Investigation of the competitive nature of eMBB and mMTC 5G services in conditions of limited communication resource

Viacheslav Kovtun1 & Krzysztof Grochla2*

The main contribution of the investigation is the Markov model of the process of resource allocation management between subscribers of eMBB and mMTC services within the 5G cluster. The proposed model considers the organization of the channel resource in the format of resource blocks. The presented model allows to estimate the average duration of IoT sessions, the average number of active multimedia/IoT sessions, the average number of channel resource units occupied by multimedia/IoT traffic, the average number of resource blocks occupied by multimedia/IoT traffic. The metrics are generalized by three management schemes of the investigated process: balanced, competitive and perspective. The first and third schemes enable static/dynamic distribution of channel resources into reserved and common segments for subscribers of eMBB and mMTC services. The proposed model is illustrated with an example showing how to assess the availability and efficiency of channel resource use of the 5G cluster of the cyber-physical system of the Situation Center of the Department of Information Technology of Vinnytsia City Council (Vinnytsia, Ukraine). The article also shows how to use the proposed model to select the 5G network parameters to keep the probabilities of rejection of multimedia and IoT requests below a set threshold.

The ubiquity of wireless networks’ presence and proliferation of cyber-physical systems increases the need for careful evaluation of reliability. While the number of IoT devices we use is growing, the complexity of interactions between different elements of wireless networks also increases. This may lead to unexpected failures with a hard to identify root cause. For example, a 5G video conference does not start or is prematurely interrupted because ten “smart” refrigerators of your neighbors in the agglomeration have decided to order food delivery on schedule. This hyperbole seems hypothetical only until we consider that one of the active sources of “Big Data” is sensor networks1–4.

Coexistence of heterogeneous services in 5G networks. The analysis shows that the complexity of IoT wireless networks will continue to grow. Experts at Juniper Research predict that by 2026 the global cellular Internet of Things (IoT) market will reach $61 billion. As a starting point: it is expected to be $31 billion this year. According to analysts, the introduction of 5G and LPWA technologies will be a prerequisite for the actual doubling of the IoT market over the next four years5–9. It is believed that the most significant influence will be provided by eMBB (enhanced Mobile BroadBand) and mMTC (massive Machine-Type Communication) services10–14. Low costs of equipment and connections for remote monitoring in key areas such as agriculture, smart cities and manufacturing will help attract new subscribers. According to forecasts, the number of 5G connections over the next four years will increase by 1200%. According to the GSMA, the number of cellular subscribers will increase to 5.7 billion by 2025 (70% of the world’s population). 5 billion people will use the mobile Internet service. The GSMA estimates that by 2022, 176 operators in 70 markets worldwide have launched commercial 5G networks; mobile technologies and services in 2021 brought 5% of world GDP or $4.5 trillion. The GSMA predicts that this contribution will increase to $5 trillion by 2025.

A potential bottleneck in the global dominance of the 5G generation is that both humans and machines will simultaneously use the eMBB and mMTC services, the distribution of which will be limited to the licensed radio frequency band. This will inevitably lead to conflicts when connecting and supporting active information exchange sessions. The problem is exacerbated by the fact that people are increasingly seeking multimedia

1Vinnytsia National Technical University, Vinnytsia 21000, Ukraine. 2Institute of Theoretical and Applied Informatics, Polish Academy of Sciences, 44-100 Gliwice, Poland. *email: kgrochla@iitis.pl
content (e.g. the recent popularization of video conferencing and monitoring), while machines exchange relatively compact text messages. In addition, the number of IoT subscribers already exceeds the number of cellular subscribers by order of magnitude, and this trend will grow. We summarize that the issue of availability of competing in the 5G ecosystem services eMBB and mMTC is relevant and requires investigation. Dependability theory9,15–18 is a powerful, constantly evolving, branch of theoretical science. One of the attributes of dependability is availability12,19–22. The basic mathematical apparatus underlying the definition of this attribute of dependability includes the provisions of probability theory and mathematical statistics, random process theory, queuing theory, mathematical logic, graph theory, optimization theory and more.

**Dependability theory.** Dependability theory9,15–18 is a powerful, constantly evolving, branch of theoretical science. One of the attributes of dependability is availability12,19–22. The basic mathematical apparatus underlying the definition of this attribute of dependability includes the provisions of probability theory and mathematical statistics, random process theory, queuing theory, mathematical logic, graph theory, optimization theory and more.

A notable element of the methodological apparatus of the theory of availability assessment, in particular, information and communication systems, is the structural-logical method12,19–22. The method describes the investigated system as a topology of interacting elements (devices, software services, persons), the set of which uniquely identifies the original object of investigation. Analytically, the relationships in the topology are characterized by the corresponding functions of the algebra of logic. The same functions in the transition to probabilistic or deterministic structural-logical models become the basis for formalizing the criteria for identifying the set of states of the investigated system. Quantitative characteristics of availability are determined based on the structural-logical model of the investigated system by replacing the minimum disjunctive normal form of logical functions with probabilistic or deterministic characteristics with the simultaneous transition from logical to arithmetic operations with them. Graphical interpretation of structural-logical models is the corresponding schemes of functional integrity20–23, depicted in the form of block diagrams, trees of faults or (and) events, ‘the bow tie’ techniques and so on. An algebra of groups of incompatible events in the paradigm of the general logical-probabilistic method12,19–22 has been developed to remove the binary constraint (the element of the model is in only one of the two defined states) characteristic of structural-logical models. Note also that the method of analysis of such trees is the basis for the formation of dynamic trees (which have a wide range of logical operators) or generalized trees (possible further conversion to Bayesian networks, Petri nets, etc.) However, these add-ons are not used in practice, due to their excessive complexity in both implementation and interpretation of the initial results. But, in the opinion of the authors, the analysis of generalized trees in the context of availability assessment seems promising due to the possibility of their direct integration with certain methods of machine learning.

**Markov chains and queuing system models.** Petri nets, Markov chains and queuing systems can serve as potential bases for modelling processes in information and communication systems. Let’s take a look at each of these approaches. Petri nets are used to model systems with discrete events, which allows for joint analysis of the behaviour and structure of these systems, improving their understanding. Petri nets allow us to describe only the logical structure of systems since they do not explicitly take into account time. Although Petri nets are widely used in the context of modelling and simulating systems with discrete events, their application to describe stochastic processes (including in information and communication systems and networks) is still methodologically limited. Markov chains are a mathematical modelling paradigm that describes systems as a stochastic process with discrete states and their order of alternation. Typically, time in such a model is a parameter, either discrete or continuous. Usual, the future state in a Markov chain does not depend on past states, but only on the current state. Markov models are used because, in addition to their simplicity, they allow analyzing the performance of various implementations of the studied system using modelling methods. Markov chains in time are stochastic processes that describe actions ending in events that generate state transitions. Queuing theory is an analytical method that uses mathematical equations to understand system problems and aims to identify and evaluate metrics that express the performance or functionality of the studied system. The queuing system can be represented by different models, but the basic process has common elements. The process begins with customers looking for a service and arriving randomly over time, forming a queue. At some point, one of the clients is called to be served according to the rule indicated in the discipline queue. The client is served using the service mechanism and then leaves the queuing system. The specification of the queue model usually requires that the performance characteristics be explicit. Only in this case, it is possible to calculate and evaluate the results for the performance of the studied system based on such properties as, for example, service time or the number of clients served in a certain time. By getting inadequate values, you can make the queuing system more efficient, avoiding the identified performance issues. The use of specific mathematical theories should be determined by the specifics of the investigated problem. Stochastic Petri nets do not have an explicit structure, as in queuing theory, but have a graphical notation, and models have a formal interpretation. Although the queuing theory facilitates the construction of the model, it has significant limitations, in particular the lack of a formal interpretation. As a result, despite the relatively high computational complexity, the authors chose the basis of Markov chains for the implementation of the research.
assessment of the availability of the investigated system with the relaxation time of the Markov chain. However, the obvious theoretical orientation of the scientific results presented in these researches is a limitation for their application. Also, the final parametric space used in these researches characterizes the investigated processes in terms of probability theory and mathematical statistics, rather than the theory of dependability. Finally, the description of the investigated process in discrete time presented in the mentioned research introduces a clear inaccuracy in the process of evaluating availability indicators.

Recent works on performance modeling of 5G. The performance evaluation using analytical models of 5G services interoperability in limited communication resource availability conditions is subject to ongoing research, with tens of works published within the last three years. Nevertheless, a recent survey on slice admission control strategies and optimization schemes shows that many works still use very simplistic slice request queuing methods and simple models, such as M/M/1 or M/M/1/N. Multiple novel models have been proposed; for example, a recent work by Mendis et al. demonstrates the use of multi-dimensional Markov models to assess the transient performance of network slicing for resource allocation with and without traffic priority. In 2020 Ginige et al. investigated ultra-reliable low-latency (URLLC) service and conventional enhanced mobile broadband (eMBB) in a smart-grid environment, formalizing a model developed using stochastic geometry and queuing theory. The paper proposes an analytical framework for the calculation of the admission ratio, modelling the occupancy distribution in both computational and communicational domains of the network resources. However, the model is presented on a system level and does not take into consideration the radio resource allocation. The main limitation of the above works is the models' narrow character. The models either represent a system as a whole, without modelling radio resources allocation or are concentrated on a single performance metric. To address this limitation, we develop a model of a 5G cluster, which allows us to estimate a large set of performance metrics. The model considers the duration of IoT sessions, the division of active sessions between multimedia and IoT, the number of channel resource units occupied, and the average number of resource blocks occupied by each traffic type. Additionally, while most of the work on analytical models of 5G services uses simulation for model validation, we evaluate the correctness of the proposed approach using real-life data from an operational network.

Motivation and objectives of the work. Considering the strengths and weaknesses of these analogues, we formulate the necessary attributes of scientific investigation. The object of investigation is the process of resource allocation management between subscribers of eMBB and mMTC services within the 5G cluster. Motivation The mechanism for efficient distribution of a limited amount of resources is the cornerstone of information-communication systems. It is the operation of this mechanism that largely ensures the achievability for systems of this type of such characteristics as availability, reliability, survivability, and functional safety. Objective confirmation of the stated thesis may be, for example, that for the 5G platform, scientists and engineers have already proposed such versions of the implementation of the mentioned mechanism as:

- Orthogonal Multiple Access (OMA) (versions: time division multiple access (TDMA) and orthogonal frequency division multiple access (OFDMA)).
- Non-orthogonal multiple access (NOMA) in the Power Division NOMA (PD-NOMA).
- Multiple access with interleaving and division (Interleave-division Multiple Access, IDMA).
- Multiple access with resource distribution (Resource Spread Multiple Access, RSMA).
- Code division multiple access (CDMA) with the extension of Low-Density Spreading (LDS).
- Multiple access with a sparse code (Sparse Code Multiple Access, SCMA).
- Multi-user shared access (Multi-User Shared Access, MUSA).
- Multiple access with pattern division (Pattern Division Multiple Access, PDMA).

All of these methods are based on the same key concept when more than one user is served in each orthogonal resource. Such a resource can be, for example, a time interval, a frequency channel, a spreading code, or an orthogonal spatial degree of freedom. At the same time, these methods are created with a focus on guaranteeing the equality of users. This approach justifies itself perfectly if the network is used by subscribers of the same class. At the same time, the introductory section of the article focuses on the fact that more and more often the interests of two fundamentally different classes of subscribers clash in public access networks—people and IoTs. A clear mathematical formalization of this feature (including the mathematical model of the studied process, a qualitative metric based on it, and the formulation of the optimization problem for one or more parameters from this metric) is extremely relevant and is investigated in a series of publications by the authors, which includes this article.

The subject of investigation is the provisions of the theory of Markov processes, probability theory and mathematical statistics and the theory of experimental planning. The aim of the investigation is the analytical formalization of the process of resource allocation management between subscribers of eMBB and mMTC services within the 5G cluster in the metrics of indicators of availability and efficiency of channel resource use. The objectives of the investigation are 1. Analytical formalization of the Markov process of resource allocation management between subscribers of eMBB and mMTC services within the 5G cluster; 2. Analytical formalization of the computationally efficient concept of calculation of characteristic parameters of metrics of indicators of availability and efficiency of use of channel resource; 3. Demonstration of the relationship between the indicators of the proposed metrics and resource allocation management schemes in the ecosystem of the target 5G cluster; 4. Substantiation of the adequacy of the created mathematical apparatus and analysis of experimental results.
The main contribution of the investigation is the Markov model of the process of resource allocation management between subscribers of eMBB and mMTC services within the 5G cluster. In contrast to the existing ones, the proposed model, in particular, takes into account the organization of the channel resource in the format of resource blocks. The presented model allows to estimate in the time domain the values of such characteristic parameters as the average duration of IoT sessions, the average number of active multimedia/IoT sessions, the average number of channel resource units occupied by multimedia/IoT traffic, the average number of resource blocks occupied by multimedia/IoT traffic.

The model of the process of resource allocation management between eMBB and mMTC subscribers within the 5G cluster, the method of computationally efficient calculation of its parameters, the metric for assessing the availability and efficiency of channel resource use of the target 5G cluster are the main highlights of the investigation.

Models and methods

Setting of the research. We present a typical info-communication 5G cluster. The key active services in it will be eMBB and mMTC. The first one is focused on meeting the information needs of human users (hereinafter—YT) and provides improved broadband mobile multimedia access. The second one is focused on meeting the information needs of IoT users and provides the possibility of the potential simultaneous connection of a large number of such subscribers. Naturally, the 5G cluster operates in the provider-subscribed frequency sub-band (hereinafter—the channel resource) within the licensed frequency band. The realities of life stipulate that active subscribers of both mentioned services will apply for the use of the channel resource at the same time. Moreover, the situation when the number of channel resources will not be enough for all subscribers of both services is not probable, but determined. This circumstance is the basis for the investigation of the process of resource allocation management between subscribers of eMBB and mMTC services within the 5G cluster.

The 5G cluster is organized around a central base station. The latter one provides different coverage radii for eMBB and mMTC services—r₁, r₂ and r₃, respectively. Moreover, r₁ < r₂ < r₃. Suppose that in one resource block RB we can organize c IoT sessions, to support each of which b units of channel resource are used. Then the radio interface of the 5G cluster on the lineup allows us to support C = S · c IoT sessions, where S is the number of RBs subscribed by the provider.

We consider the eMBB-oriented input stream of YT requests to be Poisson with intensity ν. The duration of satisfaction of the YT request is a distributed exponentially stochastic quantity with mathematical expectations μ⁻¹. The minimum amount of resources required to support a YT session at each temporal interval up the line is equivalent to the number of resources that can support d IoT sessions. The load of the eMBB service is defined as α = ν · μ⁻¹.

We consider the mMTC-oriented input stream of IoT requests to be Poisson with intensity λ. The receipt of each IoT request is characterized by the exponentially distributed service duration characteristic of the parameter θ. The mMTC service load is defined as ρ = λ · θ. The number of channel resource units publicly available for YT and IoT traffic is defined as C_L = C - R_N and C_N = C - R_L, respectively, where R₁ and R₃ are the number of channel resource units potentially available by the MBB and mMTC services, respectively.

We formulate valid schemes for managing the distribution of channel resources between subscribers of eMBB and mMTC services. The balanced scheme BS assumes that the channel resource is divided in half into public and reserved segments, each of which is claimed by subscribers of eMBB and mMTC services: C_L = R₁, C_N = R₃. The competitive scheme CS assumes that the entire channel resource is publicly available to subscribers of eMBB and mMTC services: C_L = C_N = C, R₁ = R₃ = 0. A promising scheme PS is a compromise between BS and CS schemes and is analytically described as: C_L = C - R_N, C_N = C - R_L > 0 with R₁ > 0, R₃ > 0. Note that this version of the PS scheme is rational for a general-purpose 5G cluster. If we talk about the 5G cluster as an element of critical infrastructure, then the PS scheme is defined as follows: C_L = C - R_N > 0, C_N = C - R_L with R₁ > 0, R₃ > 0.

The generalized graphic notation for the described process is presented in Fig. 1, where the variables n and m characterize the number of active YT and IoT sessions, respectively.

Let’s interpret Fig. 1. Assume that the 5G cluster base station does not support any sessions. An IoT request is received at its input. The IoT session is activated, for which RB is allocated. To support this session, b units of channel resources are allocated. If in the process of information interaction within the IoT session all available resources in RB are used, the next RB is activated.

The model of investigated process and the metric of its characteristic parameters. We present the process of resource allocation management between subscribers of eMBB and mMTC services within the 5G cluster in the mathematical apparatus of queuing systems. The input of the system receives two streams with different schemes (BS, CS, PS) for the allocation RB. It is necessary to provide for the possibility of rejections of both YT and IoT requests.

We define the state space of the created queuing system. YT request is accepted if at the time of its receipt at the disposal of the base station is not less than d ∈ C_L = −R_N units of free channel resource. If this condition is met, the YT request is accepted and the corresponding session is activated. In this case, the amount of free channel resources for the eMBB service is reduced by d units and the number of available RB is reduced by [d/c]. If at the time of receipt of the IoT request in the current RB there are free channel resources over the reserved b units, the IoT session is activated based on the resources of this RB. Otherwise, if the number of available for the mMTC service RB is exhausted, the incoming IoT request is lost, otherwise, the IoT session is activated due to the available channel resources of the new RB.
Let the functions \( n(t) \) and \( m(t) \) characterize the number of active YT and IoT sessions at a time \( t \), respectively. In this case, the base station operation process is described by a two-dimensional stochastic Markov process \( X = \{n(t), m(t), t > 0\} \) with the state space

\[
\chi = \{n \geq 0, m \geq 0 : c(m) \leq C - R_L, \quad nd \leq C - R_N, c(m) + nd \leq C\}, \tag{1}
\]

where \( c(m) = \lceil bm/M \rceil \) is the number of channel resource units occupied by IoT sessions, \( M = \lceil c/b \rceil \) is the number of IoT sessions (channel resource units) that can be serviced in any RB.

As shown in Fig. 2 UML state diagram graphically characterizes the process \( X \), where the sets \( B_L \) and \( B_N \) contain the characteristics of the process \( X \), in which the rejection of YT and IoT requests due to lack of required free channel resource or exceeding the maximum available amount of RB (hereinafter—rejection states).

We introduce the function \( p(n, m) \), which characterizes the stationary probability distribution that at a time \( t \) in the ecosystem of the 5G cluster \( n \) YT and \( m \) IoT sessions are active:
\[
p(n, m) = \lim_{t \to \infty} p(m, n, t), \quad \{n, m\} \in \chi,
\]

(2)

The stationary probability distribution of the stochastic process \(X\) satisfies the following equations of balance:

\[
\begin{align*}
\lambda p(n, m - 1) &= \frac{c(m)}{M} p(n, m), \quad n > 0, \\
\nu p(n - 1, m) &= \mu n p(n, m), \quad m > 0,
\end{align*}
\]

(3)

Express from the system (3) the function \(p(n, m)\) for \(m > 0\):

\[
p(n, m) = \frac{\lambda \theta}{c(m)} p(n, m - 1) = \frac{\rho}{c(m)} \times p(n, m - 1)
\]

\[
= \frac{\rho^2}{c(m)c(m - 1)} p(n, m - 2)
\]

\[
= \ldots = \frac{\rho^m}{c(m)c(m - 1) \ldots c(1)} p(n, 0).
\]

(4)

Substitute \(c(m) = [bm/M], M > 0\), into expression (4). We get:

\[
p(n, m) = \frac{\rho^m}{(bM)^m} \left( \frac{c(m)}{M} \cdot \frac{c(m - 1)}{M} \cdot \ldots \cdot \frac{1}{M} \right)
\]

\[
\times p(n, 0) = \left( \frac{\rho}{bM} \right)^m \frac{\rho}{M^n} = \sum_{(n, m) \in \chi} p(n, m), \quad (n, m) \in \chi
\]

(5)

Substitute expressions (4) and (5) into (3). We get:

\[
p(n, m) = \left( \frac{\rho}{bM} \right)^m \frac{1}{\prod_{i=1}^{m} \left( \frac{1}{m} \right)} \frac{1}{G(\chi)}, \quad (n, m) \in \chi
\]

(6)

where \(G(\chi) = \sum_{(n, m) \in \chi} \rho \frac{1}{\prod_{i=1}^{m} \left( \frac{1}{m} \right)} \frac{1}{G(\chi)} = \text{const}, n > 0\).

Expression (6) exhaustively characterizes the investigated process \(X\), however, does not determine the probability of rejections of YT and IoT requests \(p_L = \sum_{(n, m) \in \chi} p(n, m)\) and \(p_N = \sum_{(n, m) \in \chi} p(n, m)\), respectively. Let's get rid of this restriction.

We describe the set of rejection states \(B_L, B_N:\)

\[
B_L = \{(n, m) \in \chi : c(m, R_N) < d(n + 1)\}
\]

(7)

\[
B_N = \{(n, m) \in \chi : c(m, R_N) < c(m + 1)\}
\]

(8)

In the context of expressions (6) and (8), we define the average duration of the IoT session \(T_N\) as

\[
M[T_N] = \frac{1}{1 - R_N} \sum_{(n, m) \in \chi} mp(n, m)
\]

(9)

For the development of the mathematical apparatus, it is necessary to formalize analytical expressions for calculating the probabilities of rejections \(p_L\) and \(p_N\) taking into account expressions (7) and (8), respectively. To do this directly, for the primary states space \(\chi\), is impractical from the standpoint of computational efficiency. Let us try to circumvent this limitation by clustering the state space \(\chi\) into clusters—"superstates" \(B_L, B_N\) and \(B_L, B_N\) in the form \(\chi = \bigcup_{s=0}^{S} \chi_s\), where \(\chi_s = \{ (n, m) \in \chi : c(m) = cs \}\). In the context of this concept, based on the stationary distribution of the states of the investigated system represented by expressions (5) and (6), we define the superstates \(B_L\) and \(B_N\) as.
\[ B_L = \bigcup_{s=0}^{S} \{ \chi : C - R_N < C - d(n+1) \cup c(m) < d(n+1) \} \]

\[ = \bigcup_{s=0}^{S} \{ \chi : C - \max(c(m), R_N) < d(n+1), c(m) = cs \} \]

\[ = \bigcup_{s=0}^{S} \{ \chi : c : \left\lfloor \frac{1}{d} (C - \max(cs, R_N)) \right\rfloor < n+1, s = \left\lfloor \frac{bm}{c} \right\rfloor \} \]

\[ = \bigcup_{s=0}^{S} \left\{ \chi : n = \left\lfloor \frac{1}{d} (C - \max(cs, R_N)) \right\rfloor, s = \left\lfloor \frac{bm}{c} \right\rfloor \right\} \]

\[ = \left\{ \chi : n = \left\lfloor \frac{C - R_N}{d} \right\rfloor, m = 0 \right\} \]

\[ \cup \bigcup_{s=1}^{\infty} \left\{ \chi : n = \left\lfloor \frac{C - \max(cs, R_N)}{d} \right\rfloor, s > \left\lfloor \frac{bm}{c} \right\rfloor \right\} \]

\[ B_N = \{ \chi : C - R_L < C - c(m+1) \cup dn < c(m+1) \} \]

\[ = \bigcup_{s=0}^{S} \{ \chi : C - \max(dn, R_L) < c(m+1), c(m) = cs \} \]

\[ = \bigcup_{s=0}^{S} \left\{ \chi : c : \left\lfloor \frac{1}{d} (C - \max(cs, R_N)) \right\rfloor \leq n - 1, cs = \left\lfloor \frac{bm}{c} \right\rfloor \right\} \]

\[ = \bigcup_{s=0}^{S} \left\{ \chi : n = \left\lfloor \frac{C - \max(cs, R_N)}{d} \right\rfloor \leq n, m = \left\lfloor \frac{cs}{b} \right\rfloor \right\} \]

Based on expressions (10) and (11) we express analytically the required probabilities of rejection of YT and IoT requests:

\[ p_L = p\left(0, \left\lfloor \frac{C - R_N}{d} \right\rfloor \right) + \frac{c-R_L}{c} \sum_{m=M_s(M_s+1)}^{M_s} p\left(m, \left\lfloor \frac{C - \max(cs, R_N)}{d} \right\rfloor \right), \]

\[ p_N = \sum_{n=0}^{\infty} p\left(n, M_s \left\lfloor \frac{C - R_L}{c} \right\rfloor \right) + \sum_{n=0}^{\left\lfloor \frac{C - R_N}{d} \right\rfloor} \sum_{m=M_s}^{M_s} p(n, m). \]

Substitute expression (11) into expression (9) and determine the average duration of the IoT session:

\[ M[T_N] = \frac{\frac{c-R_L}{d} \left\lfloor \frac{C - \max(cs, R_N)}{d} \right\rfloor \sum_{n=0}^{\left\lfloor \frac{C - R_N}{d} \right\rfloor} \sum_{m=0}^{m} p(n, m)}{\lambda (1 - p_N)} \]

Determine the average number of YT and IoT sessions in the 5G cluster (\(N_L\) and \(N_N\), respectively):

\[ M[N_L] = \sum_{(n,m) \in \chi} np(n, m) = \sum_{m=0}^{\left\lfloor \frac{C - R_N}{d} \right\rfloor} \sum_{n=0}^{n} np(n, m) \]

\[ M[N_N] = \sum_{(n,m) \in \chi} mp(n, m) = \sum_{m=0}^{\left\lfloor \frac{C - R_N}{d} \right\rfloor} \sum_{n=0}^{n} mp(n, m) \]

Note that the obtained mathematical apparatus allows formulating expressions for direct calculation of the number occupied by YT traffic RB and units of channel resource. However, IoT devices are characterized by the extremely dynamic nature of information interaction. We take this fact into account when determining the following characteristic parameters of the eMBB and mMTC services.

Determine analytically the average number of units of channel resource occupied by YT and IoT traffic (\(b_L\) and \(b_N\), respectively):
\[ M[b_L] = d \sum_{m=0}^{[C - R_N]} \left\lfloor \frac{c - R_L}{m} \right\rfloor \sum_{n=0}^{[\frac{[C - R_N]}{d}]} np(n, m) \quad (17) \]

\[ M[b_N] = M \sum_{m=0}^{[C - R_L]} \left\lfloor \frac{c - R_N}{m} \right\rfloor \sum_{n=0}^{[\frac{[C - R_N]}{d}]} \left\lceil \frac{m}{M} \right\rceil p(n, m) \quad (18) \]

Determine analytically the average number of RB occupied by YT and IoT traffic \((s_L\) and \(s_N\), respectively):

\[ M[s_L] = \sum_{m=0}^{[C - R_L]} \left\lfloor \frac{c - R_N}{m} \right\rfloor \sum_{n=0}^{[\frac{[C - R_N]}{d}]} np(n, m) \quad (19) \]

\[ M[s_N] = \sum_{m=0}^{[C - R_L]} \left\lfloor \frac{c - R_N}{m} \right\rfloor \sum_{n=0}^{[\frac{[C - R_N]}{d}]} m p(n, m) \quad (20) \]

Finally, the average number of occupied units of channel resource in the 5G cluster \(b_{L+N}\) is defined as

\[ M[b_{L+N}] = M[b_L] + M[b_N] \quad (21) \]

Thus, the mathematical apparatus generalized by expression (6) characterizes a load of eMBB and mMTC services in the 5G cluster. The derivative result in expressions (12) and (13), which allow calculating the probabilities of rejection of incoming YT and IoT requests, respectively. The effectiveness of the targeted use of these services is determined by the metric of indicators, which includes: – the average duration of the IoT session \(M[T_N]\); – the average number of channel resource units occupied by YT and IoT traffic—\(b_L\) and \(b_N\), respectively; – the average number of RB occupied by YT and IoT traffic—\(s_L\) and \(s_N\), respectively.

**Experiments**

As an example, we use the mathematical apparatus presented in section “Models and methods” to estimate the load of eMBB and mMTC services of one of the 5G clusters of the cyber-physical system of the Situation Center of the Information Technology Department of Vinnytsia City Council (Vinnytsia, Ukraine) (hereinafter—SC system). This info-communication system was put into operation in 2018 and is constantly evolving in the direction of improving the implemented services and adding new ones. Currently, the SC system manages traffic lights on the roads of Vinnytsia and maintains the uninterrupted operation of the data centre, which stores video streams from more than 1 000 video cameras located in the city.

Collected in the system confidential information is assessable only for authorized employees of the Vinnytsia City Council, the National Police of Ukraine, the Security Service of Ukraine, etc. For these privileged persons to have operational access to the relevant information, a local network was created consisting of data centre servers, communication equipment, workstations and software. In normal operation, this LAN is not isolated from the World Wide Web. However, the processing, storage and audit of confidential information are carried out by a specialized relational database management system, ready for which is organized through a specialized web interface. Data, databases, management systems, web interface—all these components are located on dedicated servers.

The dedicated 5G cluster in a network structure of an SC system is an ideal object for testing the mathematical apparatus presented in section “Models and methods”. In it, the eMBB service supports multimedia streams from surveillance cameras and the mMTC service supports the operation of a sensor network, the sensors of which are located at traffic lights in the city. A detailed description of the network architecture used by the authors for experiments was made earlier in [7,13,32].

We summarized the content of the logs of the SC system for the period from 01.09.2021 to 16.09.2021 (15 full days) in the form of values of the set of established parameters for further simulation. Namely: – the number of units of channel resource in the 5G cluster \(C = 1000\); – the number of resource blocks RB \(c = 40\); – the number of channel resource units available to eMBB and mMTC services \(R_L = R_N = [0, 1000]\); – the number of channel resource units required to support the YT-session \(d = 10\); – the number of channel resource units required to support the IoT session \(b = 1\); – average speed for IoT session \(\theta = 100\) kbps; – average duration of YT-session \(\mu_{\text{YT}} = 10\) s; – the intensity of YT-requests \(\nu = 10\) p/min; – the intensity of IoT requests \(\lambda = 100\) p/min.

Based on expression (12), we estimate the dependence of the probability of YT request rejection on the number of active IoT sessions in the 5G cluster ecosystem. To estimate the number of active IoT sessions we use expression (16). We take into account that the principle of channel resource distribution between eMBB and mMTC subscribers is carried out according to the schemes generalized by the tuple \((BS, CS, PS)\), where \(BS : C_L = y R_L, C_N = y R_N\) is a balanced scheme, \(CS : C_L = C_N = C, R_L = R_N = 0\) is a competitive scheme, \(PS : C_L = C - R_N, C_N = C - R_L > 0\) with \(R_L > 0, R_N > 0, R_N \equiv y\) is a promising scheme. To control the distribution \(C_L vs. C_N \& R_L vs. R_N\) we introduce the balancing coefficient \(y \in [0, 1]\). The calculated dependencies \(p_L = f(N_N, \text{Scheme}, y)\) are visualized in Fig. 3.

We conducted an investigation symmetrical to what has just been described. Based on expression (14), we estimate the dependence of the probability of IoT request rejection on the number of active YT sessions in the...
5G cluster ecosystem. To estimate the number of active YT sessions we will use expression (15). We take into account that the principle of channel resource distribution between eMBB and mMTC subscribers is carried out according to the schemes generalized by tuple \( \langle BS, CS, PS \rangle \) taking into account the balancing coefficient \( 1 - \gamma \).

The calculated dependencies \( p_N = f(N_L, Scheme, \gamma) \) are visualized in Fig. 4.

Now we will investigate how the dynamics of growth in the number of IoT sessions affect the level of employment of the channel resource, which is represented by the coefficient \( \eta \). The coefficient \( \eta \) is determined based on expression (21):

\[
\eta = \frac{M_{BL+NN}}{C} \times 100\%.
\]

In the investigation, we take into account the possibility of changing the schemes of channel resource allocation within the tuple \( \langle BS, CS, PS \rangle \) and the fact that the number of active YT-sessions does not change: \( N_L = \text{const} \). The obtained results are presented in Fig. 5.

We conducted an investigation symmetrical to what has just been described. Let us investigate how the dynamics of growth in the number of YT sessions affect the value of the coefficient \( \eta \). In the investigation, we take into account the possibility of changing the schemes of channel resource allocation within the tuple \( \langle BS, CS, PS \rangle \) and the fact that the number of active IoT sessions does not change: \( N_N = \text{const} \). The obtained results are presented in Fig. 6.

We empirically solved the problem of finding such a value of the coefficient \( \gamma \), which would ensure the targeted operation of the eMBB and mMTC services without exceeding the specified threshold for the probabilities of rejection \( p_L \) and \( p_N \). In this case, we will not consider the CS scheme, because in its formulation the mentioned coefficient was not used. Let \( p_L = p_N = 1\% \). Define for the PS scheme the set \( \{ \gamma_{\min}, \gamma_{\max} \} \) taking into account the need to reserve the channel resource. Next, we take \( \gamma_{\min} = R_L, \gamma_{\max} = 1 - R_N \). For \( N_L = 50, N_N = 3 \cdot 10^4 \) we get \( \gamma_{\min} = 0.55, \gamma_{\max} = 0.7 \). That is, 55% of the channel resource should be given to support the subscribers of the eMBB service, and for the needs of the subscribers of the mMTC service to leave 30%. Thus, the task of

![Figure 3. Dependences \( p_L = f(N_L, Scheme, \gamma) \).](image1)

![Figure 4. Dependences \( p_N = f(N_L, Scheme, \gamma) \).](image2)
the PS scheme is the efficient distribution \( \Delta \gamma = \gamma_{\text{max}} - \gamma_{\text{min}} \) = 15% of the entire channel resource \( C \). For the BS scheme, we assume that \( \gamma \in [\gamma_{\text{min}}, \gamma_{\text{max}}] \).

Figure 7 presents the results of the calculation of the dependence \( p = p_L + p_N = f(\gamma) \) were, in turn, the value of \( \gamma \) change in the certain interval \([\gamma_{\text{min}}, \gamma_{\text{max}}]\) due to the increase in the number of YT-sessions, which was controlled by expression (17).

In turn, Fig. 8 presents the results of the calculation of the dependence \( \eta = f(\gamma) \), where the value of \( \gamma \) change in a certain interval \([\gamma_{\text{min}}, \gamma_{\text{max}}]\) due to the increase in the number of YT sessions, which was controlled by expression (17).

Finally, we conclude section "Experiments" with the answer to the question of the adequacy of the mathematical apparatus presented in section "Models and methods" to describe the processes of simultaneous operation of eMBB and mMTC services in the 5G cluster.

The obvious way to check this is to estimate the coincidence of the simulated and empirical values of the selected characteristic parameter of the target system using the appropriate statistical criterion. Characteristic parameters that are the easiest to detect for the SC system are the number of active YT and IoT sessions. These parameters are independent and distributed according to normal law. The proposed mathematical apparatus allows us to calculate these values using expressions (15) and (16), respectively. Analysis of the results of the SC system operation for the censored period allowed us to obtain a statistically representative sample \( (10^4 \text{ rows}) \) for the tuple \( \langle N_L, N_N, C, c, R_L, R_N, d, b, v, \lambda \rangle \), where the first two parameters are characteristic and the rest are stable (allow to calculate the values of characteristic parameters). Next, the obtained pairs of empirical and simulated values \( N_{\text{emp}}^L, N_{\text{mod}}^L \) and \( N_{\text{emp}}^N, N_{\text{mod}}^N \) were processed by the standard fisheries function from the specialized...
software package MATLAB Simulink. The test results showed that the variation of variances for both characteristic parameters is within the tabular values.

**Discussion**

Let’s consistently analyze the results presented in Figs. 3, 4, 5, 6, 7 and 8. Figure 3 shows the dependences of the values of the probability of rejection of the YT request from the number of active IoT sessions and the control scheme of channel resource allocation, which for the BS and PS schemes is carried out using the controlled coefficient γ. The CS scheme is freely competitive. The BS scheme is the simplest to implement and creates two independent virtual wireless communication systems in the radio space of the 5G cluster focused on eMBB and mMTC subscribers. This scheme was formulated with a focus on systems where the average load of the ordered services is quasi-stable and pre-predicted. The antagonist of this scheme is the CS scheme, where a freely available channel resource is potentially usurped by more dynamic and less resource-intensive IoT requests. It can be seen that with a small number of mMTC service subscribers, the CS scheme manages the process of connecting new eMBB service subscribers more efficiently than the BS scheme. However, with the increase in the number of connected subscribers of the mMTC service, the probability of rejection of the YT request steadily and almost linearly increases. Thus, the CS scheme should be used in infrastructure with projected low IoT traffic. Finally, the “semi-dynamic” PS scheme due to the possibility of dynamic control of the value of the coefficient γ is more promising than the BS scheme, however, inherits its functional nature, which manifests itself with an increasing number of active IoT sessions (when the PS scheme exhausts the channel resource reserve provided by managing the ratio between publicly available and reserved channel resource segments for YT sessions). Also, note that
the growth curve of the probability of rejection of the YT request for the PS scheme grows slower than the curve for the CS scheme for all values of the argument.

Let’s move on to the analysis of Fig. 4, which shows the dependence of the values of the probability of rejection of the IoT request from the number of active YT sessions and the control scheme for the allocation of channel resources. Everything we said in the address Fig. 3, is also true for Fig. 4. Note only that for the same values of the coefficient $\gamma \in \{0.6, 0.8\}$ the BS and PS schemes provide significantly smaller values of the argument. The growth dynamics of the curve CS are also slower. This can be supplemented by the fact that IoT requests are less resource-intensive. It should also be noted that when the IoT session is initiated, the remaining free units of the channel resource will be filled with IoT traffic. Note that presented in Figs. 3 and 4 the results demonstrate the representativeness of the presented mathematical apparatus for assessing the availability of the base station of the 5G cluster.

In two (BS and PS) of the three proposed control schemes, part of the channel resource is reserved for the needs of eMBB and mMTC services. This approach increases the availability of the system (which we have seen), but could potentially lead to inefficient use of prepaid channel resources. The proposed mathematical apparatus allows giving an analytical assessment of this characteristic in the form of the value of the coefficient $\eta$. Technologically, this investigation is similar to the previous one (implemented in Figs. 3 and 4).

Since the characteristics of “availability” and “efficiency of the use of channel resources” are antagonistic, the investigated schemes show a mirror image of qualitative indicators in comparison with the results of the analysis of the Figs. 3 and 4. The most aggressive for the use of channel resources the CS scheme shows the best dynamics in terms of increasing the value of the coefficient $\eta$ for both YT and IoT traffic. The graphs for the BS and PS schemes after nonlinear growth reach stability at a certain value of the coefficient $\eta$, but for both YT and IoT traffic, the more flexible PS scheme prevails. It should also be noted that for the 5G cluster of the investigated SC system, more resource-intensive YT traffic ultimately loads the entire prepaid frequency band, while IoT traffic for the selected range of argument values failed to do so. Thus, studies similar to those embodied in Figs. 5 and 6 allow to give mathematically rigorous and clear answers to the questions “Is there enough prepaid channel resources for the operation of the 5G cluster in the conditions formulated at the design stage?” and “Does the current access control scheme allow full use of the subscribed channel resource?” The answers to both of these questions are always relevant for the management of institutions and investors.

Finally, presented in Figs. 7 and 8 the results answer such an important question “Do the settings of the access control scheme allow to reach the specified level of availability in compliance with the design conditions of operation of the 5G cluster?” The controlled parameter in this investigation is the coefficient $\gamma$, which regulates the relationship between the reserved and public segments of the channel resource for eMBB and mMTC services. Figure 7 shows that the probability of rejection of IoT requests when using the BS scheme is equal to $5 \cdot 10^{-7}$, while for the PS scheme—$10^{-6}$, i.e. there is a slight advantage in favour of the first one. At the same time, the probability of rejection of YT request when using the BS scheme is equal to $10^{-3}$, while for PS scheme—$5 \cdot 10^{-4}$, i.e. PS scheme prevails in two orders of magnitude. However, the curve BS, $N_L = \text{const}$ is of more concern. It is seen that with the further growth of the argument the established restriction $p_L = p_N < 1\%$ will cease to be fulfilled. Last, the final argument in favour of the advantage of the PS scheme over BS is Fig. 8. The first one is superior to the second for the whole range of argument values. The extremum of the curve BS indicates the moment when the free channel resource in the general-purpose segment is exhausted.

Thus, the adequate mathematical apparatus presented in Section “Models and methods” proves to assess the availability and efficiency of using the 5G cluster channel resource and select the values of the controlled parameter of stable and progressive schemes so that the probabilities of rejection of YT and IoT requests in the target ecosystem do not exceed the threshold.

Conclusions
The main contribution of the investigation is the Markov model of the process of resource allocation management between subscribers of eMBB and mMTC services within the 5G cluster. In contrast to the existing ones, the proposed model, in particular, takes into account the organization of the channel resource in the format of resource blocks. The presented model allows to estimate in the time domain the values of such characteristic parameters as the average duration of IoT sessions, the average number of active multimedia/IoT sessions, the average number of channel resource units occupied by multimedia/IoT traffic, the average number of resource blocks occupied by multimedia/IoT traffic.

This metric of characteristic parameters is generalized by three schemes of management of the investigated process: balanced, competitive and perspective. The first and third of them provide for the possibility of static/dynamic distribution of channel resources into reserved and common segments for subscribers of eMBB and mMTC services. The article presents an example of using the created mathematical apparatus to assess the availability and efficiency of channel resource use of the 5G cluster of the cyber-physical system of the Situation Center of the Department of Information Technology of Vinnytsia City Council (Vinnytsia, Ukraine). The problem of such selection of the value of the controlled parameter of stable and progressive schemes so that the probabilities of rejection of multimedia and IoT requests in the target ecosystem do not exceed the set threshold is also solved.

The mathematical apparatus presented in the article is formulated for a closed 5G cluster system. The impact of other clusters of the same type, as well as wireless clusters based on other technologies, is not taken into account. Further research is planned to address this limitation.

Data availability
All the data is available on request due to restrictions e.g. privacy or ethics. Please contact the corresponding author, Krzysztof Grochla via email kgrochla@iitis.pl to request the data from this study.
36. Ojijo, M. O. & Falowo, O. E. A survey on slice admission control strategies and optimization schemes in 5g network. *IEEE Access* 8, 14977–14990 (2020).

**Acknowledgements**
The authors would like to thank anonymous reviewers who helped better present the research results.

**Author contributions**
V.K.: Concept, Design, Analysis, Writing—review & editing. K.G.: Concept, Design, Analysis, Writing—review & editing.

**Competing interests**
The authors declare no competing interests.

**Additional information**
Correspondence and requests for materials should be addressed to K.G.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher’s note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

![Open Access](https://creativecommons.org/licenses/by/4.0/)

This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit [http://creativecommons.org/licenses/by/4.0/](http://creativecommons.org/licenses/by/4.0/).

© The Author(s) 2022