Effective Rate of URLLC with Short Block-Length Information Theory

Najib Odhah, Eckhard Grass, and Rolf Kraemer

innovations
for high
performance
microelectronics

Leibniz Association

13.07.2022

Agenda

1 • Introduction

2 • Finite Block-Length Information Theory

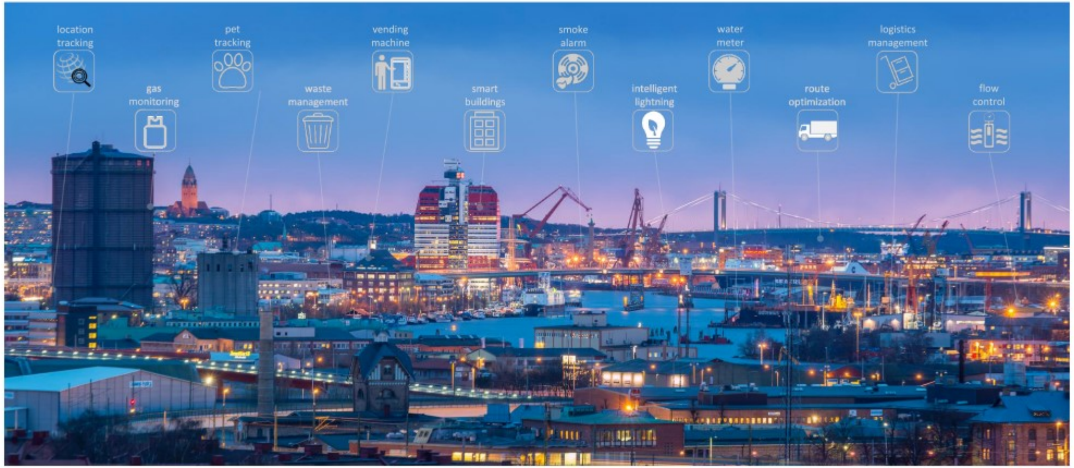
3 • Effective Capacity/Bandwidth

4 • Case Study

5 • Summary

Machine-Type Communications (MTCs)

Key enabler of future autonomous systems



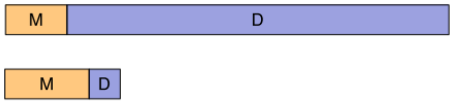
source: IoTpool

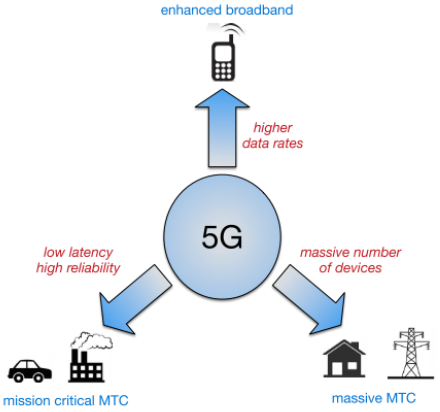
- 5G ⇒ massive MTC; ultra-reliable, low-latency comm.
- Low-power wireless-area networks ⇒ LoRa-WAN, SigFox,. . .

MTC traffic has unique characteristics: how to support it?

Machine-Type Communications (MTCs)

Unique characteristics of MTC traffic





- massive number of connected terminals
- transmitters are often idle
- short data packets
- low latency, high reliability
- high energy efficiency

Machine-Type Communications (MTCs)

Some characteristics of URLLC and mMTC traffic

massive MTC

- Small information payload (100 bits)
- High user density (10^6 devices/Km^2)
- Sporadic TX (less that 1 per minute)

  $\Rightarrow$ 120 dof per user at $B = 20\,\text{MHz}$

URLLC

- Small information payload (100 bits)
- Low latency ($100\,\mu s$)
- Low error prob. (below 10^{-5})

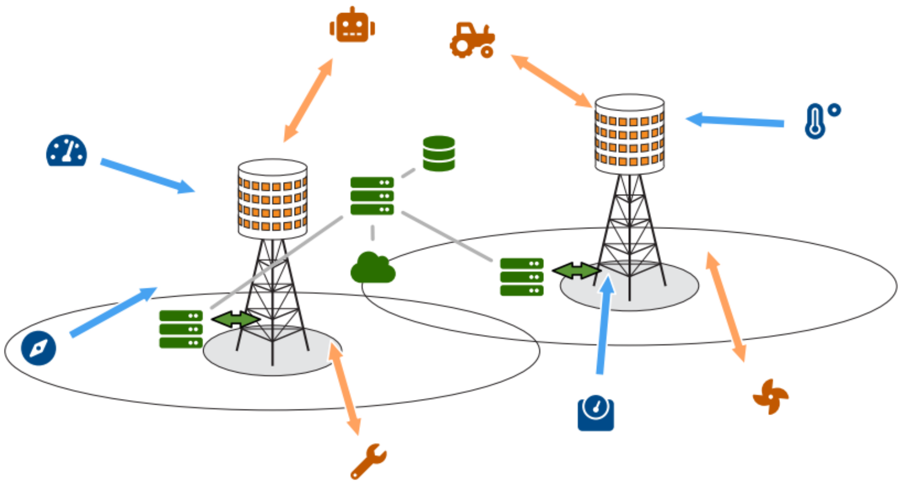
  $\Rightarrow$ 168 dof per user for 5G

Key design question

How to transmit hundreds of bits in hundreds of dof per user (short-packet regime) under stringent constraints on reliability/latency/energy efficiency?

Ultra-Reliable Low Latency Communications (URLLCs)

Wireless connectivity to the network computing fabric



massive MTC

- Uplink mostly
- High energy efficiency

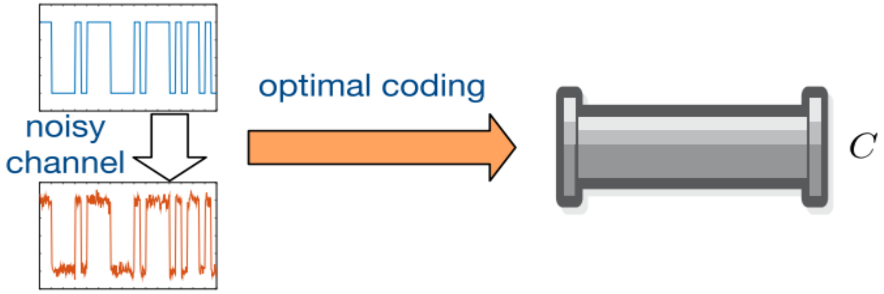
URLLC

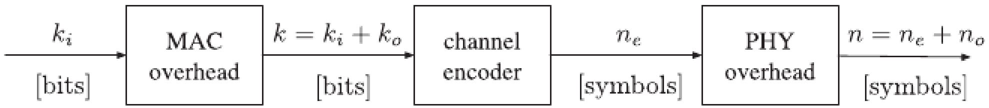
- Bidirectional
- Low latency, high reliability

Agenda

Asymptotic Information Theory

The bit-pipe approximation





- C provides a performance benchmark
- C useful to design algorithms for
  - User scheduling
  - Resource allocation
  - Power control
  - ...
- Lower bounds on C are often used, e.g.,

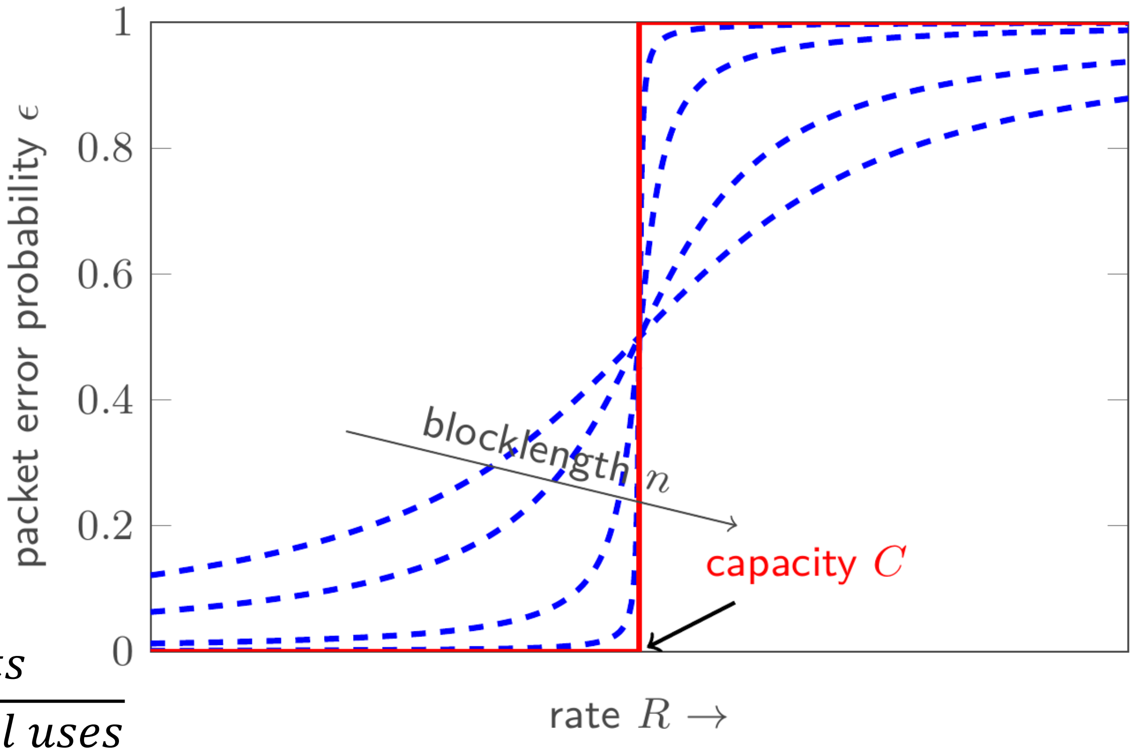
$$C \geq \mathbb{E}[\log(1 + \mathsf{sinr})]$$

Channel Capacity

$$R^*(n,\epsilon) = C - \sqrt{\frac{V}{n}}Q^{-1}(\epsilon) + \mathcal{O}\left(\frac{\log n}{n}\right)$$

$$C = \lim_{n\to\infty} R^*(n,\epsilon).$$

$$C \geq \mathbb{E}[\log(1+\mathrm{SINR})]\,\frac{bits}{channel\ uses}$$

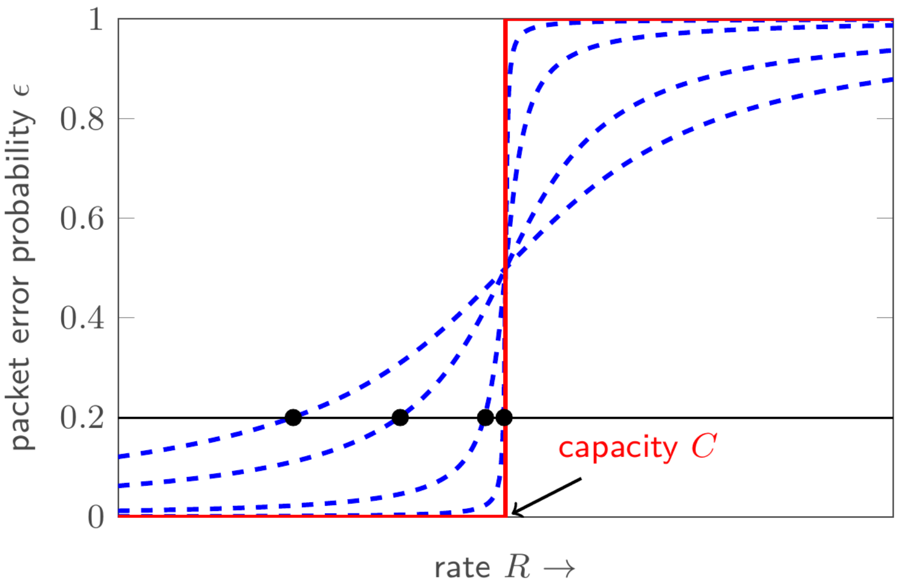


Y. Polyanskiy, H. V. Poor, and S. Verdu ́, "Channel coding rate in the finite blocklength regime," IEEE Trans. Inf. Theory, vol. 56, no. 5, pp. 2307–2359, May 2010.

Outage Channel Capacity

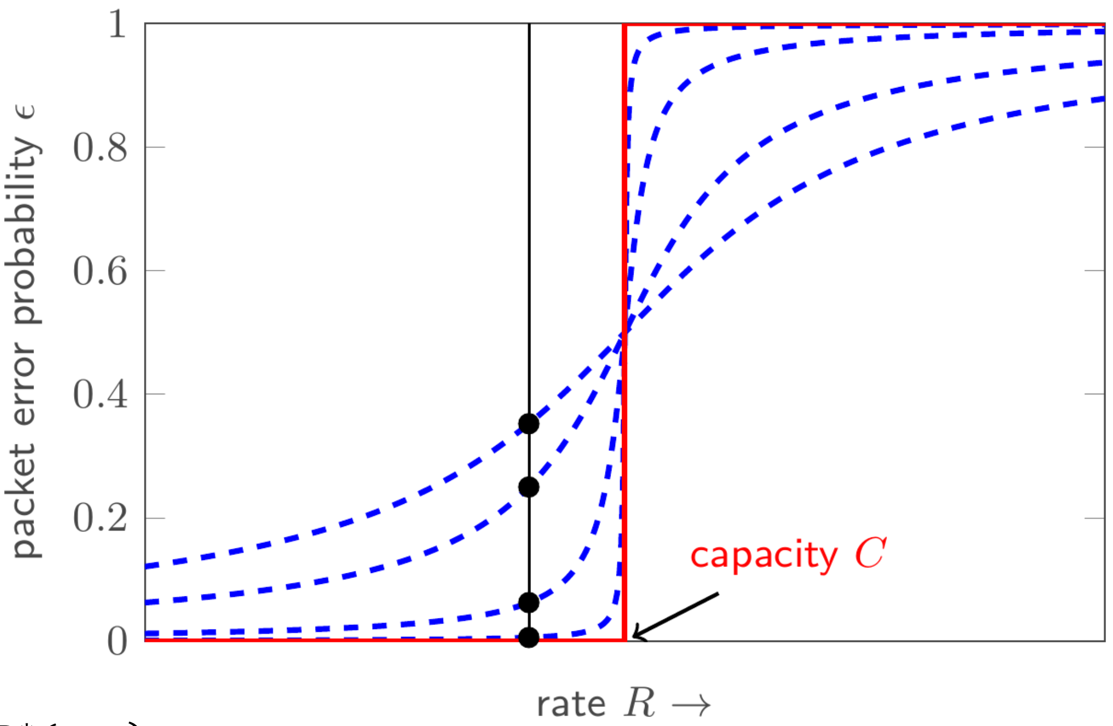
outage capacity C_ϵ is defined as the largest rate $\frac{k}{n}$ such that, for every sufficiently large packet length n, there exists an encoder/decoder pair whose packet error probability does not exceed ϵ.

$$C_\epsilon = \lim_{n\to\infty} R^*(n,\epsilon).$$



Ergodic Channel Capacity

The capacity C (in wireless communications also referred to as ergodic capacity) is defined as the **largest rate** $\frac{k}{n}$ such that there exists an encoder/decoder pair whose packet error probability can be made **arbitrarily small** by choosing the packet length sufficiently large.



$$C = \lim_{\epsilon \to 0} C_\epsilon = \lim_{\epsilon \to 0} \lim_{n \to \infty} R^*(n, \epsilon)$$

Vertical asymptotics $\Rightarrow$ error exponent

Short Block-length Information Theory

for various channels with positive capacity C, the maximal coding rate $R^*(n,\epsilon)$ can be expressed as:

$$R^*(n,\epsilon) = C - \sqrt{\frac{V}{n}}Q^{-1}(\epsilon) + \mathcal{O}\left(\frac{\log n}{n}\right)$$

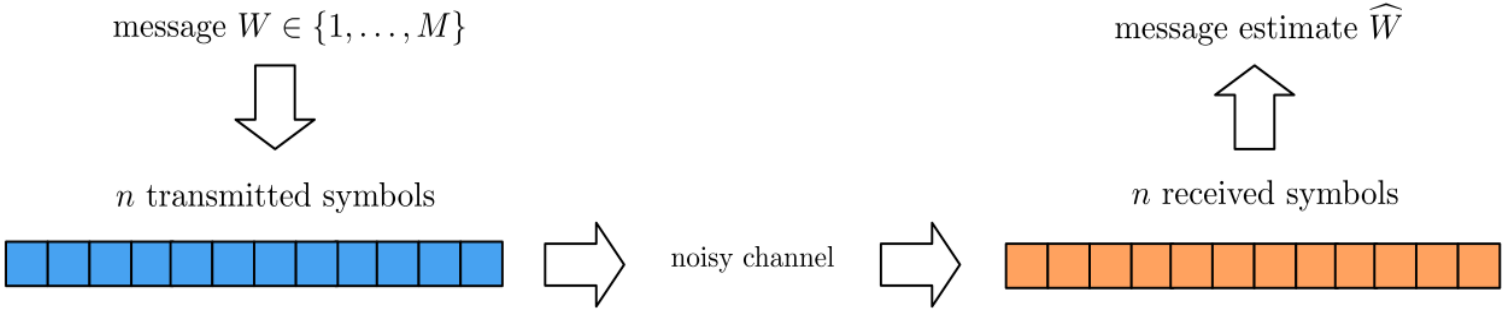
We denote by $R^*(n,\epsilon)$ The **maximum coding rate** at finite packet length n and finite packet error probability ϵ, i.e., the largest rate k/n for which there exists an encoder/decoder pair of packet length n whose packet error probability P_e does not exceed ϵ.

Where: where $\mathcal{O}\left(\frac{\log n}{n}\right)$ comprises remainder terms of order $\frac{\log n}{n}$.

$Q^{-1}(.)$ denotes the inverse of the Gaussian Q function.

V is the so-called channel dispersion due to short packet transmission.

Channel Coding Problem



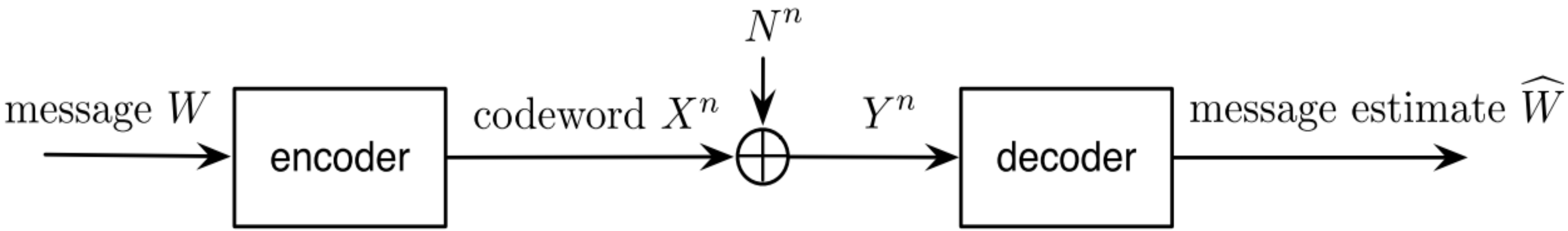
Notation

- Error probability: $\epsilon = \mathbb{P}[\widehat{W} \neq W]$
- rate: $R = (\log M)/n$

Performance metrics

- Minimum error probability: $\epsilon^*(n, R)$
- Maximum coding rate: $R^*(n, \epsilon)$

Binary Input AWGN Channel



$$Y_k = \sqrt{\mathrm{snr}} X_k + N_k, \quad k = 1, \ldots, n, \quad X_k \in \{-1, 1\}$$

- Capacity:

$$C = \lim_{n \to \infty} R^*(n, \epsilon) = \frac{1}{\sqrt{2\pi}} \int e^{-z^2/2} \left(\log 2 - \log\left(1 + e^{-2\mathrm{snr} - 2z\sqrt{\mathrm{snr}}}\right)\right)$$

- $C = 0.5$ bits/channel use at $\mathrm{snr} \approx 0.189\,\mathrm{dB}$

Normal Approximation for Binary Input AWGN Channel

Polyanskiy et al. have recently provided a unified approach to obtain tight bounds on $R^*(n,\epsilon)$. They showed that for various channels with positive capacity C, the maximal coding rate $R^*(n,\epsilon)$ can be expressed as :

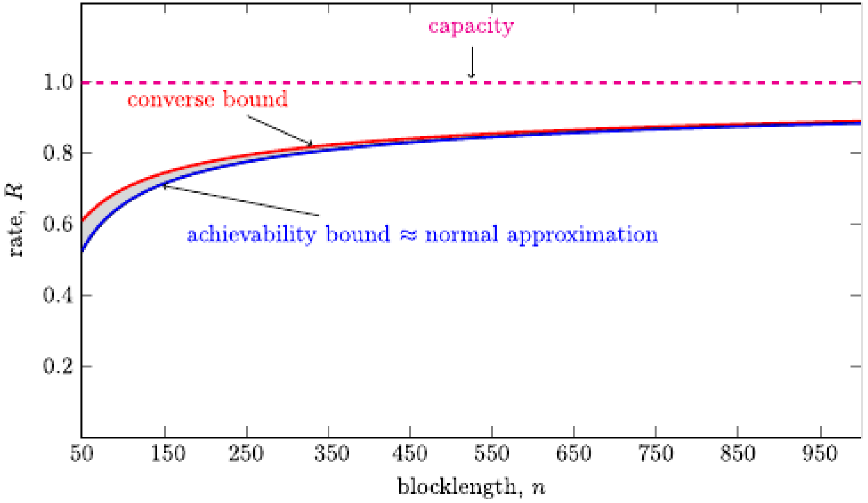
$$R^*(n,\epsilon) = C(\rho) - \sqrt{\frac{V(\rho)}{n}}Q^{-1}(\epsilon) + \frac{\log n}{2n}$$

A communication channel can be thought of as a bit pipe of randomly varying size. Specifically, the size of the bit pipe behaves as a Gaussian random variable with mean C and variance V/n. Hence, V is a measure of the channel dispersion.

$$\epsilon^*(k,n) = Q\left(\frac{nC(\rho) - k + (\log n)/2}{\sqrt{nV(\rho)}}\right)$$

$$C(\rho) = \log(1+\rho)$$

$$V(\rho) = \rho\frac{(2+\rho)}{(1+\rho)^2}(\log e)^2$$

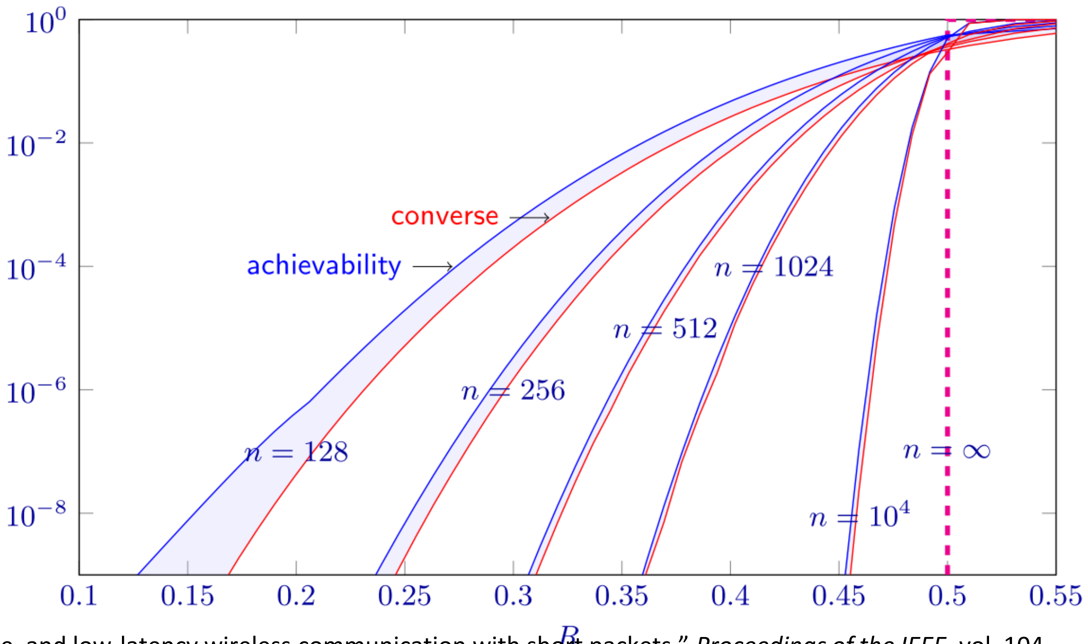


Y. Polyanskiy, H. V. Poor, and S. Verdu ́, “Channel coding rate in the finite blocklength regime,” IEEE Trans. Inf. Theory, vol. 56, no. 5, pp. 2307–2359, May 2010.

Normal Approximation for Binary Input AWGN Channel

$$R^*(n,\epsilon) = \underbrace{C - \sqrt{\frac{V}{n}}Q^{-1}(\epsilon)}_{\text{normal approximation}} + \mathcal{O}\left(\frac{\log n}{n}\right), \quad \underbrace{V = \mathbb{V}\mathrm{ar}[\imath_1(X_1,Y_1)]}_{\text{channel dispersion}}$$

- For long packets, the communication channel is modeled as a“bit pipe” that delivers reliably *C* bits per channel use. This holds under the assumption that good channel codes are used.
- For short packets, the communication channel can be thought of as a bit pipe of randomly varying size. Specifically, the size of the bit pipe behaves as a Gaussian random variable with mean *C* and variance V/n. Hence, V is a measure of the channel dispersion.
- the packet error probability is the probability that the maximal coding rate is larger than the size of the bit pipe.

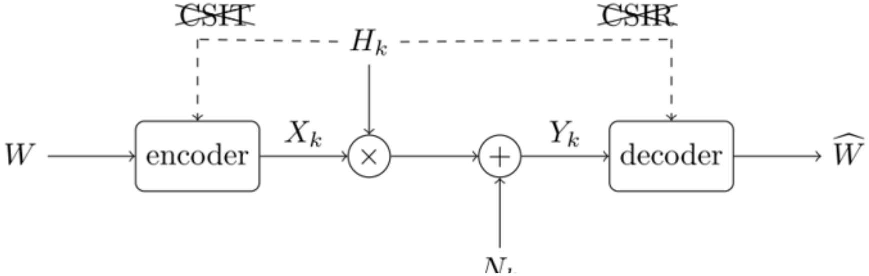


G. Durisi, T. Koch, and P. Popovski, “Toward massive, ultrareliable, and low-latency wireless communication with short packets,” *Proceedings of the IEEE*, vol. 104, no. 9, pp. 1711–1726, 2016

Binary-Input Fading Channel

The received signal at each channel use:

$$\mathbb{Y}_k = \mathbb{X}_k \mathbb{H}_k + \mathbb{W}_k$$

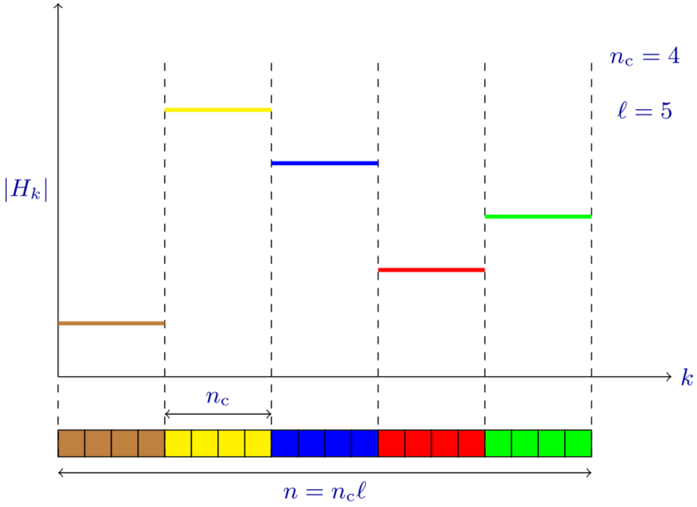


Where:

$$\mathbb{X}_k \in \mathbb{C}^{n_c \times m_t}$$

$$\mathbb{Y}_k \in \mathbb{C}^{n_c \times m_r}$$

$$\mathbb{H}_k \in \mathbb{C}^{m_t \times m_r}$$

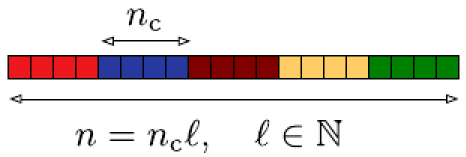


- Low latency and/or sporadic transmission $\Rightarrow$ no CSI
- Imperfect CSI can be acquired by pilot transmission and by exploiting reciprocity
- Performance depends on how $\{H_k\}$ varies within the packet

Slow Fading Channel

$$\mathbb{Y}_k = \mathbb{X}_k \mathbb{H}_k + \mathbb{W}_k \qquad \mathbb{X}_k \in \mathbb{C}^{n_c \times m_t} \qquad \mathbb{Y}_k \in \mathbb{C}^{n_c \times m_r} \qquad \mathbb{H}_k \in \mathbb{C}^{m_t \times m_r}$$

$$\ell = 1$$



Capacity Versus Outage at Finite Blocklength:

$$P_{out}(R) = \mathbb{P}[\log(1 + |H|^2 \rho) < R]$$

$$C_\epsilon = \sup\{R : P_{out}(R) \leq \epsilon\}$$

$$R^*(n, \epsilon) = C_\epsilon + \mathcal{O}\left(\frac{\log n}{n}\right)$$

$$\epsilon \approx \mathbb{E}\left[Q\left(\frac{C_\epsilon(\rho|H|^2) + (\log n)/(2n) - R^*(n,\epsilon)}{\sqrt{V(\rho|H|^2)/n}}\right)\right].$$

The outage capacity does not capture the channel-estimation overhead. Consequently, **outage capacity** is an inaccurate performance metric when the **coherence interval n_c small**.

The fading coefficients can be learned at the transmitter and **outage events can be avoided altogether by rate adaptation**.

The outage capacity is indeed a meaningful performance metric for delay-constrained communication over slowly varying fading channels.

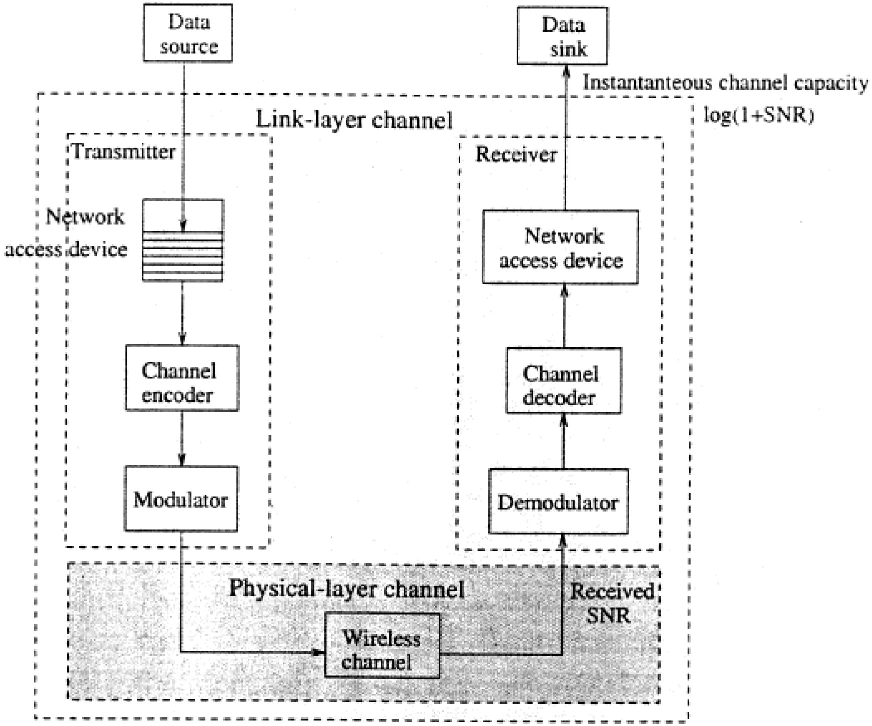
Agenda

1. Introduction
2. Finite Block-Length Information Theory
3. Effective Capacity/Bandwidth
4. Case Study
5. Summary

Effective Bandwidth / Capacity

Effective bandwidth is defined as the minimal constant service rate that is required to guarantee delay QoS requirement for a random arrival process *A(t)*.

Effective capacity is defined as the maximal constant arrival rate that can be served by a random service process *C(t)* subject to the delay QoS requirement.



D. Wu and R. Negi, “Effective capacity: a wireless link model for support of quality of service,” *IEEE Transactions on wireless communications*, vol. 2, no. 4, pp. 630–643, 2003.

Theory of Effective Capacity

$$\Lambda_A(\theta^*) + \Lambda_C(-\theta^*) = 0$$

The first term is related to arrival process, the second term is related to service process.

$$\lim_{x\to\infty} \frac{\log(\Pr\{q(\infty) \geq x\})}{x} = -\theta^*$$

The packet loss probability (The delay violation probability) can be approximated as

$$\epsilon = e^{-\theta^* x}$$

For constant capacity, i.e. constant service rate

$$\Lambda_C(-\theta^*) = \lim_{t\to\infty} \frac{1}{t} \log\left(e^{-\theta^* ct}\right) = -\theta^* c$$

From the first and last equations. To have a small packet loss probability, a capacity **(effectiv Bandwidth)** of the link equal to

$$\frac{\Lambda_A(\theta^*)}{\theta^*} = C.$$

is required, where the value of delay exponent comes from the unique solution of $\theta^* = -(\ln \epsilon)/x$

The derivation of the effective capacity will be as that of effective bandwidth, but for arrival process.

$$\Lambda_A(\theta^*) = \lim_{t\to\infty} \frac{1}{t} \log\left(e^{\theta^* \mu t}\right) = \theta^* \mu$$

Then, the **effective capacity** required for minimum violation probability is

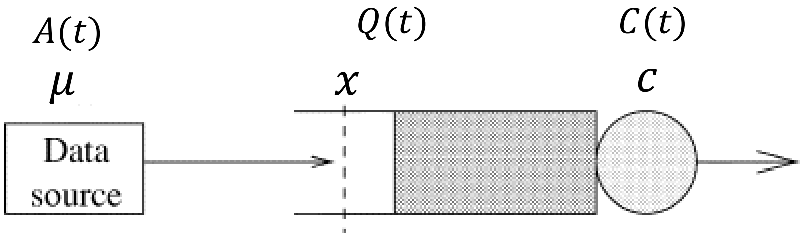
$$E_c(\theta^*) = -\frac{\Lambda_C(-\theta^*)}{\theta^*} = \mu$$

Now, an expression for the delay experienced by a packet at any time t can also be approximated as follows

$$\Pr\{D(t) > D_{\max}\} \approx \Pr\{q(\infty) > 0\} e^{-\theta^* \mu D_{\max}}$$

where the probability of non-empty buffer

$$\Pr\{q(\infty) > 0\} \approx \frac{\mathbb{E}[\mu(t)]}{\mathbb{E}[c(t)]}$$



Agenda

Case Study

The achievable transmission rate:

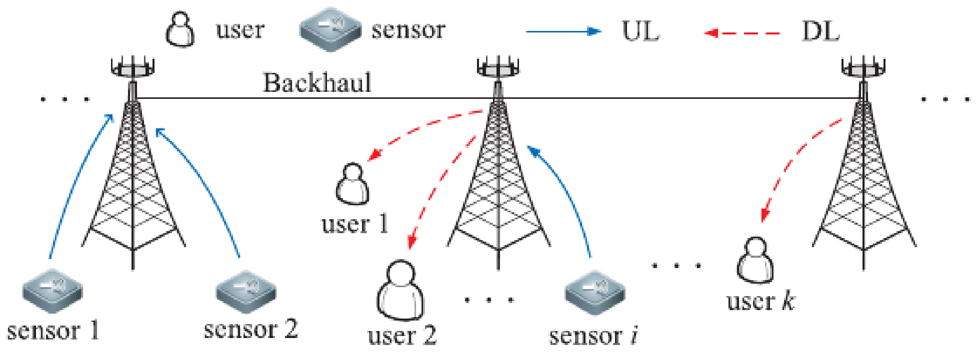
$$r_k^{\mathrm{d}} \approx \frac{\tau W_k^{\mathrm{d}}}{u \ln 2}\left[\ln\left(1+\frac{\alpha_k^{\mathrm{d}} g_k^{\mathrm{d}} P_k^{\mathrm{d}}}{\phi N_0 W_k^{\mathrm{d}}}\right) - \sqrt{\frac{1}{\tau W_k^{\mathrm{d}}}}\, Q_{\mathrm{G}}^{-1}\left(\varepsilon^{\mathrm{c,d}}\right)\right] \quad \text{(packets/frame).}$$

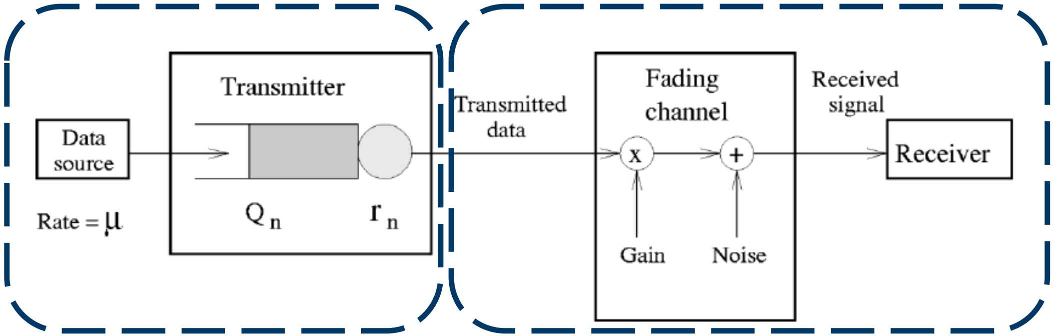
The effective bandwidth:

$$E_k^{\mathrm{B}} = \frac{T_{\mathrm{f}} \ln\left(1/\varepsilon^{\mathrm{q,d}}\right)}{D^{\mathrm{q,d}} \ln\left[\frac{T_{\mathrm{f}} \ln\left(1/\varepsilon^{\mathrm{q,d}}\right)}{\mu_k D^{\mathrm{q,d}}} + 1\right]}$$

The required SNR:

$$\gamma_k^{\mathrm{d}} = \exp\left[\frac{E_k^{\mathrm{B}} u \ln 2}{\tau W_k^{\mathrm{d}}} + \frac{Q_{\mathrm{G}}^{-1}\left(\varepsilon^{\mathrm{c,d}}\right)}{\sqrt{\tau W_k^{\mathrm{d}}}}\right] - 1$$





Queueing Model

Communication Model

Agenda

1 • Introduction

2 • Finite Block-Length Information Theory

3 • Effective Capacity/Bandwidth

4 • Case Study

5 • Summary

Summary

- The URLLC is part of real-time communications, it depends on **short packtets transmission.**
- Shannon formula for the transmission rate limit will not suitable, and a new formula must be used.
- New bounds and apporoximations were derived in the literature for short packets and were founded the so called **short blocklength information theory.**
- To study the delay constraint systems and applications, **an effective capacity** was also proposed in the literature as complement component for the achievable rate of the short packet transmission.
- A real-time RRM optimization will be fundamentally based on the **effective rate**, **i.e., the transmission rate based on effective bandwidth.**

Thank you for your attention!

Najib Odhah, Eckhard Grass, and Rolf Kraemer

IHP – Innovations for High Performance Microelectronics
Im Technologiepark 25
15236 Frankfurt (Oder)
Germany
Phone: +49 (0) 335 5625
Fax: +49 (0) 335 5625
Email: {odhah, grass, kraemer}@ihp-microelectronics.com

www.ihp-microelectronics.com

innovations
for high
performance
microelectronics

Leibniz
Association