - p<1 is better



Loss rate for far-sighted policies



Before concluding: Application to 5G New Radio

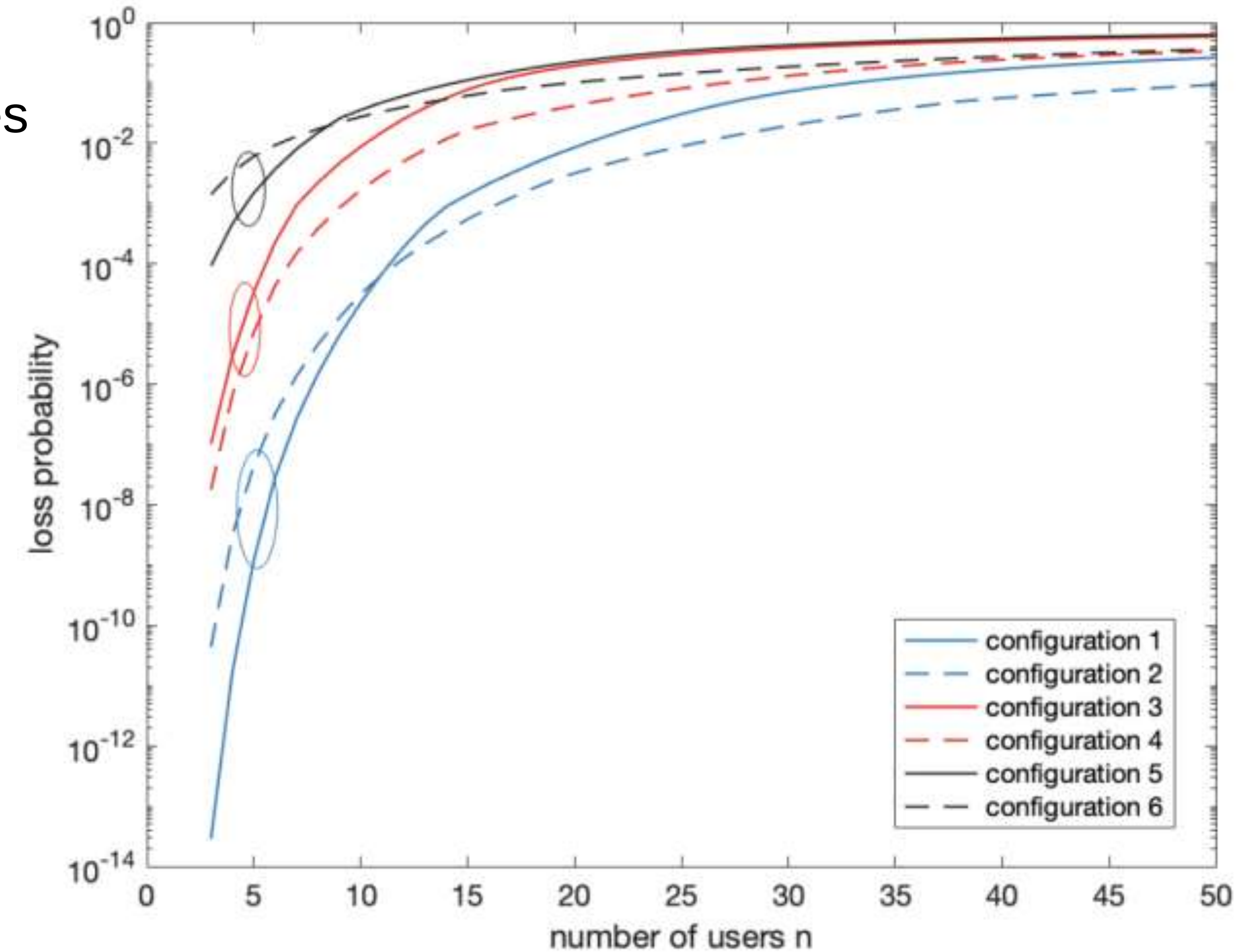
- **We consider a 5G NR system with 10 MHz reserved for URLLC**
- **Several configurations (numerologies) are possible**

	Configuration	Slot length	Capacity packets/slot	Slots (D)	ACK delay slots	ACK?	
1	(15 KHz;2 sym/slot;FDD)	0.144 ms	11	7	5	Yes	far-sighted, D=7, Δ=4
2	(15 KHz;4 sym/slot;FDD)	0.288 ms	22	3	5	No	Blind, D=3
3	(30 KHz;2 sym/slot;TDD;DDUU)	0.072 ms	5	6	7	Yes	far-sighted, D=6, Δ=3
4	(30 KHz;4 sym/slot;TDD;DDUU)	0.144 ms	11	4	7	No	Blind, D=4
5	(30 KHz;2 sym/slot;TDD;DDDU)	0.072 ms	5	3	7	Yes	far-sighted, D=3 Δ=2
6	(30 KHz;4 sym/slot;TDD;DDDU)	0.144 ms	11	1	7	No	one-shot blind

- **Conf. 1 and 2 are comparable (10 MHz for downlink)**
- **Conf. 3 and 4 are comparable (10 MHz, half time for DL, half-time for uplink)**
- **Conf. 5 and 6 are comparable (10 MHz, 75% for DL, 25% for uplink)**

Comparing 5G NR configurations

- Giving more resources for URLLC is better (obvious)
- There is no clear advantage for choosing a smaller slot.
- Optimal numerology depends on the load



Outline

- Introduction
- The blind
- The far-sighted
- **Perspectives**

We have solved the probem, but is this THE problem to solve?

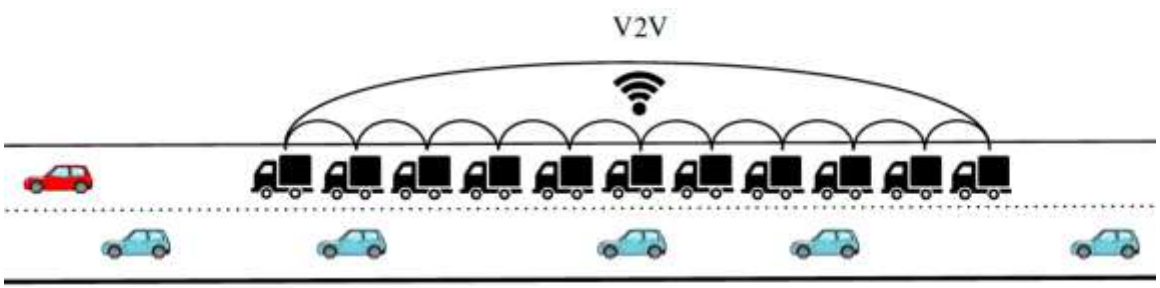
- **Back to the starting point: we solved a problem defined by 3GPP, the organism that standardizes 4G/5G and 6G…**
  - the **proportion** of packets,
  - **correctly received** by the controller
  - within the **delay budget** (e.g. 1 ms)
  - has to be larger than a **reliabilty target** (e.g. loss probability < 10^{-5})

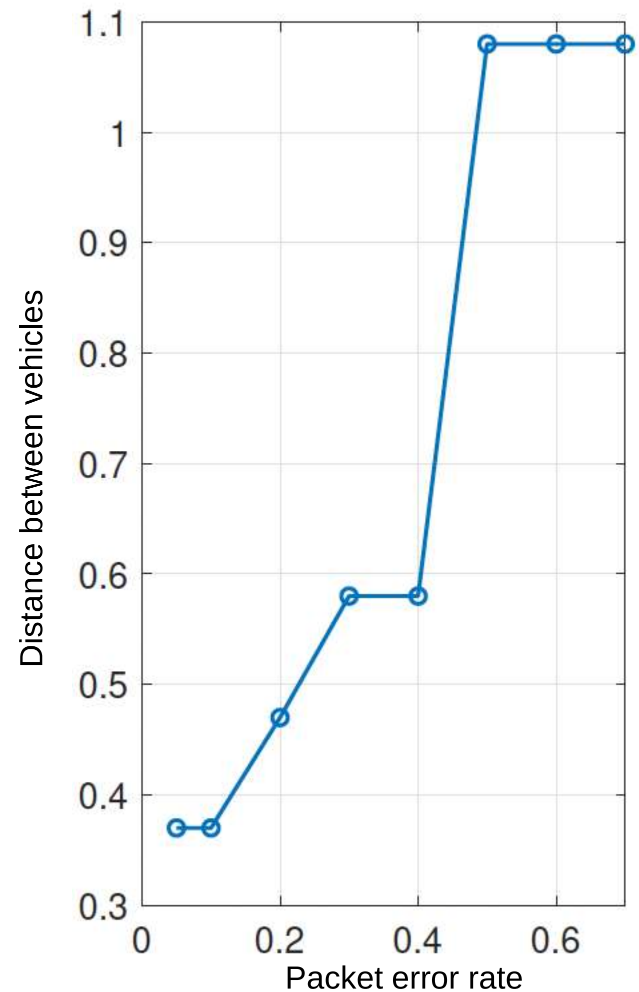
- **But why 1ms?**

- **What happens if some packets are lost?**

Problem of vehicle platooning: 10^{-5} loss target useless

- **One of the flagship 5G URLLC use cases is vehicular networks**
- **Platoons of vehicles on highways**
  - exchange velocity and acceleration
  - objective: reduce distance between vehicles
  - thus reducing fuel consumption





- **We simulated the platoon:**
  - platoon controller and communication network
- **Can support a loss rate up to 10%**
  - distance between venicles < 0.5 m
  - cannot go below this distance (safety)

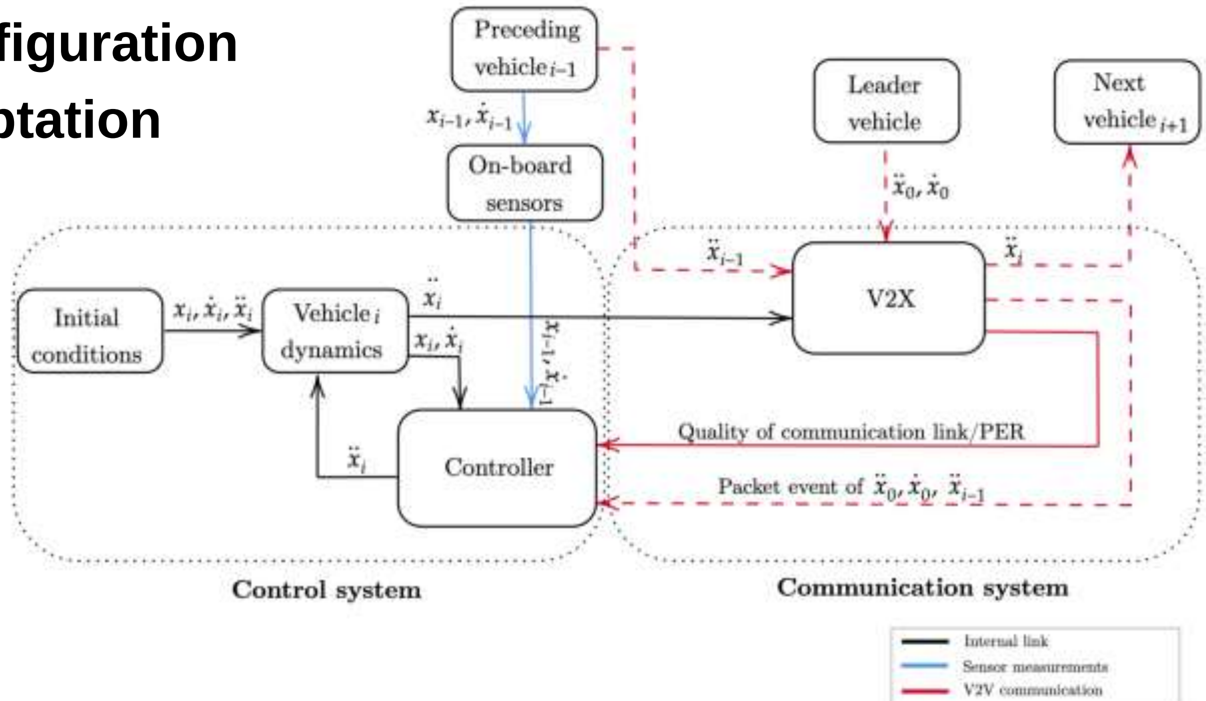
Which problem to solve?

- **Bad news:** *Pr(delay>1ms)<10*$^{-5}$ is not need in many practical cases
- **Good news:** minimize packet loss is always a valid target
- Is there anything else to do?
- **Axis 1**: Joint design of communication and control schemes
- **Axis 2:** explore different metrics, other than loss, that are more related to applications

Axis1: joint design of communication and control

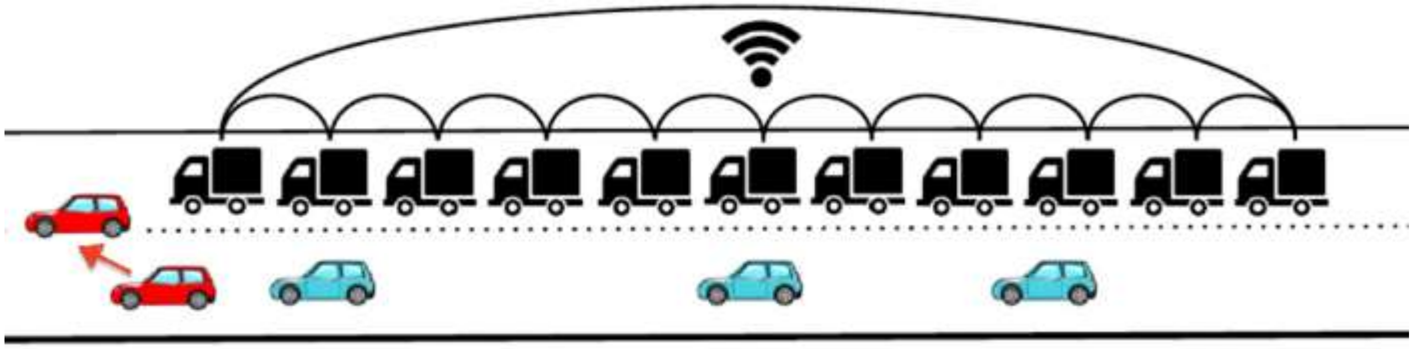
- Joint design of communication and control schemes
  - minimize the loss rate on the network
  - adapt the application to the network status

- **Step 1: Network monitoring**
- **Step 2: Network reconfiguration**
- **Step 3: Controller adaptation**



Axis 2: metrics other than latency: age, freshness, value..

- There are lots of transmitters, competing on the same channel
- But they are cooperative, in the sense that they have the same objective



- What if delay is not the most important metric?
  - red car is moving dangerously: packets of the leader have the largest **value**
  - if the **age** of a packet is large, it has less value than a new packet**. Freshness** of information is key
- We have shown that ensuring a URLLC target (e.g. 1 ms) leads to a policy that does not necessaritly ensure freshness
  - URLLC: start mild and become aggressive near the target
  - Freshness: start aggressive and reduce pace as time goes…

Many thanks to my co-authors on this topic

- **From Orange:**
  - Patrick Brown
  - Matha Deghel
  - Meriem Mhedhbi
  - Ana Galindo Serrano
- **From Telecom SudParis**
  - Tijani Chahed
- **From CentraleSupelec:**
  - Richard Combes
- **And my (former) PhD students**
  - Ayat Zaki Hindi
  - Tiago Rochas Goncalves

Some references

- **On the critical IoT resource allocation (URLLC-like)**
  - P. Brown and S.E. Elayoubi, Semi-distributed Contention-based Resource Allocation for Ultra Reliable Low Latency Communications, **IEEE Infocom 2020**, July 2020.
  - S-E. Elayoubi, M. Deghel, P. Brown and A. Galindo Serrano, Radio Resource Allocation and Retransmission Schemes for URLLC over 5G networks, **IEEE JSAC**, 37 (4), 896-904, 2019.
  - P. Brown, M. Deghel, and S.E. Elayoubi. "Wireless Communication Devices, Systems and Methods for Decoding Data Packets for the Establishment of Latency-Critical Services." **U.S. Patent** No. 17/046,932.
  - P. Brown and S.E. Elayoubi. "Wireless communication devices, systems and methods for establishing latency-critical services." **U.S. Patent** No. 17/046,934.
- **On the joint control/communications optimization**
  - T. Rochas, V. Varma and S. E. Elayoubi, “Performance of Vehicle Platooning Under Different V2X Relaying Methods,” in IEEE PIMRC, september 2021.
  - T. Rochas, V. Varma and S. E. Elayoubi, “Performance and design of robust platoons under different communication technologies,” **IEEE VTC-Spring**, 2021.
- **On the age minimization versus URLLC**
  - A. Zaki-Hindi, S. Elayoubi and T. Chahed, "Transmission policy design for critical services under different objectives," **IEEE GLOBECOM**, 2021.