МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО»

УТВЕРЖДАЮ
Зав. кафедрой системного анализа и
автоматического управления
к. ф.-м. н., доцент
_____ И. Е. Тананко

ОТЧЕТ О ПРАКТИКЕ

студента 1 курса 171 группы факультета КНиИТ
направления 09.04.01 — Информатика и вычислительная техника
Сербина Владислава Андреевича

вид практики: производственная
кафедра: системного анализа и автоматического управления
курс: 1
семестр: 2
продолжительность: рассредоточенная, с 06.02.23 по 02.04.23, с 02.05.23 по 16.06.23

Руководитель практики,
доцент, к. ф.-м. н. _____ Е. П. Станкевич

Саратов 2023

Тема практики:
«Научно-исследовательская работа»

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Традиционные архитектуры сетей сотовой связи сталкиваются с огромными проблемами из-за беспрецедентного увеличения трафика мобильной передачи данных, ограниченной доступности спектра частот и высокого энергопотребления. В свете этого, отрасли, а также исследовательские сообщества находятся в постоянном поиске фундаментальных достижений в разработке новых сетевых архитектур для поддержки растущего пользовательского спроса при одновременном снижении капитальных и эксплуатационных расходов для сетевых операторов. Архитектура облачной сети радиодоступа (C-RAN) – это такая концепция смены парадигмы для сотовых сетей, которая также активно рассматривается в качестве основного кандидата для будущих сотовых систем $5G$.

Целью данной работы является подбор литературы и изучение математических моделей сетей облочного радиодоступа ($C-RAN$)

Работа состоит из двух разделов.

В первом разделе описываются некоторые математические модели $C-RAN$.

Во втором разделе описывается математическая модель, используемая для моделеирования и анализа $C-RAN$.

1 Математические модели C-RAN

С точки зрения моделирования, каждая антенна (RRH) представляет собой источник требований в направлении восходящей линии связи, в то время как для направления нисходящей линии связи требования поступают из базовой сети, которая обеспечивает подключение к внешним сетям (например, интернету или другим сервисным платформам). Затем для каждого сектора сотовой связи создаются две очереди заданий, по одной в каждом направлении. Поскольку ограничение времени на обработку субкадров нисходящих каналов составляет половину ограничения для восходящих каналов, они могут выполняться отдельно на выделенных процессорных блоках. Однако выделение процессоров для каждой очереди не является эффективным способом использования ограниченных ресурсов.

В работах [1–4] авторы сосредоточены на оптимизации некоторых параметров $C - RAN$.

В работе [2] рассматривается пул модулей базовой полосы частот (BBU) в C-RAN как набор виртуальных машин (VMs). Каждое пользовательское оборудованию (UE) может связываться с несколькими виртуальными машинами в пуле BBU и каждая удаленная радиоголовка (RRH) может обслуживать только ограниченное число UE. В рамках этой модели предложен вариант оптимизации использования виртуальных машин в пуле BBU и формирования разреженного луча в скоординированном кластере RRH, который ограничен пропускной способностью $fronthaul$, чтобы минимизировать системные затраты на $C - RAN$. Задача опатимизации формулируется как задача нелинейного программирования.

Предполагается, что кластер $C-RAN$ состоит из N одноантенных UEs и L $RRHs$, каждый с K антеннами. Множество всех UEs и всех $RRHs$ обозначим как $N = \{1, .., N\}$ и $L = \{1, .., L\}$ соответсвенно. В пуле BBU размещено M одинаковых виртуальных машин. Каждая из них обладает вычислительной мощностью μ и требует затрат ресурсов виртуальной машины $\varphi > 0$, когда она активна. Количество активных виртуальных машин обозначается как $m \in N$, где $m \leq M$. Эта модель отражает популярные модели коммерческих облачных сервисов, например, Amazon Elastic Compute Cloud (EC2).

В нисходящей линии связи C-RAN (рисунок 1), весь входящий трафик UEs сначала обрабатывается диспетчером. Предполагается, что интенсивность

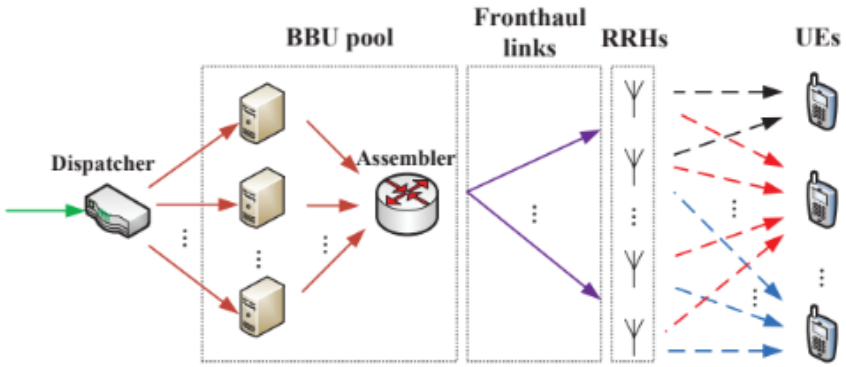

Рисунок 1 – Схема $C-RAN$

входного потока данных UE_i к диспетчеру равна $\lambda_i, \forall i \in N$ и пусть $\alpha = \sum_{i \in N} \lambda_i$. Затем каждый транспортный блок (или даже блок кода внутри каждого транспортного блока) в потоке данных от UE_i может быть направлен диспетчером на одну из m активных виртуальных машин для обработки (например, турбокодирования) с вероятностью $1/m$. Следовательно, интенсивность входящего трафика, направляемого на каждую активную виртуальную машину равна α/m.

В части беспроводной передачи авторы рассматривают совместную передачу как метод $CoMP$ в $C-RAN$, т.е. данные каждого UEs могут совместно использоваться всеми скоординированными $RRHs$, в то время как $RRHs$ имеют ограниченную пропускную способность линии $fronthaul$ (линия $fronthaul$ соединяет пул BBU и $RRHs$, может быть выполнена в виде линии оптоволоконной связи, медным кабелем или беспроводной передачей). После обработки виртуальными машинами данные каждого UEs пересылаются на UE не более чем через L $RRHs$ (поскольку данные распределяются между ограниченным количеством $RRHs$). Пусть достижимая скорость беспроводной передачи в UE_i равна c_i.

Каждая активная виртуальная машина в пуле BBU может быть смоделирована как система массового обслуживания. В частности, для каждой системы интенсивность входящего потока равна α/m, а интенсивность обслуживания равна μ. На протяжении всей статьи предполагается, что требования в каждой системе обслуживаются по принципу "первый пришел - первый ушел"($FIFO$), а длина очереди бесконечна.

Авторами рассматривается двухуровневая сеть массового обслуживания для представления поведения каждого UE при обработке и передаче данных по нисходящему каналу $C-RAN$, рисунок 2. В частности, в пуле BBU транс-

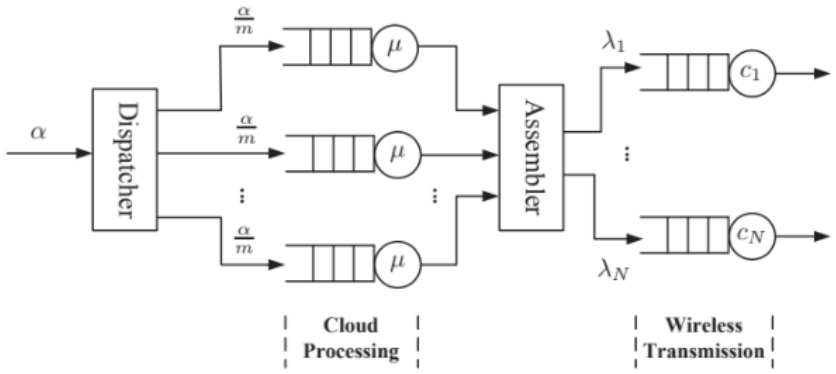

Рисунок 2 – Модель сети массового обслуживания, представляющая облачную обработку $C-RAN$ и беспроводную передачу данных

портные блоки для каждого UE обрабатываются (например, кодируются) m параллельными активными виртуальными машинами, каждая из которых абстрагируется как система массовго обслуживания со средней скоростью обслуживания μ. Затем обработанные данные передаются на UE_i через $RRHs$ по беспроводным каналам, которые моделируются системами массового обслуживания со средней скоростью обслуживания c_i.

Средняя задержка обработки данных для UE_i в пуле BBU обозначается как b_i. Пусть d_i - средняя задержка передачи данных на UE_i в системе беспроводной передачи (т.е. ожидаемая задержка, возникающая в системе до того, как данные будут полностью переданы). Предполагается, что процесс поступления пакета UE_i к диспетчеру является пуассоновским процессом с интенсивностью λ_i. Следовательно, процесс поступления требований в каждую виртуальную машину также формирует пуассоновский процесс со средней скоростью поступления α/m. Предполагается, что время обслуживания пакетов данных в каждой системе соответствует экспоненциальному распределению со средним значением $1/\mu$, для $\mu > \alpha/m$. Тогда интенсивность поступления для пакетов данных каждого UEs в системы беспроводной передачи такая же, как и для диспетчера. Предполагается, что время обслуживания каждого пакета данных в системах беспроводной передачи соответствует экспоненциальному распределению со средним значением $1/c_i$. Таким образом, обработку и передачу данных в данной модели $C-RAN$ можно рассматривать как две последовательные сети массового обслуживания, состоящие из систем $M/M/1$. Выражения для вычисления станционрых характеристик системы выглядят слудющим образом:

$$b_i = \frac{m}{m\mu - \alpha},$$

$$d_i = \frac{1}{c_i - \lambda_i}.$$

Пусть u_i об

альнейшая задача оптимизации решается как задача нелинейного программирования:

$$min - - > m\phi + \eta \sum_{i=1}^{N} \sum_{j=1}^{L} ||w_{ij}||_{2,1} + \eta \sum_{i=1}^{N} \sum_{j=1}^{L} P_j ||w_{ij}||_{2,0},$$

$$\frac{m}{m\mu - \alpha} + \frac{1}{c_i - \lambda_i} \leq \tau, \quad \forall i \in N$$

$$\alpha < m\mu, \quad \lambda_i < c_i, \forall i \in N,$$

$$0 < m \leq M, \quad m \in N,$$

$$c_i \leq B_i log(1 + SINR_i), \quad \forall i \in N$$

$$\sum_{i=1}^{N} ||w_{ij}||_{2,1} \leq E_j, \forall j \in L,$$

$$\sum_{i=1}^{N} ||w_{ij}||_{2,0} \leq S_j, \forall j \in L.$$

В работе [3] с целью определения оптимального размера $C - RAN$ модель представлена многоприборной системой массового обслуживания с неординарным входным поток $M^{[X]}/M/c$, с помощью модели оценивается необходимая вычислительная мощность в центре обработки данных при соблюдении требований к задержке ответа. Модель, в частности, рассматривает выполнение функций базовой полосы частот в соответствии с принципом высокопроизводительного параллельного программирования в многоядерных системах, т.е. параллельно выполняемые задачи выполняются в один и тот же физический момент на отдельных ядрах. Кроме того, в работе рассмотрена $M^{[X]}/M/1\ PS$ для моделирования параллельной обработки в параллельных средах, т.е. когда различные задачи совместно используют один блок обработки путем чередования этапов выполнения каждого процесса с помощью фрагментов с разделением времени (также называемых временными интервалами)

В статье [1] авторы сосредоточены на оптимизации энергопотребления и

использования ресурсов за счет использования всего потенциала архитектуры C-RAN. В частности, предлагается новая система "эластичного"предоставления ресурсов под названием «Elastic-Net», позволяющую минимизировать энергопотребление как на сотовых станциях, так и в облаке, одновременно учитывая колебания спроса на пропускную способность для каждого пользователя. В данном решении, учитывая региональные колебания спроса на мощность, концепция динамически адаптирует активную плотность RRH, мощность передачи и размер виртуальных машин (VM) на основе колебаний трафика, чтобы минимизировать энергопотребление при максимальном использовании ресурсов. Авторы вводят идею «VBS-кластера», в которой они объединяют VBS, обслуживающие регион, в единый VBS-кластер, в то время как антенны RRH в регионе действуют как единая когерентная антенная решетка, распределенная по региону.

В данной работе рассматривается нисходящей поток связи системы $C-RAN$ и предполагается, что каждый пользователь обслуживается ближайшим активным RRH. $RRHs$ и пользователи распределяются в соответствии с двумя независимыми процессами Пуассона в $(R)^2$, обозначаемыми как Φ_r и $\Phi_u(t)$ соответственно. Распределение пользователей является функцией времени из-за их временных изменений. Пусть λ_r и $\lambda_u(t)$ обозначают плотность RRH и плотность пользователей, зависящую от времени, соответственно.Множество всех RRH обозначается через $\Omega = \{1, ..., L\}$ и $A \subseteq \Omega$ это множество активных RRH и $Z = \omega/A$ – множество неактивных RRH. Пусть также $\mu_a(t) \in [0, 1]$ обозначает коэффициент активности RRH, который указывает отношение активных RRH ко всем RRH, где $\lambda_r^a(t) = \mu_a(t)\lambda_r$ - зависящая от времени плотность активных RRH, а $\lambda_r^s(t) = (1 - \mu_a(t))\lambda_r$ – зависящая от времени плотность неактивных RRH. Общая полоса пропускания обозначается через B, а полоса пропускания для каждого пользователя задается через $B_u(t) = B\frac{\lambda_r^a(t)}{\lambda_u(t)}$. Хоть аналогичный анализ может быть применен для многоантенных систем, для простоты предполагается, что все $RRHs$ и пользователи оснащены одной антенной.

Авторы также концентрируются на эффекте затухания сигнала и используют обычно применяемую модель распространения сигнала следующим образом,

$$P_r = Gr^{-\alpha}hP.$$

где P_r, P, r и α обозначают принимаемую мощность, передаваемую мощность,

расстояние распространения и показатель потерь на пути, соответственно. Кроме того, G - это коэффициент потери пути, а случайная величина h используется для моделирования медленного затухания и она подчинается логарифмически нормальному распределению. В соответствии с этими предположениями принятый сигнал для типичного пользователя, обозначаемого как пользователь u^{th}, задается,

$$y_u = r_u^{-\alpha/2}\sqrt{Gh_uP}s_u + \sum_{j\neq u, j\in A} r_j^{-\alpha/2}\sqrt{Gh_jP}s_j + n_0$$

где r_u - расстояние между пользователем и обслуживающим его RRH, r_j расстояние между пользователем и j-м создающим помехи RRH и $n_0 \in C$ – аддитивный белый гауссовский шум ($AWGN$) в приемнике, обозначаемый как $n0 \sim CN(0, \sigma_n^2)$. Из пердыдущей формулы вычисляется отношение сигнал/помеха плюс шум (SINR):

$$SINR_u = \frac{h_u g(r_u)P}{\sum\limits_{j\neq u, j\in A} h_j g(r_j)P + \sigma_n^2},$$

где σ_n^2 – мощность шума, и $g_r = Gr^{-\alpha}$. Сбой происходит, если полученное значение $SINR$ падает ниже порогового значения γ, и операция проходит успешно, если $SINR_u > \gamma$. Взаимосвязь между вероятностью отказа (P_{out}) и вероятностью покрытия (P_{cov}) равна,

$$P_{cov} = 1 - P_{out} = Pr(SINR_u > \gamma)$$

Средняя пропускная способность каждого активного RRH, обозначаемая как R, определяется по формуле,

$$R = B(E)[log_2(1 + SINR_u)],$$

где $(E)[.]$ обозначает ожидаемое значение. Также определяется пропускная способность пользователя как средняя пропускная способность на одного пользователя, заданная с помощью,

$$R_u(t) = B_u(t)(E)[log_2(1 + SINR_u)].$$

Поскольку в $C - RAN$ разделена на блоки: $RRHs$ и $VBSs$, энергопо-

требление сети разделяется на две части: (i) энергопотребление RRH и транспортной сети и (ii) энергопотребление пула VBS. Для моделирования энергопотребление RRH, рассматривается линейая модель мощности:

$$P_{rrh} = \begin{cases} P^a_{rrh} + \frac{1}{\eta} P & P > 0, \\ P^s_{rrh} & P = 0, \end{cases}$$

где P^a_{rrh} – потребляемая мощность активной цепи, η – КПД усилителя мощности, P – мощность передачи и P^s_{rrh} – потребляемая мощность RRH в спящем режиме.

Поскольку данные, передаваемые между $RRHs$ и пулом VBS, представляют собой цифровые потоки ввода-вывода с избыточной дискретизацией в режиме реального времени порядка Гбит/с, энергопотребление транспортной сети оказывает значительное влияние на энергопотребление всей сети. В [1] рассматривается пассивная оптическая сеть (PON) для обеспечения недорогих соединений с высокой пропускной способностью и низкой задержкой между пулами $RRHs$ и VBS. PON содержит терминал оптической линии (OLT), который находится в пуле VBS и соединяет набор связанных оптических сетевых модулей (ONU) через одно оптоволокно. Реализация спящего режима в $ONUs$ является многообещающим решением для энергосбережения в PON; однако OLT не может перейти в спящий режим, и его энергопотребление фиксировано. В этой статье рассматривается режим быстрого/циклического сна, в котором состояние ONU чередуется между активным состоянием (когда RRH находится в активном состоянии) и спящим состоянием (когда RRH находится в спящем состоянии). Следовательно, энергопотребление транспортной сети выглядит следующим образом:

$$P_{tn} = P_{olt} + P_{onu},$$

где P_{olt} – энергопотребление OLT в пуле VBS, P_{onu} – энергопотребление ONU, заданное как,

$$P_{onu} = |A|P^a_{tl} + |Z|P^s_{tl},$$

где P^a_{tl} и P^s_{tl} – потребляемая мощность каждым ONU в активном и спящем режимах соответственно. Поскольку P_{olt} потребляется в пуле VBS, оно учитывается в энергопотребление пула VBS. Следовательно, RRH области и энергопотреб-

ление транспортной сети определяется по формуле:

$$P_{area} = \lambda_r^a(t)(P_{rrh}^a + \frac{1}{\eta}P + P_{tl}^a) + \lambda_r^s(t)(P_{rrh}^s + P_{tl}^s).$$

Работы [5–12], посвещаны решению задач анализа моделей C-RAN.

В работе [11] для оценки максимального расстояния от удаленных радиомодулей (RRHs) до центра обработки (пул BBUs) в облачных сетях радиодоступа используется система массового обслуживания $M/G/k$. Дистанционная обработка влечет за собой задержку прохождения сигнала в обоих направлениях между пулом BBU и RRH, которая включается в себя сумму задержек: 1) передачи, 2) постановки в канал, 3) ожидания в очереди, 4) обработки и 5) компонетов задержки распространения. В работе [11] характеризуется взаимосвязь между состоянием канала и максимальным расстояние между RRH и пулом BBU, что имеет последствия для балансировки нагрузки обработки и архитектурных решений относительно размещения центром обработки данных, в которых размещается пул блоков обработки базовых частот. Среднее время ожидания ω_q расчитывается следюущей формулой:

$$\omega_q = \frac{\rho\frac{b}{r_2}(\nu_a^2 + \nu_b^2)}{2(1-\rho)}f(\nu_a),$$

где $\rho = \lambda\frac{b}{r_2} < 1$ – загрузка системы, λ, ν_a – интенсивность потока заявок и коэффициент вариации интервалов между поступающими в систему требованиями; b, ν_b – среднее значение и коэффициент вариации длительности обслуживания заявок; $f(\nu_a)$ – корректирующая функция, рассчитываемая в зависитмости от значения коэффицента вариации ν_a, r_2 – пропускная способность участка:

$$f(\nu_a) = \begin{cases} e^{[-\frac{2(1-\rho)}{3\rho}\frac{(1-\nu_a^2)^2}{\nu_a^2+\nu_b^2}]}, & \nu_a < 1, \\ e^{[-(1-\rho)\frac{\nu_a^2-1}{\nu_a^2+4\nu_b^2}]} & \nu_a \geq 1 \end{cases}$$

Задержка обработки в пуле BBU – это время, затрачиваемое на обработку радиосигнала, например, демодуляцию, кодирование и обратное преобразования радиоресурса. Вычисление декодирования имеет свою производительность, непосредственно связанную с количеством циклов, выполняемых FEC - прямая коррекция ошибок является техникой кодирования/декодирования сигнала

с возможностью обнаружения ошибок и коррецией информации методом упреждения. И задержка обработки может быть выражена как $\frac{kbF}{pO} + J$. Пул BBU выполняет k циклов алгоритма FEC для каждого кодового болока, b – длина кодового блока в битах, которая может варьироваться в зависимости, например, от используемой технологии, скорости кодирования и алгоритма регулировки скорости "прокалывания" бита. Каждый бит кодовго блока обычно обрабатывается через два идентичных составляющих декодера объединенной сложности F, выраженных в операциях на бит. Тактовая частота процессора, выделенная для обработки каждого кодового блока, обозначается p (в Гц). Выделенный процессор имеет эффективность O в операция за такт. J – время, необходимое для обработки других беспроводных функций. Объединяя все компоненты задержки, рассмотренные выше, общая задержка между пулом BBU и RRH может быть выражена:

$$M\xi_T = 2\left(\frac{d_1}{c_0} + \frac{b}{r_1} + \frac{\rho\frac{b}{r_2}(\nu_a^2 + \nu_b^2)}{2(1-\rho)}f(\nu_a) + \frac{d_2}{c_0} + \frac{b}{r_2}\right) + \frac{kbF}{pO} + J$$

В статье [12] рассматривается облачная сеть радиодоступа ($C-RAN$), в которой серверы обработки сигналов базовой полосы (BBU), отделены от удаленных радиоголовок ($RRHs$). $RRHs$ образуют единый кластер, в то время как BBU образуют пул ресурсов. Каждый RRH может принимать случайный (пуассоновский), квазислучайный или скачкообразный трафик. Последнее аппроксимируется с помощью сложного пуассоновского процесса, в соответствии с которым пакеты вызовов с обычно распределенным размером пакета следуют пуассоновскому процессу. Вызов требует вычислительных ресурсов от $BBUs$ и блока радиоресурсов от обслуживающего RRH. Если какой-либо из двух аппаратов недоступен, происходит блокировка вызова. В противном случае новый вызов принимается в RRH. Авторы моделируют $C-RAN$ как систему с потерями и рассматриваются два разных случая: i) все RRH учитывают скачкообразный трафик и ii) некоторые RRH учитывают случайный трафик, некоторый квазислучайный трафик, а остальные RRH учитывают скачкообразный трафик. В обоих случаях показывается, что для вероятностей стационарного состояния существует решение, и предлагаются эффективные алгоритмы свертки для точного расчета времени и вероятностей перегрузки вызовов. Точность этих

алгоритмов проверяется с помощью моделирования.

В работе [6] предлагается для моделирования C-RAN использовать единую систему массового обслуживания с общим пулом обслуживающих устройств, а именно многоприборную систему с k - приборами. Глобальный планировщик выделяет вычислительные ресурсы для каждого выполняемого задания кодирования (нисходящий канал) или декодирования (восходящий канал).

Авторы предполагают, что в vBBUs (в частности, функции виртуального кодирования / декодирования) поступает пуассоновский поток вызовов, т.е. время между приходами выполняемых заданий BBU распределено экспоненциально. Это разумно отражает тот факт, что существует достаточно большое количество антенн, которые не синхронизированы. Возникновение заданий является результатом суперпозиции независимых точечных процессов. Это оправдывает предположение пуассовонского вхоядщего потока требований. На практике кадры приходят с фиксированными относительными фазами. Таким образом, рассмотрение пуассновского входящего потока в некотором смысле является предположением наихудшего случая. Поступления требований не синхронизированы, поскольку RRHs находятся на разных расстояниях от пула BBU. Кроме того, при отсутствии выделенных каналов задержка на входе (время между прибытиями) может сильно варьироваться из-за сетевого трафика.

Параллельное выполнение задач кодирования и декодирования в многоприборной системе с k - обслуживающими приборами, может быть смоделировано системой массового обслуживания $M^{[X]}/M/k$. Время выполнения каждой подзадачи зависит от рабочей нагрузки, а также от сетевой подфункции, которую она реализует. Количество параллельно выполняемых подзадач, принадлежащих сетевой подфункции, является переменным. Таким образом, рассматривается объем нефиксированного размера, который поступает в момент поступления каждого запроса. Время между прибытиями требований экспоненциально с параметром λ. Размер пакета B не зависит от состояния системы.

В случае Cloud-RAN полный функциональный параллелизм невозможен, поскольку некоторые процедуры базового блока (например, IFFT, модуляция и т.д.) требуют последовательного выполнения. Однако параллелизм данных функций BBU (в частности, декодирования и кодирования) обещает значительное повышение производительности. Эти утверждения тщательно изучены в [7,8]. Результаты показывают, что время выполнения функций BBU может быть

значительно сокращено при выполнении параллельной обработки в подкадре, т.е. за счет параллельного выполнения либо UEs – пользвательского треминала, либо даже меньших блоков данных, так называемых CBs – блоков кода.

A. Параллельная обработка со стороны UEs

В LTE несколько UEs – пользовательских треминалов могут обслуживаться в субкадре продолжительностью 1 миллисекунда. Максимальное и минимальное количество UEs, запланированных для каждого подкадра, определяется пропускной способностью eNB – базовая станция сети LTE. LTE поддерживает масштабируемую полосу пропускания 1.25, 2.5, 5, 10 и 20 МГц. В подкадре каждый запланированный UE получает TB - транспортный блок (а именно, группу радиоресурсов в форме RB - ресурсный блок) либо для передачи, либо для приема. Например, при рассмотрении eNB с частотой 20 МГц доступно 100 RBs. Согласно LTE [13], минимальное количество RBs, выделенных на UE, равно 6. Следовательно, максимальное количество подключенных UE на подкадр задается $b_{max} = 100/6$. TBs определяется планировщиком радиосвязи в зависимости от условий индивидуального радиоканала UEs, а также от объема трафика в ячейке.

Из вышеописанного следует, что параллельная обработка базовой полосы (в частности, канальное кодирование) субкадров может быть смоделирована как система массового обслуживания $M^{[X]}/G/c$. При рассмотрении распараллеливания для каждого UE количество требований в пакете соответствует количеству UE, запланированных в подкадре LTE, например, количеству требований декодирования в миллисекунду в диапазоне eNB 20 МГц от 1 до 16. Затем подкадр содержит переменное количество UE, которое представлено случайной величиной B.

Предполагается, что время обработки задания (а именно TB) экспоненциально. Это предположение предназначено для учета случайности во времени обработки UEs из-за недетерминированного поведения функции кодирования канала. Например, время выполнения декодирования одного UE может варьироваться от нескольких десятков микросекунд до почти всего временного бюджета, т.е. 2000 микросекунд [14]. На практике это время обслуживания охватывает время отклика каждого компонента системы облачных вычислений, т.е. процессорных блоков, оперативной памяти, внутренних шин, механизма виртуализации, каналов передачи данных и т.д. В дальнейшем предполагается,

что время обслуживания TB (т.е. требования) распределяется экспоненциально со средним значением $1/\mu$. Если предположить, что количество В UEs на подкадр геометрически распределено со средним значением $1/(1-q)$ (то есть $\mathbb{P}(B=k)=(1-q)q^{k-1}$ для $k \geq 1$), полное время обслуживания кадра затем экспоненциально распределяется со средним значением $1/((1-q)\mu)$.

Геометрическое распределение как дискретный аналог экспоненциального распределения отражает изменчивость запланированных UEs в подкадре. Размер B зависит как от количества UEs, требующих обслуживания в соте, так и от условий радиоканала каждого из них. Кроме того, B тесно связан со стратегией планирования радиопередач (например, круговой отбор, пропорциональная выбор и т.д.). Количество UEs всегда варьируется от 1 до b_{max}, где последнее количество зависит от пропускной способности $eNBs$. В LTE b_{max} достигается, когда пользователи испытывают плохие условия радиосвязи, т.е. при использовании надежной модуляции в виде $QPSK$ и высокой степени избыточности. Для средних условий радиосвязи и ненасыщенных $eNBs$ более вероятно наличие небольших партий UEs. Геометрическое распределение предназначено для отражения сочетания условий радиосвязи в UEs и их потребностей в передаче.

Если предположить, что вычислительная платформа имеет неограниченный буфер, стабильность системы требует:

$$\rho = \frac{\lambda}{\mu C} < 1. \tag{1}$$

Если время пребывания превышает некоторый порог (т.е. 1 миллисекунду для кодирования и 2 миллисекунды для декодирования), то подкадр теряется. Если вероятность того, что время пребывания превысит пороговое значение, была небольшой, то можно аппроксимировать скорость потери субкадра по этой вероятности. Стоит отметить, что в LTE подтверждения передачи и приема обрабатываются для каждого подкадра гибридным процессом автоматического повторного запроса ($HARQ$). Когда TB теряется, весь подкадр отправляется повторно.

В. Параллельная обработка со стороны CBs

В LTE, когда TB слишком велик, он разбивается на более мелкие блоки данных, называемые CBs. Если предположить, что время обработки CB экспоненциально со средним значением $1/\mu$, снова получим модель $M^{[X]}/M/c$, где

размер партии равен количеству CB в TB. Если это число распределено геометрически, то время обслуживания TB экспоненциально, как предполагалось выше. Ключевое отличие теперь заключается в том, что отдельные CBs обрабатываются параллельно ядрами C. Планировщик способен выделить обслуживающее устройство для каждого CB благодаря более атомарной декомпозиции подкадров и TBs.

C. Отсутсвие параллельной обработки

Если обработка TBs или CBs не является параллельной, планирование основывается на подкадрах, как представлено в [15]. Все еще предполагая многоприборную систему, где подкадры поступают в соответствии с процессом Пуассона, рассмотрим систему массового обслуживания $M/G/k$. Делая экспоненциальные допущения для времени обслуживания CBs и TBs, а также предполагая геометрическое число CBs на TB, мы получаем систему массового обслуживания $M/M/k$ с очередью, которая хорошо известна в литературе по организации массового обслуживания [16].

2 Анализ модели облачной сети радиодоступа

Из анализа, проведенного в предыдущем разделе, модель $M^{[X]}/G/C$ может быть разумно использована для оценки времени обработки подкадра в облачной архитектуре, основанной на многоядерной платформе. В то время как время пребывания произвольного требования в партии было проанализировано в [17], время пребывания всей партии, по-видимому, получило меньше внимания в технической литературе. В этом разделе мы выводим преобразование Лапласа для этой величины; это в конечном итоге позволяет нам получить асимптотическую оценку вероятности превышения большого порогового значения.

Рассмотрим система $M^{[X]}/G/C$ с пакетами размера B, поступающими в соответствии с пуассоновским процессом с интенсивность λ. Время обслуживания задачи в пакете экспоненциально со средним значением $1/\mu$. Мы предполагаем, что условие устойчивости (1) выполняется так, что существует стационарный режим. Количество N заданий в системе в стационарном режиме таково, что [24]

$$\phi(z) \stackrel{\text{def}}{=} \mathbb{E}\left(z^N\right) = \frac{\sum_{k=0}^{C-1}(C-k)p_k z^k}{C - \frac{\lambda}{\mu} z\left(\frac{1-B(z)}{1-z}\right)} \tag{2}$$

где $p_k = \mathbb{P}(N = k)$ и $B(z)$ – производящая функция с размера пакета B, т.е. $B(z) = \sum_{k=0}^{\infty} \mathbb{P}(B = k)z^k$. Как показано в [17], вероятности p_k для $k \geq 1$ удовлетворяют уровнениям баланса:

$$\begin{aligned}
p_1 &= \frac{\lambda}{\mu} p_0, \\
p_k &= \left(1 + \frac{\lambda - \mu}{k\mu}\right) p_{k-1} - \frac{\lambda}{\mu k} \sum_{\ell=0}^{k-2} p_\ell b_{k-1-\ell} \quad 2 \leq k \leq C, \\
p_k &= \left(1 + \frac{\lambda}{\mu C}\right) p_{k-1} - \frac{\lambda}{\mu C} \sum_{\ell=0}^{k-2} p_\ell b_{k-1-\ell} \quad k \geq C,
\end{aligned}$$

где b_l - вероятность того, что размер партии равен l. Мы видим, в частности, что вероятности p_k для $k = 2, .., C$ линейно зависят от p_0, которые в конечном итоге могут быть вычислены с использованием условия нормализации $\sum_{k=0}^{C-1}(C - k)p_k = C(1 - \rho)$

Рассматривается пакет размера b, поступающий в момент времени t_0 и обнаруживающий n заданий в очереди. Мы рассмотрим два случая, рисунок 3:

Случай $n \geq C$: В этом случае первое задание помеченного пакета должно подождать, прежде чем оно поступит на обслуживающий прибор.

Случай $n < C$: В этом случае $b \wedge (C - n) \stackrel{\text{def}}{=} \min(b, C - n)$ задания помеченного пакета немедленно поступают в эксплуатацию; $0 \vee (b + n - C) \stackrel{\text{def}}{=} \max(0, b + n - C)$ задания должны подождать, прежде чем приступить к обслуживанию.

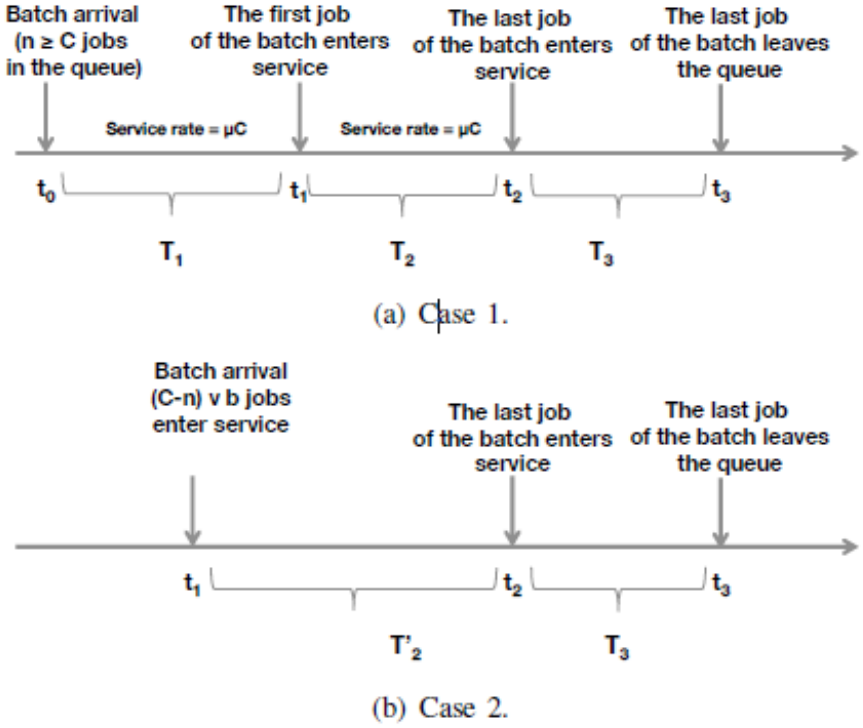

Рисунок 3 – Варианты поступления пакетов

Анализ первого случая.

В случае $n \geq C$ помеченный пакет должен будет подождать определенное время, прежде чем первое задание поступит на обслуживание. Пусть t_1 обозначает время, в которое первое задание из помеченного пакета начинает свое обслуживание. Очевидно, что мы имеем, что $T_1 = t_1 - t_0$ равно сумме $n - C + 1$ независимых случайных величин, экспоненциально распределённых со средним значением $1/(\mu C)$. Преобразование Лапласа T_1 определяется для $R(s) \geq 0$ с помощью

$$\mathbb{E}_b\left(e^{-sT_1}\right) = \left(\frac{\mu C}{s + \mu C}\right)^{n-C+1} \tag{3}$$

где E_b - математическое ожидания размера партии b.

Пусть t_2 обозначает время, в которое последнее задание из пакета по-

ступает на обслуживание. Разница $T_2 = t_2 - t_1$, очевидно, является суммой $b-1$ независимых экспоненциальных случайных величин со средним значением $1/(\mu C)$ (величина μC является скоростью обслуживания системы); преобразование Лапласа этой разницы равно

$$\mathbb{E}_b\left(e^{-sT_2}\right) = \left(\frac{\mu C}{s+\mu C}\right)^{b-1} \tag{4}$$

Чтобы полностью определить время пребывания помеченного пакета, необходимо знать количество y_b заданий, которые принадлежат этому пакету и которые находятся в очереди, когда пакет загружен. Последнее задание пакета начинает свое обслуживание. Пусть $t_1 = \tau_1 < \tau_2 < ... < \tau_b = t_2$ обозначает время завершения обслуживания заданий (не обязательно относящихся к помеченному пакету) в интервале $[t_1; t_2]$. (Обращаем внимание, что точка t_1, соответствующая времени поступления на обслуживание первого задания из помеченной партии, сама по себе является временем завершения обслуживания одного задания, присутствующего в очереди при поступлении помеченной партии.) По определению τ_n – это время, в которое n-е задание помеченного пакета поступает в эксплуатацию.

Обозначим через y_n количество заданий, принадлежащих помеченному пакету в момент времени τ_n^+. Тогда последовательность (y_n) представляет собой цепочку Маркова, для которой условные вероятности перехода могут быть выражены в терминах чисел Стирлинга второго рода $S(n;k)$ определяется для $0 \leq k \leq n$ с помощью

$$S(n,k) = \sum_{j=0}^{k} \frac{(-1)^{k-j}}{(k-j)!j!} j^n \tag{5}$$

Числа Стирлинга таковы, что $S(n,n) = 1$ для $n \geq 0$, $S(n,1) = 1$ и $S(n,0) = 0$ и удовлетворяют рекурсии для $n \geq 0$ и $k >= 1$

$$S(n+1,k) = kS(n,k) + S(n,k-1) \tag{6}$$

Чтобы сформулировать результаты, альтернативно используются многочлены $\mathcal{A}_{n,p}(x)$ определяющиеся с помощью чисел Стирлинга следующим образом:

$$A_{n,p}(x) = p! \sum_{j=0}^{n} \left(\begin{array}{c} n \\ j \end{array} \right) S(j,p) x^{n-j}. \tag{7}$$

Многочлены $\mathcal{A}_{n,p}(x)$ удовлетворяют рекурсии для $n, p \geq 0$

$$\begin{cases} \mathcal{A}_{n,p}(x) = (x+p)\mathcal{A}_{n-1,p}(x) + p\mathcal{A}_{n-1,p-1}(x), \\ \mathcal{A}_{n,p}(0) = p! S(n,p). \end{cases} \tag{8}$$

Имеем следующий результат. Лемма 1: Условные вероятности перехода цепи Маркова (y_n) задаются для $k \geq l$ с помощью

$$\begin{aligned} &\mathbb{P}\left(y_n = k \mid y_1 = \ell\right) = \\ &= \frac{(C-\ell)!}{(C-k)!C^{n-1}} \sum_{m=0}^{n-1} \left(\begin{array}{c} n-1 \\ m \end{array} \right) S(m, k-\ell) \ell^{n-1-m} = \\ &= \frac{1}{C^{n-1}} \left(\begin{array}{c} C-\ell \\ k-\ell \end{array} \right) \mathcal{A}_{n-1,k-\ell}(\ell). \end{aligned} \tag{9}$$

Из приведенной выше леммы выводится тождество

$$\begin{aligned} &\frac{1}{C^{n-1}} \sum_{k=0}^{C} \left(\begin{array}{c} C-\ell \\ k-\ell \end{array} \right) k \mathcal{A}_{n-1,k-\ell}(\ell) = \\ &= C\left(1 - \left(1 - \frac{\ell}{C}\right)\left(1 - \frac{1}{C}\right)^{n-1}\right), \end{aligned} \tag{10}$$

где используется тот факт, что

$$E\left(y_n \mid y_1 = \ell\right) = C\left(1 - \left(1 - \frac{\ell}{C}\right)\left(1 - \frac{1}{C}\right)^{n-1}\right) \tag{11}$$

С учетом приведенных выше результатов, когда b-е задание помеченного пакета поступает на обслуживание, в очереди есть y_b заданий этого пакета. Время T_3 для выполнения этих заданий равно

$$T_3 = \mathcal{E}\left(y_b \mu\right) + \mathcal{E}\left(\left(y_b - 1\right)\mu\right) + \ldots + \mathcal{E}(\mu) \tag{12}$$

где $\mathcal{E}(k\mu)$ для $k = 1, ..., y_b$ - независимые случайные величины со средним

значением $1/(k\mu)$. Преобразование Лапласа для T_3, зная, что y_b равно

$$\mathbb{E}_b\left(e^{-sT_3} \mid y_b = k\right) = \frac{k!}{\prod_{\ell=1}^{k}\left(\frac{s}{\mu}+\ell\right)} = \frac{k!}{\left(\frac{s}{\mu}+1\right)_k}, \tag{13}$$

где $(x)_k$ - символ Почхаммера (он же возрастающий факториал) определяется через $(x)_k = x(x+1)...(x+k-1)$. Используя лемму 1,следует, что преобразование Лапласа времени пребывания T партии размером b в системе, когда в очереди по прибытии находится $n \geq C$ клиентов, равно

$$\begin{aligned}&\mathbb{E}_b\left(e^{-sT} \mid N = n \geq C\right) = \\ &= \quad \frac{C!}{C^b}\left(\frac{\mu C}{s+\mu C}\right)^{n+b-C}\sum_{k=0}^{C}\frac{S(b,k)}{(C-k)!}\frac{k!}{\left(\frac{s}{\mu}+1\right)_k},\end{aligned} \tag{14}$$

который может быть переписан с помощью многочленов $\mathcal{A}_{n,p}(x)$, определенных уравнением (7) как

$$\begin{aligned}&\mathbb{E}_b\left(e^{-sT} \mid N = n \geq C\right) = \\ &= \frac{1}{C^{b-1}}\left(\frac{\mu C}{s+\mu C}\right)^{n+b-C}\sum_{k=0}^{C}\left(\begin{array}{c} C-1 \\ k-1 \end{array}\right)\mathcal{A}_{b,k-1}(1)\frac{1}{\left(\frac{s}{\mu}+1\right)_k}.\end{aligned} \tag{15}$$

Анализ второго случая.

Когда количество n заданий в очереди становится меньше C при поступлении помеченной партии размером b, тогда клиенты $b \wedge (C-n)$ немедленно начинают свое обслуживание. Давайте сначала предположим, что $b+n > C$. Принимая поступление помеченной партии в качестве источника времени, последнее задание помеченной партии поступает на обслуживание в случайное время T_2' с преобразованием Лапласа

$$\mathbb{E}\left(e^{-sT_2'}\right) = \left(\frac{\mu C}{s+\mu C}\right)^{n+b-C}. \tag{16}$$

Количество заданий помеченного пакета, присутствующих в системе на момент поступления последнего задания на обслуживание, равно Y_n таким образом, что

$$\mathbb{P}\left(Y_n = k\right) = \mathbb{P}\left(y_{b+n-C} = k \mid y_1 = C - n\right) =$$
$$= \quad \frac{1}{C^{n+b-C-1}} \begin{pmatrix} n \\ k+n-C \end{pmatrix} \mathcal{A}_{n+b-C-1,k+n-C}(C-n), \tag{17}$$

используя уравнение (9). Для заданного значения $Y_n = k$ время T_3, необходимое для обслуживания всех заданий помеченного пакета, имеет преобразование Лапласа, заданное уравнением (13). Используя лемму 1, приходим к выводу, что в предположении $n < C$ и $b + n >$C время пребывания T помеченной партии имеет преобразование Лапласа

$$\mathbb{E}_b\left(e^{-sT} \mid N = n, b + N > C, N < C\right) =$$
$$= \quad \left(\frac{\mu C}{z + \mu C}\right)^{n+b-C} \sum_{k=C-n}^{C} \mathbb{P}\left(Y_n = k\right) \frac{k!}{\left(\frac{s}{\mu} + 1\right)_k} \tag{18}$$

и, следовательно,

$$\mathbb{E}_b\left(e^{-sT} \mid N = n < C, b + N > C\right) =$$
$$= \left(\frac{\mu C}{z + \mu C}\right)^{n+b-C} \tau(n, b; s), \tag{19}$$

где

$$\tau(n, b, ; s) = \frac{1}{C^{n+b-C-1}} \sum_{k=C-n}^{C} \begin{pmatrix} n \\ k+n-C \end{pmatrix} \times \mathcal{A}_{n+b-C-1,k+n-C}(C-n) \frac{k!}{\left(\frac{s}{\mu} + 1\right)_k}. \tag{20}$$

Когда $b + n \leq C$, все задания помеченного пакета поступают на обслуживание сразу после прибытия и преобразования Лапласа времени пребывания является

$$\mathbb{E}_b\left(e^{-sT} \mid N = n, b + n \leq C\right) = \frac{b!}{\left(\frac{s}{\mu} + 1\right)_b}. \tag{21}$$

Основной результат.

Используя результаты предыдущих разделов, определяем преобразование Лапласа $\Phi(s) = \mathbb{E}(e^{-sT})$ времени пребывания пакета в очереди $M^{[X]}/M/C$.

Теорема 1: Преобразование $\Phi(s)$ Лапласа задается формулой

$$\begin{aligned}\Phi(s) =&\beta(s)\left(\phi\left(\frac{\mu C}{s+\mu C}\right)-\phi_C\left(\frac{\mu C}{s+\mu C}\right)\right)\\&+\mathbb{E}\left(\frac{B!}{\left(\frac{s}{\mu}+1\right)_B}\mathbb{P}(N\leq C-B)\right)\\&+\sum_{n=0}^{C-1}p_n\mathbb{E}\left(\tau(n,B;s)\left(\frac{\mu C}{s+\mu C}\right)^{n+B-C}\right),\end{aligned}\tag{22}$$

где

$$\beta(s)=\\ \mathbb{E}\left(\frac{1}{C^{B-1}}\left(\frac{\mu C}{s+\mu C}\right)^{B-C}\sum_{k=0}^{C}\left(\begin{array}{c}C-1\\k-1\end{array}\right)\frac{\mathcal{A}_{B,k-1}(1)}{\left(\frac{s}{\mu}+1\right)_k}\right),\tag{23}$$

функция $\phi_C(z)$ задается формулой

$$\phi_C(z)=\sum_{n=0}^{C-1}p_n z^n\tag{24}$$

и $\tau(n,b,s)$ определяемые уравнением (20)

Доказательство: Определяя размер партии b, получаем из двух предыдущих разделов:

$$\begin{aligned}\mathbb{E}_b\left(e^{-sT}\right)=&\beta_b(s)\sum_{n=C}^{\infty}p_n\left(\frac{\mu C}{s+\mu C}\right)^n\\&+\sum_{n=0}^{C-1}p_n\tau(n,b;s)\left(\frac{\mu C}{s+\mu C}\right)^{n+b-C}\\&+\frac{b!}{\left(\frac{s}{\mu}+1\right)_b}\mathbb{P}(N\leq C-b)\end{aligned}\tag{25}$$

с

$$\beta_b(s)=\frac{C!}{C^b}\left(\frac{\mu C}{s+\mu C}\right)^{b-C}\sum_{k=0}^{C}\frac{S(b,k)}{(C-k)!}\frac{k!}{\left(\frac{s}{\mu}+1\right)_k}\tag{26}$$

и $\tau(n,b,s)$ определяется уравнением (20). Используем тот факт, что $\tau(n,b,s)=0$, если $b<C-n$ в приведенном выше уравнении. Путем декондиционирования

в зависимости от размера партии получается уравнение (22).

Следуя [17], определим $z1$ как корень с наименьшим модулем в уравнении

$$V(z) \stackrel{\text{def}}{=} C - \frac{\lambda}{\mu} z \left(\frac{1 - B(z)}{1 - z} \right) = 0; \tag{27}$$

корень $z1$ действителен и больше 1. Отрицательное действительное число

$$s_1 = -\mu C \left(1 - \frac{1}{z_1} \right) \tag{28}$$

является особенностью с наименьшим модулем преобразования $\Phi(s)$ Лапласа, если $s_1 > -\mu$ (а именно, $z_1 < C/(C-1)$).

Следствие 1: Если $s_1 > -\mu$, то когда t стремится к бесконечности

$$\mathbb{P}(T > t) \sim \frac{\mu C U(z_1) \beta(s_1)}{s_1 z_1^2 V'(z_1)} e^{s_1 t} \tag{29}$$

где $U(z) = \sum_{k=0}^{C-1} (C-k) p_k z^k$. Если $s_1 < -\mu$ тогда хвост распределения T таков, что когда t стремится к бесконечности

$$\mathbb{P}(T > t) \sim \kappa e^{-\mu t}, \tag{30}$$

где

$$\begin{aligned} \kappa &= \mathbb{E}(B\mathbb{P}(N+ \mid B \le C)) \\ &+ C\mathbb{E}\left(\left(\frac{C}{C-1} \right)^B - 1 \right) \sum_{n=0}^{\infty} p_{C+n} \left(\frac{C}{C-1} \right)^n \\ &+ \sum_{n=0}^{C-1} p_n C \mathbb{E}\left(\mathbb{1}_{\{B > C-n\}} \left(\left(\frac{C}{C-1} \right)^{n+B-C} - \frac{n}{C-1} \right) \right). \end{aligned} \tag{31}$$

Доказательство: Когда $s1 > -\mu$, корень с наименьшим модулем преобразования Лапласа $\Phi(s)$ равен $s1$, а оценка (29) непосредственно за этим следует использование стандартных результатов для преобразований Лапласа.

Когда $s1 > -\mu$, корень с наименьшим модулем равен $-\mu$. У нас есть для s - это окрестности $-\mu$

$$\begin{aligned}\beta(s) &\sim \frac{\mu}{\mu+s} \\ &\times \mathbb{E}\left(\frac{1}{C^{B-1}}\left(\frac{\mu C}{C-1}\right)^{B-C}\sum_{k=0}^{C}\left(\begin{array}{c}C-1\\k-1\end{array}\right)k\mathcal{A}_{B,k-1}(1)\right) \\ &= \frac{\mu}{\mu+s}C\mathbb{E}\left(\left(\frac{C}{C-1}\right)^{B}-1\right)\sum_{n=0}^{\infty}p_{C+n}\left(\frac{C}{C-1}\right)^{n},\end{aligned} \tag{32}$$

где мы использовали уравнение (10) для $l = 1$.

Кроме того, при тех же условиях,

$$\begin{aligned}&\mathbb{E}\left(\frac{B!}{\left(\frac{s}{\mu}+1\right)_B}\mathbb{P}(N \le C-B)\right) \\ &\qquad\qquad \frac{\mu}{\mu+s}\mathbb{E}(B\mathbb{P}(N+B \le C))\end{aligned} \tag{33}$$

и

$$\begin{aligned}&\sum_{n=0}^{C-1}p_n\mathbb{E}\left(\tau(n,B;s)\left(\frac{\mu C}{s+\mu C}\right)^{n+B-C}\right) \\ &\sim \frac{\mu}{\mu+s}\sum_{n=0}^{C-1}p_n\mathbb{E}\left(\left(\frac{C}{C-1}\right)^{n+B-C}\frac{1}{C^{n+B-C-1}}\right. \\ &\left.\times\sum_{k=C-n}^{C}\left(\begin{array}{c}n\\k+n-C\end{array}\right)k\mathcal{A}_{n+B-C-1,k+n-C}(C-n)\right) \\ &= \frac{\mu}{\mu+s}\sum_{n=0}^{C-1}p_nC \\ &\times\mathbb{E}\left(1_{\{B>C-n\}}\left(\left(\frac{C}{C-1}\right)^{n+B-C}-\frac{n}{C-1}\right)\right),\end{aligned} \tag{34}$$

где используется уравнение (10) для $l = C-n$. Объединение приведенных выше вычислений остатков приводит к уравнению (30).

Следствие 1 гласит, что когда пропускная способность системы достаточно велика, в конце времени пребывания пакета преобладает время обслуживания одного задания. Также стоит отметить, что вопреки тому, что указано в [18], тот же результат справедлив для скорости затухания времени пребывания задания

в системе. Наконец, при большом C и при умеренных значениях загрузки и среднего размера партии, $\kappa \sim \mathbb{E}(B\mathbb{P}(N+B \leq C)) \sim \mathbb{E}(B)$. Это означает, что существует примерно мультипликативный коэффициент $\mathbb{E}(B)$ между конечным временем пребывания пакета и временем выполнения задания.

Когда размер партии геометрически распределен со средним значением $1/(1-q)$ мы имеем $s_1 = -(1-q)\mu C(1-\rho)$ и

$$z_1 = \frac{C}{qC + \frac{\lambda}{\mu}} > 1 \text{ for } C > \frac{\lambda}{(1-q)\mu} \tag{35}$$

У нас есть $z_1 < \frac{C}{C-1}$ тогда и только тогда, когда $\rho > 1 - \frac{1}{C(1-q)}$.

ЗАКЛЮЧЕНИЕ

В данной работе приведен обзор математических моделей, используемых для анализа и оптимизации $C - RAN$. Описана система массового обслуживания типа $M^{[X]}/M/c$, используемая для моделирования $C - RAN$.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1 *Hajisami, A.* Elastic resource provisioning for increased energy efficiency and resource utilization in cloud-RANs / A. Hajisami, X. T. Tuyen, A. Younis, D. Pompili // *Computer Networks*. — 2020. — Vol. 172. — P. 107.

2 *Tang, J.* System cost minimization in cloud RAN with limited fronthaul capacity / J. Tang, P. T. Wee, Q. S. Tony, B. Liang. — Institute of Electrical and Electronics Engineers (IEEE), 2017. — Pp. 3371–3384.

3 *Rodriguez, V. Q.* Contribution to the design and the implementation of a cloud radio access network / V. Q. Rodriguez, F. Guillemin // *ArXiv*. — 2019. — Vol. 5. — P. 28.

4 *Bsebsu, A.* Fast optimization of cache-enabled cloud-RAN using determinantal point process / A. Bsebsu, G. Zheng, S. Lambotharan // *Physical Communication*. — 2021. — Vol. 46. — P. 101.

5 *Rodriguez, V. Q.* On dimensioning cloud-RAN systems / V. Q. Rodriguez, F. Guillemin // Proceedings of the 11th EAI International Conference on Performance Evaluation Methodologies and Tools. — Venice: ACM, 2017. — Pp. 132–139.

6 *Rodriguez, V. Q.* Cloud-RAN modeling based on parallel processing / V. Q. Rodriguez, F. Guillemin // *IEEE Journal on Selected Areas in Communications*. — 2018. — Vol. 36, no. 3. — Pp. 457–468.

7 *Rodriguez, V. Q.* VNF modeling towards the cloud-RAN implementation / V. Q. Rodriguez, F. Guillemin // International Conference on Networked Systems (NetSys). — Gottingen: IEEE, 2017. — Pp. 1–8.

8 *Rodriguez, V. Q.* Towards the deployment of a fully centralized Cloud-RAN architecture / V. Q. Rodriguez, F. Guillemin // 13th International Wireless Communications and Mobile Computing Conference (IWCMC). — Valencia: IEEE, 2017. — Pp. 1055–1060.

9 *Rodriguez, V. Q.* Performance analysis of VNFs for sizing cloud-RAN infrastructures / V. Q. Rodriguez, F. Guillemin // Conference on Network Function Virtualization and Software Defined Networks (NFV-SDN). — Berlin: IEEE, 2017. — Pp. 1–6.

10 *Rodriguez, V. Q.* Performance analysis of resource pooling for network function virtualization / V. Q. Rodriguez, F. Guillemin // 17th International Telecommunications Network Strategy and Planning Symposium (Networks). — Montreal: IEEE, 2016. — Pp. 158–163.

11 *Дармолад, А.В.* Оценка максимального расстояния от удаленных радиомодулей до центра обработки в облачных сетях радиодоступа / А.В. Дармолад, Д.Н. Биксалина, Э.С. Сопин // *Институт проблем информатики, ФИЦ УИ РАН*. — 2020. — Vol. 1. — Pp. 73–78.

12 *Chousainov, I. A.* An analytical framework of a c-RAN supporting random, quasi-random and bursty traffic / I. A. Chousainov, I. Moscholios, P. Sarigiannidis, A. Kaloxylos, M. Logothetis // *Computer Networks*. — 2020. — Vol. 180. — P. 107.

13 *Dahlman, E.* 4G: LTE/LTE-Advanced for Mobile Broadband / E. Dahlman, S. Parkvall, J. Skold. — Amsterdam: Elsevier, 2014. — P. 510.

14 *Nikaein, N.* OpenAirInterface / N. Nikaein, M. K. Marina, S. Manickam, A. Dawson, R. Knopp, C. Bonnet // *ACM SIGCOMM Computer Communication Review*. — 2014. — Vol. 44, no. 5. — Pp. 33–38.

15 *Bhaumik, S.* CloudIQ / S. Bhaumik, S. P. Chandrabose, M. K. Jataprolu, G. Kumar, A. Muralidhar, V. Srinivasan, T. Woo // Proceedings of the 18th annual international conference on Mobile computing and networking. — Istanbul: ACM, 2012. — Pp. 125–136.

16 *Kleinrock, L.* Queueing systems: computer applications / L. Kleinrock. — New York: Wiley-Interscience, 1976. — P. 576.

17 *Nelson, R.* Performance analysis of parallel processing systems / R. Nelson, D. Towsley, A.N. Tantawi // *IEEE Transactions on Software Engineering*. — 1988. — Vol. 14, no. 4. — Pp. 532–540.

18 *Cromie, M. V.* Further results for the queueing system m x /m/c / M. V. Cromie, M. L. Chaudhry, W. K. Grassmann // *The Journal of the Operational Research Society*. — 1979. — Vol. 30, no. 8. — P. 755.