Информационные процессы, Том 24, № 1, 2024, стр. 1–15.

ПЕРЕДАЧА ИНФОРМАЦИИ В КОМПЬЮТЕРНЫХ СЕТЯХ

Модель резервирования канального ресурса при обслуживании разнородного трафика в сети маломощных мобильных абонентских терминалов на базе КА с высокой пропускной способностью

А.А. Маслов[*], Г.В. Себекин[*], М.С. Степанов[**], С.Н. Степанов[,***], А.О. Щурков[*],**

[*] *Московский физико-технический институт (национальный исследовательский университет), Москва*

[**] *Московский технический университет связи и информатики, Москва*

[***] *Институт радиотехники и электроники им. В.А. Котельникова Российской академии наук (ИРЭ им. В.А. Котельникова РАН), Москва*

Поступила в редколлегию 01.03.2024 г. Принята 10.04.2024 г.

Аннотация—Резервирование канального ресурса в сетях маломощных мобильных абонентских терминалов на базе космических аппаратов с высокой пропускной способностью позволяет повысить эффективность реализации функций планирования и управления работой сетей за счет либо выравнивания потерь заявок разных типов либо создания преимуществ для передачи более важного трафика. Для оценки показателей качества предоставления сервисов в сетях с резервированием требуется разработка математических моделей, которые должны учитывать основные особенности сетей, включая совместное обслуживание нескольких типов трафика реального времени и эластичного трафика данных, неравномерность размещения терминалов по локальным зонам, ограниченное их количество, возможность одновременного обслуживания каждым терминалом заявок на передачу трафика двух типов, один из которых является эластичным трафиком данных, другой трафиком реального времени, ограничение на скорость передачи информации терминалом. Цель работы состоит в решении задачи моделирования с учетом перечисленных факторов. Построена модель на основе многомерных ступенчатых марковских процессов. Определен перечень задач, которые могут решаться с их использованием, включающий как задачи определения минимально необходимого ресурса сети для выполнения заданных требований по качеству, так и задачи отыскания предпочтительных распределений ресурса между локальными зонами по критерию минимизации среднего времени обслуживания заявок на передачу данных. Представлены численные примеры решения задач. Разработанные модели могут быть использованы для построения систем управления сетью передачи информации на базе космических аппаратов с высокой пропускной способностью.

КЛЮЧЕВЫЕ СЛОВА: космические аппараты с высокой пропускной способностью, канальный ресурс, мультисервисный трафик, трафик реального времени, эластичный трафик, мобильные абонентские терминалы, резервирование ресурса.

DOI: 10.53921/18195822_2024_24_1_1

1. ВВЕДЕНИЕ

Важным направлением развития спутниковых систем широкополосного доступа на базе космических аппаратов (КА) с высокой пропускной способностью на высокоэллиптической и геостационарной орбитах [1] является создание сетей маломощных мобильных абонентских терминалов (АТ), устанавливаемых на транспортные средства [2]. Проводимые работы

подтверждают возможности создания таких АТ, и в частности, антенн [3, 4]. Анализ и оценка показателей качества предоставления сервисов в таких сетях обычно проводятся с использованием теории телетрафика [5]. Обслуживание порождаемого при получении сервисов трафика осуществляется посредством выделения необходимого канального ресурса, который, как правило, выражается целым числом единиц скорости передачи информации (бит/с) [5]. При этом для создания преимуществ в передаче наиболее важных типов трафика или для выравнивания потерь заявок разных типов используется резервирование канального ресурса. В статье проводится разработка математического аппарата для оценки эффективности процедур резервирования в сети маломощных АТ (рисунок 1), описанной в [6] и обладающей следующими особенностями:

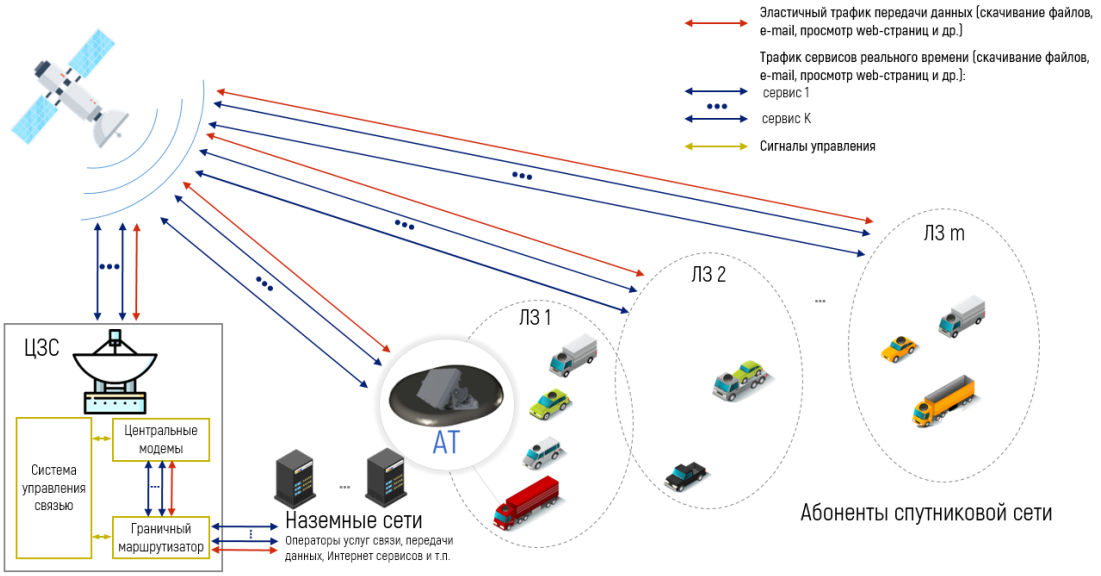

Рис. 1. Схема сети

- АТ, установленные на транспортные средства и предназначенные для получения водителями сервисов связи и передачи данных, неравномерно размещены в некотором количестве локальных зон (ЛЗ) обслуживания многолучевой антенны КА. Предоставление сервисов требует передачи трафика реального времени или эластичного трафика в зависимости от типа сервиса.
- Сеть имеет топологию «звезда». Доступ к сервисам осуществляется через центральную земную станцию (ЦЗС), выполняющую роль шлюза между спутниковыми каналами и наземными телекоммуникационными сетями. Определением возможности предоставления сервиса по заявке водителя и собственно назначением ресурса занимается программа планировщик из состава системы управления ЦЗС.
- Ввиду маломощности максимальная скорость передачи АТ ограничена, поэтому один АТ в процессе обмена информацией может использовать ресурс каналов ретрансляции лишь частично.
- АТ являются персональными. Поэтому можно допустить, что водители в каждый момент времени получают не более одного сервиса реального времени, хотя одновременно может осуществляться обмен данными (например, можно проводить телефонный разговор и скачивать файлы или просматривать web страницы и т. п.).

Качество функционирования сети в [6] оценивается такими показателями как доля заявок на передачу трафика каждого типа, потерянных из-за отсутствия свободного канального ресурса, а также среднее время обслуживания заявки на передачу данных. Правила резервирования ресурса, отрабатываемые планировщиком, традиционно включают принятие решения о приеме заявки на обслуживание исходя из количества уже обслуживаемых заявок такого типа и ограничение доступа заявок исходя из занятого канального ресурса [5].

Для решения задач планирования и распределения канального ресурса в таких сетях необходима разработка математических моделей обслуживания, позволяющих оценить значения показателей качества в зависимости от объема ресурса и его распределения между ЛЗ с учетом состава и характеристик получаемых сервисов, количества и интенсивностей работы АТ, их размещения по ЛЗ, принятых процедур предоставления и резервирования ресурса. Цель настоящей статьи состоит в разработке указанной модели, учитывающей перечисленные особенности сетей.

Моделированию процессов обслуживания мультисервисного трафика посвящено большое количество работ (например, [6–22]). Вопросы моделирования рассматриваются применительно к различным телекоммуникационным сетям и приложениям, например:

- к узлам доступа сетей мобильной связи (см. [7, 15–17]);
- к сетям Интернета вещей (см. [11, 12, 15, 18]);
- к системам дистанционного наблюдения (см. [7, 10]);
- к облачным компьютерным системам (см. [14]);
- к системам спутниковой связи (см. [6, 20–22]).

Общий подход, в т. ч., и при использовании процедур резервирования, предусматривает принятие предположения о пуассоновском характере входных потоков заявок и экспоненциальном распределении времен обслуживания и объемов передаваемых файлов, что позволяют строить модели на основе многомерных ступенчатых марковских процессов, которые учитывают основную специфику анализируемых систем. Необратимость этих процессов затрудняет получение аналитических результатов. Оценки показателей качества основываются на численном решении систем уравнений равновесия (СУР). Непосредственное использование результатов работ [5–22] невозможно, т. к. в них комплексно не учитываются факторы, связанные с использованием резервирования, неравномерным размещением АТ в совокупности ЛЗ, другими особенностями сети. Обслуживание потоков, генерируемых АТ нескольких ЛЗ, рассматривается в [6, 22] без учета резервирования. Для достижения поставленной цели в разделе 2 строится математическая модель совместного обслуживания мультисервисного трафика в сети. Характеристике задач, которые могут решаться с использованием разработанной модели, посвящен раздел 3. В разделе 4 приводятся численные примеры решения таких задач с использованием модели.

2. МОДЕЛЬ СЕТИ

Введем обозначения. Пусть ресурс, выделенный отдельной ЛЗ, составляет v единиц. Каждый АТ может одновременно обслуживать трафик одного из K сервисов реального времени и эластичный трафик данных (рисунок 2).

По аналогии с [6] состояние АТ обозначим парой (k,q), где k – номер предоставляемого сервиса реального времени от 1 до K или 0, q – единица или ноль в зависимости от того, обслуживается или нет в текущий момент трафик данных. Очевидно, общее количество состояний АТ равно $2(K+1)$. Пусть требуемые для обслуживания заявок объемы ресурса задаются вектором $(b_1,b_2,\ldots,b_K,b_e)$, где $b_k (k = 1, 2, \ldots, K)$ значения ресурсов для заявок на

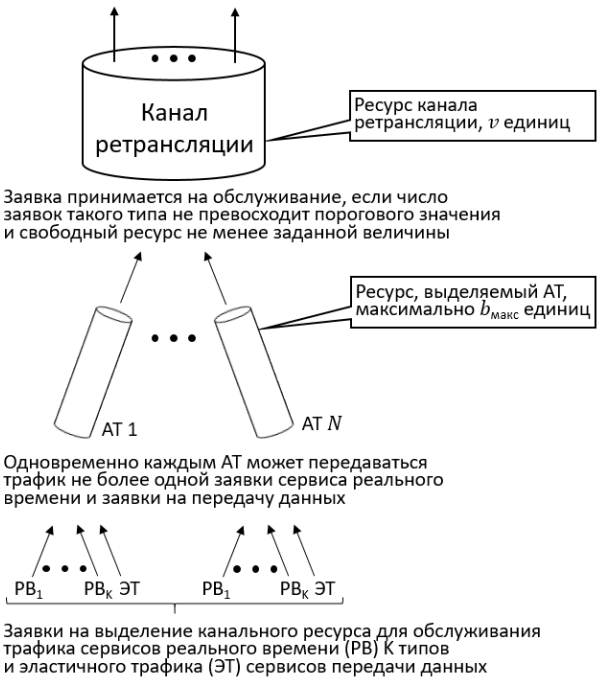

Рис. 2. Особенности совместного обслуживания трафика реального времени и эластичного трафика в отдельной ЛЗ

передачу трафика сервисов реального времени и b_e минимальный ресурс для обслуживания заявок на передачу данных. При этом максимальный ресурс для заявки на передачу данных составляет b_{max} и определяется максимальной скоростью передачи АТ. Введем вектор $\mathbf{b} = (0,b_1,\ldots,b_K,b_e,(b_1+b_e),\ldots,(b_K+b_e))$, состоящий из $2(K+1)$ компонент, первые $(K+1)$ компонент которого соответствуют состояниям АТ от (0,0) до $(K,0)$, а вторые $(K+1)$ компонент состояниям АТ от (0,1) до $(K,1)$. Обозначим также через $n_{(k,q)}$ количество АТ, находящихся в текущий момент времени в состоянии (k,q), и через $\mathbf{n} = (n_{(0,0)},\ldots,n_{(K,0)},n_{(0,1)},\ldots,n_{(K,1)})$ вектор количеств АТ в рассматриваемой ЛЗ, находящихся в соответствующих состояниях. Сумма компонент $\mathbf{n}$ равна количеству АТ в ЛЗ N. Очевидно, минимальный ресурс, необходимый для обслуживания всех АТ зоны в текущем состоянии, равен $l = \mathbf{nb}^T$, т. е., если суммарный ресурс ЛЗ составляет v, то поступающим заявкам может выделяться ресурс из оставшейся части $v - l$. При этом ресурс для передачи трафика данных отдельным АТ может быть увеличен до b_{max}, если этот АТ находится в состоянии (0,1), и до $(b_{max} - b_k)$, если АТ находится в состоянии $(k,1), k = 1,2,\ldots,K$. Отметим также, что количество обслуживаемых в системе заявок на передачу трафика реального времени k-го типа равно $(n_{(k,0)} + n_{(k,1)}), k = 1,2,\ldots,K$, а количество заявок на передачу данных $\sum_{k=0}^{K} n_{(k,1)}$. Как и в [6], полагаем, что потоки заявок на передачу трафика реального времени и данных независимы, интервалы между моментами поступления заявок подчиняются экспоненциальному распределению. Если АТ не занят в текущий момент передачей трафика реального времени, то новые заявки поступают с интенсивностями $\beta_k, k = 1,2,\ldots,K$. По сути, это означает, что интервал до момента поступления новой заявки на передачу трафика реального времени имеет экспоненциальное распределение с параметром $\beta = \sum_{k=1}^{K} \beta_k$, и вероятность, что это будет заявка k-го типа, равна $\frac{\beta_k}{\beta}$. Также, если АТ не занят в текущий момент передачей трафика данных, то новые заявки возникают с интенсивностью β_e. Время обслуживания заявок тоже подчиняется экспоненциальному распределению с параметрами $\mu_k, k = 1,2,\ldots,K$, и μ_e (μ_e соответствует передаче трафика данных на минимальной скорости, т. е. при выделении ресурса b_e).

Для учета правил резервирования обозначим через G множество типов заявок, за счет которых оно осуществляется. Введем ограничения на:

- предельные количества одновременно обслуживаемых заявок каждого типа $c_k,k \in G$ и c_e (если заявки на передачу эластичного трафика данных входят в G), т. е. $(n_{(k,0)} + n_{(k,1)}) \leqslant c_k,k \in G$ и $\sum_{k=0}^{K} n_{(k,1)} \leqslant c_e$;
- предельные объемы занятого ресурса $\Theta_k, k \in G$ и Θ_e, при превышении которых поступающие заявки каждого типа получают отказ.

В частности, если допустить, что после приема на обслуживание заявок неприоритетных типов оставшийся доступным для распределения ресурс составляет не менее r единиц, то $\Theta_k = v - r - b_k$ и $\Theta_e = v - r - b_e$, т. е. $\mathbf{nb}^T + b_k \leqslant v - r,k \in G$ и $\mathbf{nb}^T + b_e \leqslant v - r$.

С учетом перечисленных ограничений множество состояний сети имеет вид:

$$S = \Bigg\{ \mathbf{n} : \mathbf{nb}^T \leqslant v; n_{(k,q)} \geqslant 0,k = 0,1,\ldots,K,q = 0,1; \tag{1}$$

$$(n_{(k,0)} + n_{(k,1)}) \leqslant \min\left(c_k, \left\lfloor \frac{(v-r)}{b_k} \right\rfloor\right),k \in G, \sum_{k=0}^{K} n_{(k,1)} \leqslant \min\left(c_e, \left\lfloor \frac{(v-\delta r)}{b_e} \right\rfloor\right),$$

$$\sum_{k=0}^{K} (n_{(k,0)} + n_{(k,1)}) = N \Bigg\}.$$

Здесь $\delta = 1$, если заявки на передачу эластичного трафика данных входят в G, и $\delta = 0$ в противном случае.

Далее определим множества состояний, являющихся подмножествами S, при нахождении в которых имеют место потери поступающих заявок.

Для заявок k-го типа при $k \notin G$, множества состояний сети, для которых поступающие заявки k-го типа получают отказ, записываются в виде:

$$U_k = \left\{ \mathbf{n} : \mathbf{n} \in S,\mathbf{nb}^T > v - b_k, (n_{(0,0)} + n_{(0,1)}) > 0 \right\},k \notin G. \tag{2}$$

В (2) условие $(n_{(0,0)} + n_{(0,1)}) > 0$ означает, что в сети есть АТ, способные порождать заявки типа k. Если заявки на передачу эластичного трафика данных не входят в G, соответствующее множество состояний сети имеет вид:

$$U_e = \left\{ \mathbf{n} : \mathbf{n} \in S,\mathbf{nb}^T > v - b_e, \sum_{k=0}^{K} n_{(k,0)} > 0 \right\}. \tag{3}$$

Аналогично, в (3) условие $\sum_{k=0}^{K} n_{(k,0)} > 0$ означает, что в сети есть АТ, способные порождать заявки на передачу трафика данных.

Для типов трафика из G множества состояний системы, для которых поступающие заявки получают отказ, записываются в виде:

$$U_k^G = \Bigg\{ \mathbf{n} : \mathbf{n} \in S,\mathbf{nb}^T > (v - b_k - r) \cup (n_{(k,0)} + n_{(k,1)}) = \tag{4}$$

$$= \min\left(c_k, \left\lfloor \frac{(v-r)}{b_k} \right\rfloor\right), (n_{(0,0)} + n_{(0,1)}) > 0 \Bigg\}, k \in G,$$

и

$$U_e^G = \Big\{ \mathbf{n} : \mathbf{n} \in S, \mathbf{n}\mathbf{b}^T > (v - b_e - r) \cup \sum_{k=0}^{K} n_{(k,1)} = \tag{5}$$

$$= \min \left(c_e, \left\lfloor \frac{(v-r)}{b_e} \right\rfloor \right), \sum_{k=0}^{K} n_{(k,0)} > 0 \Big\}.$$

Для определения возможных переходов между состояниями сети целесообразно также ввести множество U_{in}^G состояний, в которые невозможно попасть при поступлении заявок из G:

$$U_{in}^G = \left\{ \mathbf{n} : \mathbf{n} \in S, \mathbf{n}\mathbf{b}^T > (v - r) \right\}. \tag{6}$$

Здесь надо иметь в виду, что сеть может перейти в состояния $\mathbf{n} \in U_{in}^G$ за счет поступления заявок важных типов, и поэтому в указанных состояниях возможны события завершения обслуживания заявок типов $k \in G$.

Обозначим через $p(\mathbf{n}), \mathbf{n} \in S$, вероятности нахождения сети в состоянии $\mathbf{n}$ в установившемся режиме. Указанные вероятности удовлетворяют СУР:

$$p(\mathbf{n}) \sum_{\mathbf{n}_{out} \in S_{out}(\mathbf{n})} a(\mathbf{n}, \mathbf{n}_{out}) = \sum_{\mathbf{n}_{in} \in S_{in}(\mathbf{n})} p(\mathbf{n}_{in}) a(\mathbf{n}_{in}, \mathbf{n}), \mathbf{n} \in S \tag{7}$$

Здесь $S_{in}(\mathbf{n})$ – множество состояний $\mathbf{n}_{in}$, из которых можно попасть в $\mathbf{n}$ в результате наступления событий поступления или завершения обслуживания заявок с интенсивностями переходов $a(\mathbf{n}_{in}, \mathbf{n})$, а $S_{out}(\mathbf{n})$ – множество состояний $\mathbf{n}_{out}$, в которые можно попасть из $\mathbf{n}$ с интенсивностями переходов $a(\mathbf{n}, \mathbf{n}_{out})$. Для решения СУР, как видно из (7), необходимо определить множества $S_{in}(\mathbf{n})$, $S_{out}(\mathbf{n})$, а также соответствующие $a(\mathbf{n}_{in}, \mathbf{n})$, и $a(\mathbf{n}, \mathbf{n}_{out})$. Как и в [6], обозначим через $\mathbf{e}_{(k,q)}$ единичный вектор размерности вектора $\mathbf{n}$, содержащий 1 в позиции, соответствующей компоненте $n_{(k,q)}$ в векторе $\mathbf{n}$, и нули в остальных позициях. Всем событиям поступления и завершения обслуживания заявок в сети соответствует разность двух единичных векторов. Например, поступлению заявки k-го типа соответствует переход сети из состояния $\mathbf{n}$ в состояние $\mathbf{n} - \mathbf{e}_{(0,q)} + \mathbf{e}_{(k,q)}$, т. е. группа АТ, в которой не обслуживаются заявки реального времени, уменьшается на один АТ, а к группе АТ, в которой обслуживаются заявки k-го типа, добавляется один АТ (в обеих группах АТ обслуживают или не обслуживают заявки на передачу трафика данных в соответствии со значением q).

Определяя множество $S_{in}(\mathbf{n})$ и связанные с ним интенсивности переходов, отметим, что попадание в состояние $\mathbf{n}$ при поступлении заявок должно рассматриваться, как для случая заявок на передачу трафика реального времени, так и для случая заявок на передачу данных. Если поступила заявка на передачу трафика реального времени k-го типа, т.е. переход осуществлен из состояния $\mathbf{n} + \mathbf{e}_{(0,q)} - \mathbf{e}_{(k,q)}$ и $n_{(k,q)} > 0$, то интенсивность такого перехода равна $(n_{(0,q)} + 1)\beta_k$. Если при этом $k \in G$, то дополнительно должно выполняться условие $\mathbf{n} \in S \setminus U_{in}^G$. Аналогично, если поступила заявка на передачу данных, т.е. переход осуществлен из состояния $\mathbf{n} + \mathbf{e}_{(k,0)} - \mathbf{e}_{(k,1)}$ и $n_{(k,1)} > 0$, то интенсивность перехода равна $(n_{(k,0)} + 1)\beta_e$. Если при этом $\delta = 1$, то дополнительно должно выполняться условие $\mathbf{n} \in S \setminus U_{in}^G$. Далее рассмотрим случаи завершения обслуживания заявок. Если обслужена заявка на передачу трафика реального времени k-го типа, т.е. переход осуществлен из состояния $\mathbf{n} - \mathbf{e}_{(0,q)} + \mathbf{e}_{(k,q)}$, $\mathbf{n} \in S \setminus U_k$ и $n_{(0,q)} > 0$, то интенсивность такого перехода равна $(n_{(k,q)} + 1)\mu_k$. Если же обслужена заявка на передачу данных, т.е. переход осуществлен из состояния $\mathbf{n} - \mathbf{e}_{(k,0)} + \mathbf{e}_{(k,1)}$, $\mathbf{n} \in S \setminus U_e$ и $n_{(k,0)} > 0$, то интенсивность такого перехода равна $(n_{(k,1)} + 1)\mu_e^k(\mathbf{n} - \mathbf{e}_{(k,0)} + \mathbf{e}_{(k,1)})$. Здесь функция $\mu_e^k(\mathbf{n})$ имеет смысл интенсивности обслуживания заявок на передачу данных для групп АТ, которые

либо обслуживали только такие заявки (случай $k = 0$), либо одновременно с заявками на передачу данных обслуживали заявки на передачу трафика реального времени типа k. Процедура расчета $\mu_e^k(\mathbf{n})$ описана в [6].

Количество элементов множества $S_{in}(\mathbf{n})$ определяется числом возможных переходов в состояние $\mathbf{n}$, которое может включать случаи:

- поступления заявок на передачу трафика реального времени с переходом АТ из состояния (0,0) в состояние $(k,0), k = 1,2,\ldots,K$, и с переходом АТ из состояния (0,1) в состояние $(k,1), k = 1,2,\ldots,K$, т.е. $2K$ случаев;
- поступления заявок на передачу данных с переходом АТ из состояния $(k,0)$ в состояние $(k,1), k = 0,1,\ldots,K$, т.е. $K+1$ случаев;
- завершения обслуживания заявок на передачу трафика реального времени с переходом АТ из состояния $(k,0)$ в состояние $(0,0), k = 1,2,\ldots,K$, и с переходом АТ из состояния $(k,1)$ в состояние $(0,1), k = 1,2,\ldots,K$, т.е. $2K$ случаев;
- завершения обслуживания заявок на передачу данных с переходом АТ из состояния $(k,1)$ в состояние $(k,0), k = 0,1,\ldots,K$, т.е. $K+1$ случаев.

Всего максимально $6K+2$ случаев, т.к. не все случаи могут удовлетворять перечисленным выше условиям.

Подобным образом определяются множество S_{out} и связанные с ним интенсивности переходов. При поступлении заявки на передачу трафика реального времени k-го типа она принимается на обслуживание, если $\mathbf{n} \in S \setminus U_k^G$ при $k \in G$ или $\mathbf{n} \in S \setminus U_k$ в противном случае. Переход осуществляется в состояние $\mathbf{n} - \mathbf{e}_{(0,q)} + \mathbf{e}_{(k,q)}$ при условии $n_{(0,q)} > 0$ с интенсивностью $n_{(0,q)}\beta_k$. Аналогично при поступлении заявки на передачу данных она принимается на обслуживание, если $\mathbf{n} \in S \setminus U_e^G$ при $\delta = 1$ или $\mathbf{n} \in S \setminus U_e$ в противном случае. Переход осуществляется в состояние $\mathbf{n} - \mathbf{e}_{(k,0)} + \mathbf{e}_{(k,1)}$ при условии $n_{(k,0)} > 0$ с интенсивностью $n_{(k,0)}\beta_e$. При завершении обслуживания заявок на передачу трафика реального времени k-го типа происходит переход в состояние $\mathbf{n} + \mathbf{e}_{(0,q)} - \mathbf{e}_{(k,q)}$. Такой переход возможен, если $n_{(k,q)} > 0$. Интенсивность перехода $n_{(k,q)}\mu_k$. А при завершении обслуживания заявок на передачу данных происходит переход в состояние $\mathbf{n} + \mathbf{e}_{(k,0)} - \mathbf{e}_{(k,1)}$. Такой переход возможен, если $n_{(k,1)} > 0$. Интенсивность перехода $n_{(k,1)}\mu_e^k(\mathbf{n})$.

Для решения СУР, следуя подходам, разработанным в [5], используется метод Гаусса-Зейделя, процедура применения которого описана в [6].

Далее перейдем к определению показателей качества предоставления сервисов с учетом найденных $p(\mathbf{n}), \mathbf{n} \in S$. Значения показателей для отдельной ЛЗ зависят от принятых для этой ЛЗ параметров резервирования $c_k, k \in G$, c_e при $\delta = 1$ и r. Обозначим $\pi_{k,cr}$ и $\pi_{e,cr}$ – доли потерянных заявок на передачу трафика k-го типа и эластичного трафика для заданных параметров резервирования, а π_k и π_e – доли потерянных заявок на передачу трафика k-го типа и эластичного трафика при использовании неких предпочтительных параметров резервирования, которые будут выбираться ниже в зависимости от критериев решения задач. Также обозначим W_{cr} и W – среднее время обслуживания заявки на передачу данных для заданных параметров резервирования и при использовании предпочтительных параметров резервирования, соответственно. Ограничимся оценкой $\pi_{k,cr}$, $\pi_{e,cr}$ и W_{cr} с позиций пользователя [6]. Для заявок на обслуживание трафика реального времени k-го типа получаем

$$\pi_{k,cr} = \frac{\sum_{\mathbf{n} \in U_k^{den}} \big(p(\mathbf{n})(n_{(0,0)} + n_{(0,1)})\big)}{\sum_{\mathbf{n} \in S} \big(p(\mathbf{n})(n_{(0,0)} + n_{(0,1)})\big)} \tag{8}$$

и для заявок на передачу данных

$$\pi_{e,cr} = \frac{\sum_{\mathbf{n}\in U_e^{den}} \left(p(\mathbf{n}) \sum_{k=0}^{K} n_{(k,0)} \right)}{\sum_{\mathbf{n}\in S} \left(p(\mathbf{n}) \sum_{k=0}^{K} n_{(k,0)} \right)}. \tag{9}$$

Здесь U_k^{den} это U_k при $k \notin G$, и U_k^G в противном случае. Аналогично, U_e^{den} это U_e, если $\delta = 0$, и U_e^G в противном случае. Среднее время обслуживания заявки на передачу данных вычисляется с использованием формулы Литтла, как отношение среднего количества обслуживаемых заявок данных y_e к интенсивности $\lambda_e(1 - \pi_e)$ потока заявок на передачу данных, принятых на обслуживание, и равно:

$$W_{cr} = \frac{y_e}{\lambda_e(1 - \pi_e)}, \tag{10}$$

$$y_e = \sum_{\mathbf{n}\in S} \left(p(\mathbf{n}) \sum_{k=0}^{K} n_{(k,1)} \right),$$

$$\lambda_e = \beta_e \sum_{\mathbf{n}\in S} \left(p(\mathbf{n}) \sum_{k=0}^{K} n_{(k,0)} \right).$$

Показатели для сети, т.е. на множестве ЛЗ, где расположены АТ, определяются в зависимости от решаемой задачи. Можно, в частности, стремиться к достижению наилучшего значения того или иного показателя для сети в целом, допуская некоторое снижение качества в отдельных ЛЗ, а можно к обеспечению равных условий по качеству обслуживания для всех ЛЗ. Как и в [6], рассмотрим эти два варианта на примере среднего времени обслуживания заявки на передачу данных, W_{c1} и W_{c2}, соответственно. Показатель W_{c1} вычисляется, как осредненный на множестве ЛЗ, в виде:

$$W_{c1} = \frac{1}{\lambda_{sum}} \sum_{m=1}^{M} \lambda_e^m (1 - \pi_e^m) W^m = \frac{\sum_{m=1}^{M} y_e^m}{\lambda_{sum}}, \tag{11}$$

$$\lambda_{sum} = \sum_{m=1}^{M} \lambda_e^m (1 - \pi_e^m).$$

Здесь принадлежность показателей и параметров к той или иной ЛЗ обозначается верхним индексом $m, m = 1,2,\ldots,M$, M – общее число ЛЗ.

Для второго варианта среднее время обслуживания заявки на передачу данных независимо от ЛЗ дислокации АТ не хуже, чем

$$W_{c2} = \max_{m\in(1,M)} W^m. \tag{12}$$

3. ХАРАКТЕРИСТИКА ЗАДАЧ, КОТОРЫЕ МОГУТ РЕШАТЬСЯ С ИСПОЛЬЗОВАНИЕМ РАЗРАБОТАННЫХ МОДЕЛЕЙ

Разработанные модели могут использоваться для определения минимально необходимого ресурса сети и поиска предпочтительных решений по его распределению между ЛЗ. Исходными данными для решения задач являются:

- количество АТ в каждой ЛЗ $N^m, m = 1,2,\ldots,M$;
- максимальная скорость передачи АТ в единицах ресурса b_{max};

- перечень сервисов реального времени, их характеристики $(b_k, \mu_k, k = 1,2,\ldots,K)$;
- характеристики эластичного трафика данных (b_e, μ_e);
- требования по качеству предоставления сервисов $(\pi_k^m \leqslant \pi_k^*, k = 1,2,\ldots,K, \pi_e^m \leqslant \pi_e^*$ и $W^m \leqslant W^*, m = 1,2,\ldots,M)$;
- интенсивности возникновения на АТ заявок каждого типа $\beta_k, k = 1,2,\ldots,K$ и β_e;
- данные о приоритетности типов трафика.

Отметим, что при заданных характеристиках АТ показатели качества для отдельных ЛЗ являются функциями количества АТ N^m и выделенного ЛЗ ресурса v^m, $m = 1,2,\ldots,M$. Это позволяет использовать для решения задач двухэтапные процедуры. На первом этапе осуществляется получение зависимостей показателей качества для отдельных ЛЗ от возможных значений ресурса v^m, включая шаги:

- отыскание для каждого v^m предпочтительных параметров резервирования $c_k, k \in G, c_e$ и r с учетом критериев решения задачи (на множестве $\pi_{k,cr}^m$ и $\pi_{e,cr}^m$ определяются π_k^m и π_e^m, а на множестве W_{cr}^m находятся W^m);
- собственно построение зависимостей на основе результатов первого шага.

На втором этапе с учетом этих зависимостей отыскиваются предпочтительные решения для сети в целом.

Задача минимизации требуемого ресурса записывается в виде:

$$v = \sum_{m=1}^{M} v^m \to \min \tag{13}$$

$$\pi_k^m(v^m,N^m) \leqslant \pi_k^*, m = 1,2,\ldots,M, k = 1,2,\ldots,K,$$

$$\pi_e^m(v^m,N^m) \leqslant \pi_e^*, W^m(v^m,N^m) \leqslant W^*, m = 1,2,\ldots,M.$$

Очевидно, суммарный ресурс будет минимальным, если минимальны ресурсы (v^m), предоставляемые каждой ЛЗ. Отыскание v^m может осуществляться путем пошагового изменения возможных значений объема ресурса с проверкой перебором существования параметров резервирования $c_k, k \in G, c_e$ и r, для которых выполняются ограничения задачи (13).

Задача поиска предпочтительных решений по распределению ресурса между ЛЗ в постановке минимизации среднего времени обслуживания заявки на передачу данных в сети W_{c1} с учетом (11) имеет вид:

$$W_{c1} \to \min \tag{14}$$

$$\sum_{m=1}^{M} v^m = v,$$

$$\pi_k^m(v^m,N^m) \leqslant \pi_k^*, m = 1,2,\ldots,M, k = 1,2,\ldots,K,$$

$$\pi_e^m(v^m,N^m) \leqslant \pi_e^*, W^m(v^m,N^m) \leqslant W^*, m = 1,2,\ldots,M.$$

Задача поиска предпочтительных решений по распределению ресурса между ЛЗ в постановке обеспечения равных условий по качеству обслуживания для АТ всех ЛЗ записывается в виде:

$$W_{c2} \to \min \tag{15}$$

$$\sum_{m=1}^{M} v^m = v,$$

$$\pi_k^m(v^m,N^m) \leqslant \pi_k^*, m = 1,2,\ldots,M, k = 1,2,\ldots,K,$$

$$\pi_e^m(v^m,N^m) \leqslant \pi_e^*, W^m(v^m,N^m) \leqslant W^*, m = 1,2,\ldots,M.$$

Очевидно, при решении задач (14) и (15) сначала следует отыскать минимальные ресурсы ЛЗ $v_{min}^m, m = 1,2,\ldots,M$, обеспечивающие выполнение ограничений по предельным значениям показателей качества, т. е. необходимо решить задачу в постановке (13). При этом для каждого v_{min}^m в качестве параметров резервирования $c_k, k \in G, c_e$ при $\delta = 1$ и r принимаются те, которые минимизируют W^m. Далее распределению между ЛЗ подлежит оставшийся ресурс $v_{left} = v - \sum_{m=1}^{M} v_{min}^m$. С этой целью строятся зависимости $W^m(v^m,N^m), m = 1,2,\ldots,M$, для значений ресурса от v_{min}^m до$v_{min}^m + v_{left}$. Количество вариантов прямого перебора равно числу сочетаний из $(v_{left} + M - 1)$ по $(M - 1)$, т. е. $C_{v_{left}+M-1}^{M-1}$. Перебор может использоваться, когда v_{left} и M невелики. Например, если АТ расположены в семи ЛЗ, а объем нераспределенного ресурса составляет 30-40 единиц, то необходимо перебрать порядка 2-10 миллионов вариантов [6]. В случае, если интенсивности потоков заявок каждого типа можно приближенно считать постоянными (как в модели Эрланга), а потери заявок на передачу данных пренебрежимо малыми, для решения задачи (14) можно использовать метод динамического программирования по аналогии с тем, как это сделано в [22]. Относительно решения задачи (15) необходимо отметить, что зависимости среднего времени обслуживания заявки на передачу данных от ресурса для каждой ЛЗ являются невозрастающими. При этом ресурс следует распределять пошагово, направляя очередную единицу в ЛЗ, где текущее достигнутое среднее время обслуживания максимально.

4. ЧИСЛЕННЫЕ ПРИМЕРЫ ИСПОЛЬЗОВАНИЯ МОДЕЛЕЙ

В качестве примера рассмотрим случай размещения АТ в трех ЛЗ, причем $N^1 = 100, N^2 = 50$ и $N^3 = 30$. Максимальная скорость передачи АТ в единицах ресурса $b_{max} = 8$. Водителям доступны два сервиса реального времени с параметрами: $b_1 = 1, b_2 = 4, \mu_1 = 0{,}3$ мин$^{-1}, \mu_2 = 0{,}15$ мин$^{-1}$, а также сервисы передачи эластичного трафика с параметрами $b_e = 2, \mu_e = 0{,}3$ мин$^{-1}$. Интенсивности возникновения на каждом АТ заявок на передачу трафика реального времени $\beta_1 = 0{,}03$ мин$^{-1}, \beta_2 = 0{,}002$ мин$^{-1}$, а для эластичного трафика передачи данных $\beta_e = 0{,}01$ мин$^{-1}$. Трафик реального времени первого типа является приоритетным (допустим, водитель использует этот сервис для докладов в центр управления перевозками, сообщает о возникающих нештатных ситуациях и т. п.). Соответственно, требования по качеству обслуживания заданы в виде $\pi_1^m \leqslant 0{,}001, \pi_2^m \leqslant 0{,}03, \pi_e^m \leqslant 0{,}03, m = 1,2,3$. Суммарный ресурс сети $v = 86$ единиц. Пусть должна решаться задача в постановке (15), т. е. обеспечения равных условий обслуживания во всех ЛЗ.

На первом шаге отыскиваются минимальные значения ресурса для каждой ЛЗ (таблица 1).

Видно, что суммарный минимальный ресурс равен 71. При этом в каждой ЛЗ параметры резервирования различны. Оставшийся ресурс v_{left} составляет 15 единиц. Далее получаем зависимости среднего времени обслуживания заявок на передачу эластичного трафика данных от выделенного ресурса для каждой ЛЗ и осуществляем пошаговое распределение указанных 15 единиц (см. таблицу 2). Получаем, что первой ЛЗ следует добавить 7 единиц ресурса, второй и третьей – по 4 единицы.

Итоговые результаты решения задачи представлены в таблице 3.

Наиболее трудоемким элементом процедуры решения задачи является получение π_k и π_e по $\pi_{k,cr}$ и $\pi_{e,cr}$, а также W по W_{cr} перебором по возможным вариантам параметров резервирования для каждой ЛЗ при каждом значении ресурса. Однако, количество вариантов перебора

Таблица 1. Результаты расчета минимальных значений ресурса и предпочтительных параметров c_2, c_e и r для отдельных ЛЗ

№ЛЗ, m	1	2	3
v^m_{min}	33	22	16
c_2	5	3	3
c_e	5	3	3
r	2	2	1
π^m_1	0,00078	0,00029	0,00092
π^m_2	0,02776	0,02972	0,02652
π^m_e	0,01880	0,02151	0,01379
W^m	0,99040	0,93641	0,97226

Таблица 2. Иллюстрация последовательных шагов распределения ресурса между ЛЗ

АТ в ЛЗ	Доп. ресурс для обслуживания АТ в ЛЗ и W^m, m = 1,2,3							
	0	1	2	3	4	5	6	7
100	**0,99040**	**0,94340**	**0,91807**	**0,90304**	**0,89133**	**0,88209**	**0,87495**	**0,86941**
	шаг 1	шаг 3	шаг 6	шаг 8	шаг 11	шаг 13	шаг 15	
50	**0,93640**	**0,91304**	**0,89609**	**0,88377**	**0,87469**	0,86812	0,86322	0,85861
	шаг 5	шаг 7	шаг 10	шаг 12				
30	**0,97226**	**0,93539**	**0,89804**	**0,87874**	**0,86883**	0,86110	0,85724	0,85554
	шаг 2	шаг 4	шаг 9	шаг 14				

Таблица 3. Результаты расчета предпочтительных параметров c_2, c_e и r для отдельных ЛЗ

№ЛЗ, m	1	2	3
v^m_{min}	40	26	20
c_2	7	3	4
c_e	4	3	2
r	10	6	4
π^m_1	1,13E-07	6,02E-07	3,68E-06
π^m_2	0,02583	0,02731	0,01335
π^m_e	0,02079	0,01569	0,02841
W^m	0,86941	0,87469	0,86883

может быть существенно сокращено, если отбросить зоны значений параметров резервирования, где их изменение практически не влияет на показатели качества. Например, из рисунков 3 и 4 видно, что интерес представляют значения $c_2 < 7$ и $c_e < 8$.

5. ЗАКЛЮЧЕНИЕ

Для сетей маломощных абонентских терминалов на базе космических аппаратов с высокой пропускной способностью построена модель резервирования канального ресурса, осуществляемого с целью либо выравнивания потерь заявок разных типов, либо создания преимуществ для передачи более важного трафика. Модель позволяет комплексно учесть важнейшие факторы функционирования сетей, включающие совместное обслуживание нескольких типов трафика реального времени и эластичного трафика данных, неравномерность размещения терминалов по локальным зонам, ограниченное их количество, возможность одновременного обслуживания каждым терминалом заявок на передачу трафика двух типов, один из которых является эластичным трафиком данных, другой трафиком реального времени, ограничение на скорость передачи информации терминалом. При построении модели использован аппарат многомер-

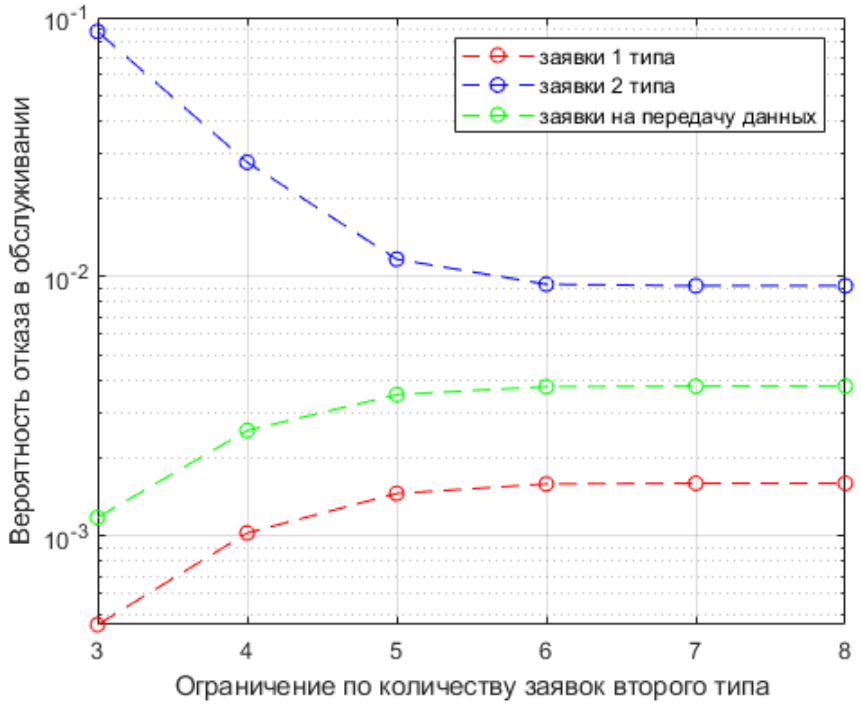

Рис. 3. Зависимости вероятности отказа в обслуживании заявок разных типов π_1^1,π_2^1,π_e^1 от ограничения по количеству обслуживаемых заявок второго типа c_2 при $\nu^1 = 35$ и $c_e = 17$

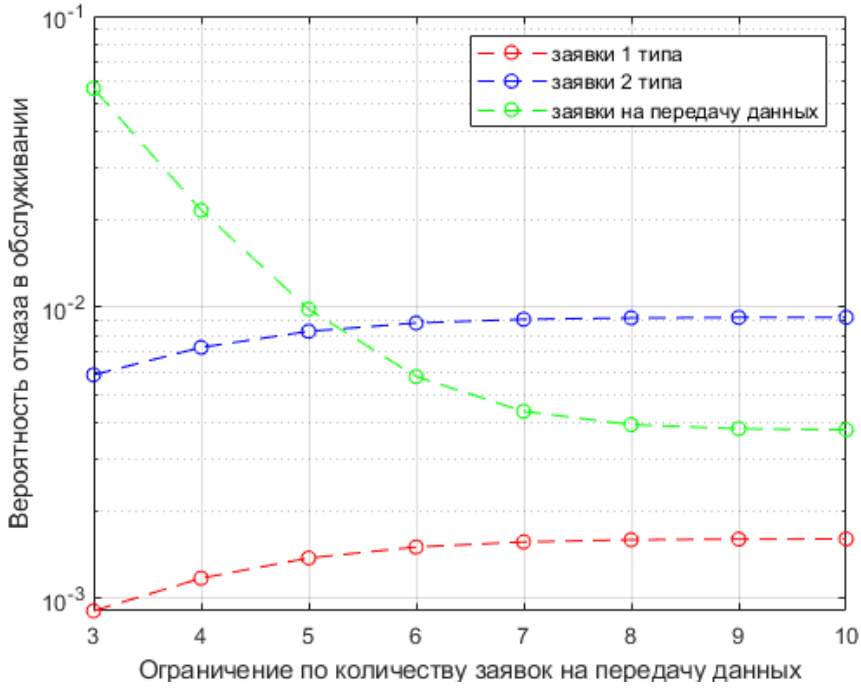

Рис. 4. Зависимости вероятности отказа в обслуживании заявок разных типов π_1^1,π_2^1,π_e^1 от ограничения по количеству обслуживаемых заявок на передачу данных c_e при $\nu^1 = 35$ и $c_e = 8$

ных ступенчатых марковских процессов и метод Гаусса-Зейделя для нахождения вероятностей состояний сети в установившемся режиме. Перечень задач, которые могут решаться с использованием модели, включает как задачи определения минимально необходимого ресурса сети

для выполнения заданных требований по качеству, так и задачи отыскания предпочтительных распределений ресурса между локальными зонами. Сформированы процедуры нахождения предпочтительных параметров резервирования и приведены численные примеры их применения. Полученные результаты позволяют использовать модели при планировании и управлении работой сетей на базе космических аппаратов с высокой пропускной способностью.

СПИСОК ЛИТЕРАТУРЫ

1. Урличич Ю.М. Перспективные технологии и сервисы спутниковой связи. Доклад на XXVII Международной конференции операторов и пользователей сети спутниковой связи и вещания в Российской Федерации, 2021.

2. Алешин В.С., Догаев С.Г. Оценка реальных потребностей России в услугах подвижной спутниковой связи. *Электросвязь*, 2021. № 2. С. 48-51.

3. Vladimir D. Burtsev, Tatyana S. Vosheva, Dmitry S. Filonov, Petr V. Semkin, Stanislav A. Kuznetsov. 3D printed optimized phased array with volumetric emitters. *2023 DAYS on DIFFRACTION (DD)*, 2023, pp. 25-30.

4. Anton V. Nikulin, Aleksey N. Kosmynin, Vladimir D. Burtsev, Sergey Y. Prokhorov, Tatyana S. Vosheva, Dmitry S. Filonov. Towards Phased Array Antenna Operating in Ku-band for Satellite Communications. *Photonics and Nanostructures - Fundamentals and Applications*, 2023, pp. 25-30.

5. Степанов С.Н. Теория телетрафика: концепции, модели, приложения. *М.:Горячая линия – Телеком*, 2015. С. 868.

6. Maslov A.A., Sebekin G.V., Stepanov S.N., Shchurkov A.O., Vasilyev A. Model of processes for joint maintenance of real-time multiservice traffic and elastic data traffic in a network of low-power mobile subscriber terminals based on high-throughput satellites. *T-Comm*, 2024 (в печати).

7. Stepanov M.S., Kanishcheva M.G., Malikova E.E., Ndayikunda J., Kaigorodov N.G. The Development and Analysis of a Service Model for the Traffic of a Surveillance System Operator by a Dedicated Resource of an LTE Cell. *2022 Systems of Signals Generating and Processing in the Field of on Board Communications*, 2022, pp. 1-6.

8. Stepanov M.S., Stepanov S.N., Kroshin F.S. Effective Algorithm of Estima-tion the Performance Measures of Group of Servers with Dependence of Call Repetition on the Type of Call Blocking. *DCCN 2022. Lecture Notes Computer Science, Springer, Cham.*, 2022.

9. Gibadullina Elena E., Viskova Elena V., Stepanov Sergey N. Automated Service Configuration Management in IP/MPLS Networks. *4th International Science and Technology Conference "Modern Network Technologies - 2022", MoNeTec*, 2022.

10. Andrabi U.M., Kanishcheva M., Stepanov S.N. Observation system resource planning in presence of access control based on volume of resource occupied by traffic flows. *T-Comm*, 2022, vol. 16, no.8, pp.54-62.

11. Dawood T., Stepanov M. S., Naoussi C., Joao B., Yahia K. The Mathematical Model of the Internet of Things Traffic Servicing in Case of its Impulse Nature. *2023 Systems of Signals Generating and Processing in the Field of on Board Communications, Moscow, Russian Federation*, 2023, pp. 1-8.

12. Ndimumahoro F., Stepanov M.S., Muzata A.R., Tammam D., Popov V.G. Using the Principles of Mobile Systems Modeling for LoRaWAN Characteristics Estimation. *2022 Systems of Signals Generating and Processing in the Field of on Board Communications*, 2022, pp. 1-8.

13. Stepanov M.S., Stepanov S.N., Andrabi U., Petrov D., Ndayikunda J. The Increasing of Resource Sharing Efficiency in Network Slicing Implementation. *Vishnevskiy V., Samouylov K., Kozyrev D. (eds) Distributed Computing and Computer Networks. DCCN 2021. Lecture Notes Computer Science, Springer, Cham.*, 2022. — V.1552. — P. 18 – 35.

14. Volkov A.O., Korobkina A.V., Stepanov S.N. Development of a Model and Algorithms for Servicing Real-Time and Data Traffic in a Cloud Computing System. *2022 Systems of Signals Generating and Processing in the Field of on Board Communications*, 2022, pp. 1-6.

15. Stepanov M., Ndimumahoro F., Dawood T., Mutumba B., Alberto J., Naoussi C. The Construction and Analysis of the LTE Network Model with Joint Service of Traffic of LTE and NB-IoT Devices. *4th International Science and Technology Conference "Modern Network Technologies - 2022", MoNeTec*, 2022.

16. Umer M. A., Stepanov M.S. Collective servicing of heterogenous traffic streams over 3GPP LTE network and application of access control. *T-Comm*, 2022, vol. 16, no.3, pp. 43-49.

17. Stepanov M.S., Stepanov S.N., Kanischeva M.G., Kroshin F.S. Analysis of Procedures for Joint Servicing of Multiservice Traffic in Access Nodes. *DCCN 2023. Lecture Notes Computer Science, Springer, Cham.* ,2023. P.1-15.

18. Stepanov S.N., Korotkova V.I., Kanischeva M.G., Pshenichnikov A.P. Construction and Analysis of a Model of Group Receipt of IoT Elastic Data in a Wireless Access Node. *2023 Systems of Signals Generating and Processing in the Field of on Board Communications, SOSG, Conference Proceedings*, 2023 pp. 1-8.

19. U. M. Andrabi, and S. N. Stepanov. The Model of Conjoint Servicing of Real-Time Traffic of Surveillance Cameras and Elastic Traffic Devices with Access Control. *2nd International Informatics and Software Engineering Conference (IISEC). Ankara, Turkey*, 2023 pp. 1-6.

20. Stepanov S.N., Andrabi U.M., Stepanov M.S., Ndayikunda J. Reservation Based Joint Servicing of Real Time and Batched Traffic in Inter Satellite Link. *Proc. of 2020 Systems of Signals Generating and Processing in the Field of on Board Communications. Moscow, Russia*, 2020, pp. 1-5.

21. Stepanov S.N., Romanov A.I., Stepanov M.S., Vasilyev A.P. Analytical modeling of call admission control for real-time and elastic services in inter satellite link. *2018 Systems of Signals Generating and Processing in the Field of on Board Communications*, 2018, C. 8350637.

22. Маков С.В., Маслов А.А., Себекин Г.В. Оценка эффективности использования ресурса пропускной способности каналов ретрансляции космических аппаратов класса HTS на геостационарной и высокоэллиптической орбите для организации сети передачи трех типов трафика. *Наноиндустрия. Специальный выпуск 9s, ТОМ 16(119)*, 2023, С. 613-619.

Model for Reserving Channel Resource When Servicing Heterogeneous Traffic in a Network of Low-power Mobile Subscriber Terminals Based on High-Throughput Satellites

A.A. Maslov, G.V. Sebekin, M.S. Stepanov, S.N. Stepanov, A.O. Shchurkov

Reserving a channel resource in networks of low-power mobile subscriber terminals based on high-throughput satellites makes it possible to increase the efficiency of network planning and management functions by either equalizing losses of different types of applications or creating advantages for transmitting more important traffic. To assess the quality of service provision in networks with resource reservation, it is necessary to develop mathematical models that should take into account the main features of networks, including the joint maintenance of several types of real-time traffic and elastic data traffic, uneven placement of terminals in local zones, a limited number of them, the possibility of simultaneous servicing by each terminal of requests for two types of traffic, one of which is elastic data traffic, the other is real-time traffic, the limit on the speed of information transfer by the terminal. The purpose of the work is to solve the modeling problem taking into account the listed factors. The model is based on multidimensional stepwise Markov processes. A list of tasks that can be solved using them is defined, including both the tasks of determining the minimum required network resource to meet specified quality requirements, and the tasks of finding preferred resource distributions between local zones according to the criterion of minimizing the

average time for servicing data transfer requests. Numerical examples of task solving are presented. The developed models can be used to build information transmission network management systems based on high-throughput satellites.

KEYWORDS: high-throughput satellites, channel resource, multiservice traffic, real-time traffic, elastic traffic, mobile subscriber terminals, resource reservation.