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

Кафедра системного анализа и автоматического управления

МАТЕМАТИЧЕСКИЕ МОДЕЛИ ОБЛАЧНОЙ СЕТИ РАДИОДОСТУПА

КУРСОВАЯ РАБОТА

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

Научный руководитель
доцент, к. ф.-м. н. _____ Е. П. Станкевич

Заведующий кафедрой
к. ф.-м. н., доцент _____ И. Е. Тананко

Саратов 2023

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.

В настоящее время телекоммуникационный сектор разрабатывает сетевую архитектуру для предполагаемых сотовых систем $5G$. Облачные сети радиодоступа ($C-RAN$) представляют собой новые архитектуры мобильной связи, разработанные для поддержки высоких скоростей передачи данных экономичным образом и, как ожидается, обеспечивающие низкие задержки, высокую гибкость, спектральную эффективность и низкое энергопотребление, для соблюдения требований $5G$.

Целью данной работы является изучение математических моделей сетей облачного радиодоступа ($C-RAN$).

Работа состоит из двух разделов.

В первом разделе рассматриваются основные принципы функционирования $C-RAN$.

Во втором разделе описываются некоторые математические модели $C-RAN$.

1 Основные понятия и принципы функционирования облачной сети радиодоступа

$Cloud\ RAN$ – архитектура мобильной сети, которая хорошо подходит для решения проблемы роста расходов на сеть доступа, что до сих пор является одной из главных проблем в бизнесе операторов связи. Кроме того, эта новая концепция позволяет сетевым операторам предлагать инновационные услуги и разворачивать варитуализированные сети по требованиями заказчика. Эта задача соответствует общей концепции виртуализации сетевых функций (NFV) [1], которая как раз и состоит в замене сетевых функций, работающих аппартано на выделенном частном оборудовании, открытыми программными приложениями, работающими на общих коммерческих готовых серверах ($COTS$) на облачных платформах. Таким образом, сетевые операторы могут создавать экземпляры виртуальных сетевых функций ($VNFs$) с высокой скоростью и эффективностью в различных сетевых местоположениях в соответствии с требованиями заказчика [2].

Основополагающий принцип виртуализации заключается в размещении сетевых функций на одной или нескольких виртуальных машинах (VMs). $VNFs$ развертываются поверх виртуализированной инфраструктуры, которая может охватывать более одного физического местоположения и даже инфраструктуру облачных вычислений. В идеале $VNFs$ должны располагаться там, где они наиболее эффективны с точки зрения производительности и стоимости. $VNFs$ могут размещаться в центрах обработки данных, сетевых узлах или даже на устройствах конечных пользователей в зависимости от требуемой производительности (в частности, задержки) и ресурсов (пропускная способность, хранение и вычисления).

Однако облачные технологии и коммерческое использования сетевых функций порождают новые проблемы, особенно при виртуализации сетей беспроводного доступа. Это, в частности, относится к Cloud-RAN (C-RAN), целью которой является реализация обработки радиосигналов в базовых блоках (BBU) с использованием программного обеспечения. BBU - представляет из себя небольшой сервер, который может быть установлен либо в телекоммуникационной стойке (если существует какое-либо выделенное помещение), либо в климатическом шкафу на крыше здания т.е. в непосредтсвенной близости к eNB - базовой станции сети стандарта LTE. В настоящее время изучается несколь-

ко функциональных разделений, наиболее амбициозным из них, безусловно, является полная централизация функций BBU [3]. Эта архитектура основана на распределенных антеннах и базовых блоках ($BBUs$), сгруппированных в центральном офисе (CO). BBU включает в себя критически важные функции нижнего уровня, такие как кодирование и декодирование каналов, модуляцию и демодуляцию, а также планирование радиосвязи, управление радиоканалом и сравнение данных. Гибкое перераспределение ресурсов между базовыми станциями сети доступа позволяет лучше адаптироваться к временным и сезонным изменениям трафика.

Другими словами, в традиционной архитектуре RAN функции обработки и радиосвязи в основной полосе частот размещены внутри базовой станции (eNB) в месте сотовой связи, в то время как в $C-RAN$ функицональные возможности базовой станкции отделены от узла сотовой связи и распределены между RRH и пулом BBU, которые расположены далеко друг от друга. Функции обработки в основной полосе частот перемещены и виртуализированы в пул BBU в центральном облаке. RRH содержит антенны малой мощности и выполняет все радиочастотные функции, необходимые для излучения сигнала в ячейке сотовой связи. Они выполняют усиление, аналого-цифровое преобразование радиосигналов и направляют оцифрованные радиосигналы в центральный пул BBU, где принятые сигналы обрабатываются в большом масштабе, а ресуры облака динамически распределяются по требованию.

На рисунке 1 представлена схематическая модель C-RAN, состоящая из базовых станций сети - $eNBs$, расположенных в различных места, на станции присутсвует радимодуль RRH, предназначенный для приема, передачи и управления сигналом. Передача радиосигналов между BBU и RRH осуществляется через фронтхол ($fronthaul$), роль которого пару десятков лет исполнял коаксиальный кабель. Затухание в этом кабеле заметно ухудшало отношение сигнал/шум, надежность и качество связи. Замена коаксикального кабеля оптическим волокном, с одной стороны, привела к переносу функций оцифровки сигналов (АЦП/ЦАП) в радиомодули. С другой стороны, она сняла прежние ограничения на допустимое расстояние между модулями RRU и BBU. В соответсвие, с концепцией $C-RAN$, на рисунке представлена полная централизация BBU в дата центре и виртуальная реализация их фукнций для эффективного распределения нагрузки.

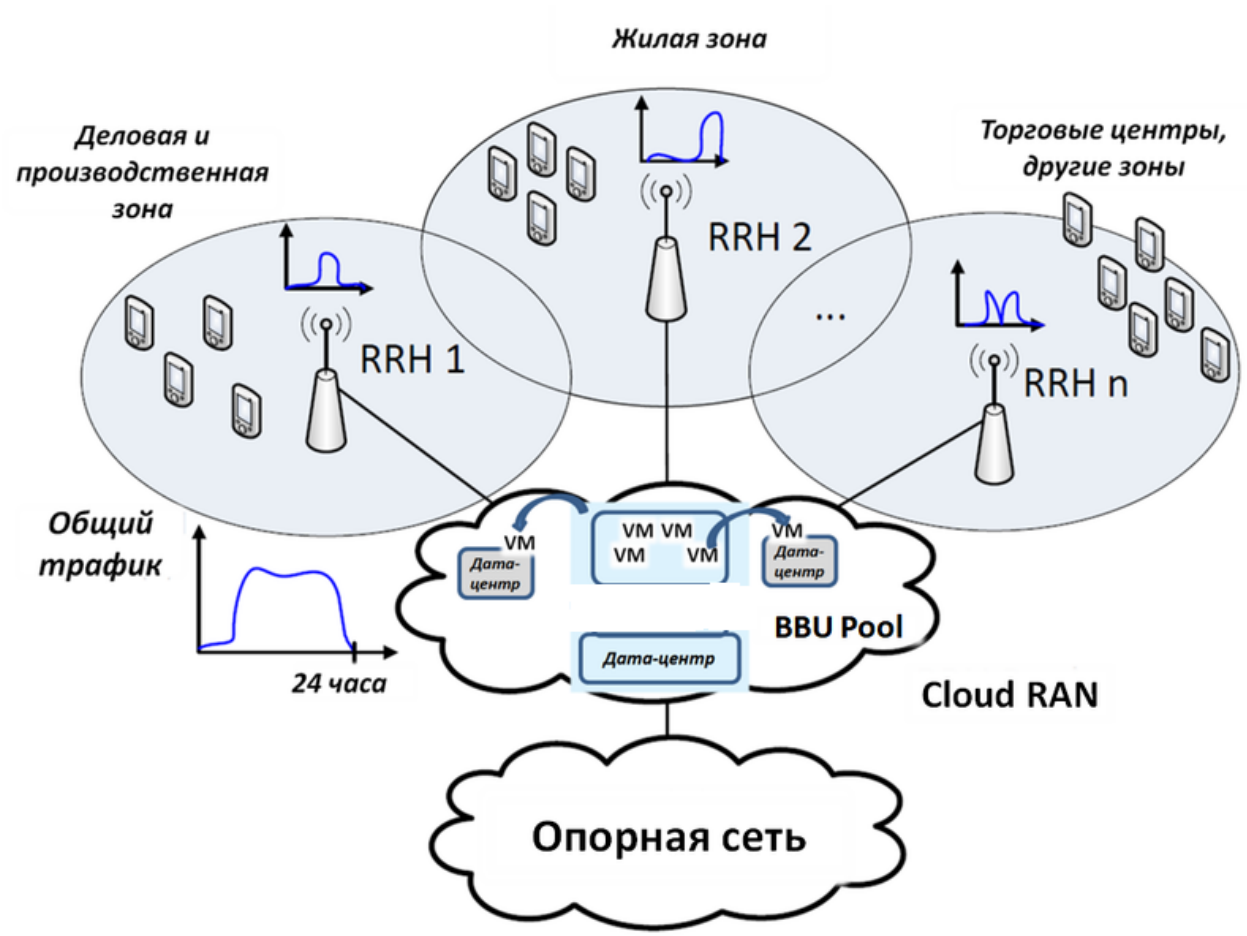

Рисунок 1 – Схематическое представление C-RAN

Например, в дневные часы большинство абонентов находятся на работе в бизнес-центрах, различных организациях и на предприятиях ближе к центру города. В рабочие часы трафик в этой области увеличивается, и там требуется больше ресурсов BBU. В жилых зонах в рабочие часы трафик невысок. Напротив, в вечернее, ночное и утреннее время, концентрация трафика пользователей в жилых зонах больше, чем в центре города.

При традиционном построении сети доступа, для каждой зоны потребуется количество ресурсов BBU из расчёта на максимальный трафик в определённые часы и дни недели. В другое время ресурсы BBU недозагружены, т.к. оборудование физически распредлено по зоне покрытия сети и не имеет возможность перераспределять вычислительные ресурсы. Это приводит к росту стоимости владения сети радиодоступа, поскольку количество BBU рассчитывается из максимальных значений.

Полностью централизованная архитектура может обеспечить многосоставное радиосвязное взаимодействие и управление помехами, что обеспечит

как лучшую эффективность использования спектра частот, так и пользовательский опыт [4, 5]. Однако полная централизация должна соответствовать строгим требованиям к задержке, определенным стандартами Long Term Evolution (LTE) [6]. Фактически, обработка основной полосы частот должна быть завершена в течение 1 миллисекунды в направлении нисходящей линии связи - поток информации направленный в сторону абонента и 2 миллисекунд в направлении восходящей линии связи - поток информации направленный от сотовой вышки на обработку в сеть. Как следствие, высокопроизводительные процессоры и их эффективное использование являются обязательными условиями при построении C-RAN.

Многие усилия уже были направлены на внедрение традиционной сети радиодоступа (RAN) с помощью технологии виртуализации сетевых функций. Некоторые исследования показывают, что функции, принадлежащие физическому уровню, особенно функция кодирования канала, потребляют наибольшее количество времени обработки и вычислительных ресурсов [7]. Таким образом, помимо требования к высокопроизводительным процессорам, можно рассматривать методы параллельного программирования для обеспечения гибкости, применяемой NFV [8, 9]. Более того, одна из основных проблем производительности C-RAN связана с недетерминированным поведением функции кодирования канала, т.е. изменчивостью времени, которое требуется для выполнения процесса кодирования и декодирования. Основная причина нестабильности обусловлена характеристиками радиоканала каждого пользовательского оборудования (UE), подключенного к eNB - базовая станция сети, требуемой скоростью передачи данных на UE, а также объемом трафика в ячейке.

1.1 Виртуальные функции

VNF в C-RAN — это не что иное, как виртуализированный BBU (vBBU), который программно реализует все сетевые функции, принадлежащие трем нижним уровням стека протоколов LTE. Эти функции в основном касаются таких функций PHY, как генерация сигнала, IFFT /FFT, модуляция и демодуляция, кодирование и декодирование, планирование радиосвязи, объединение /сегментация протокола управления радиоканалом (RLC) и процедуры шифрования/дешифрования протокола эволюции пакетных данных (PDCP) для нисходящей и восходящей линий связи [10]. Задача состоит в том, чтобы выполнять виртуальные функции BBU достаточно быстро, чтобы увеличить расстояние

между RRHs и функциями BBU (а именно, пулом BBU) и, таким образом, повысить уровень концентрации BBU в дата-центре для экономии капитальных и операционных затрат.

Как показано на рисунке 2, vBBU может быть создан в верхней части облачной инфраструктуры и смоделирован с помощью графа пересылки подфункций, которые, в свою очередь, могут быть разделены на параллельно выполняемые задания. Задания BBU могут выполняться на многоядерной платформе в соответствии со стратегией планирования, которая находится в ядре операционной системы (OS) хоста. В [11] предлагается архитектура C-RAN, в которой используются преимущества производительности, обеспечиваемые контейнерами, которые, в отличие от виртуальных машин, содержат единую ОС, т.е. все ядра управляются глобальным планировщиком.

Когда пользовательское оборудование (UE) требует либо передачи, либо приема данных, вызывается экземпляр vBBU. Как следствие, различные экземпляры виртуального BBU выполняются одновременно на вычислительной платформе. В сотовых системах на основе LTE блок данных передачи (а именно, подкадр) может состоять из данных различных UE. Вся обработка подкадра в основной полосе частот должна выполняться в течение 2 миллисекунд и 1 миллисекунды в направлении восходящей линии связи и нисходящей линии связи соответственно. Поскольку субкадр генерируется каждую миллисекунду в обоих направлениях (рисунок 3) [12], обработка в BBU всех ячеек, принадлежащих облачной системе, скорее всего, требует высокопроизводительных параллельных вычислений.

При выполнении функций BBU в многоядерной системе могут быть предусмотрены различные стратегии планирования, такие как планирование для каждого подкадра LTE [13] или с более высокой степенью детализации, например, для UE или для блока кода (CB) [8, 9]. Стратегия планирования должна повысить производительность пула BBU с точки зрения задержки. Как отмечено в [8,9], прирост производительности наиболее заметен, когда планировщик имеет дело с короткими заданиями (например, обработка CBs), вместо того, чтобы выделять вычислительные ресурсы для тяжелых задач (например, обработка подкадров) [12].

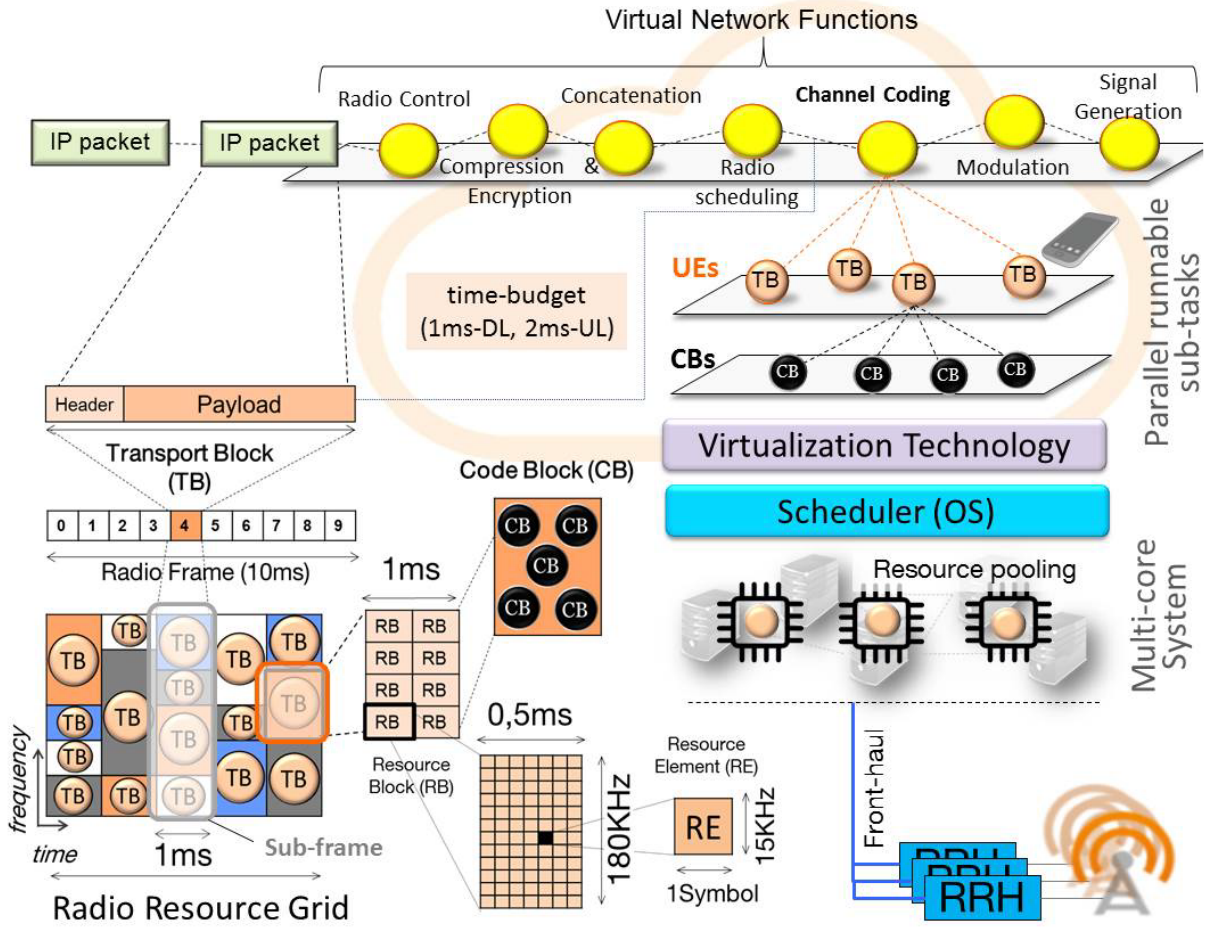

Рисунок 2 – Архитектура C-RAN

1.2 Параллельная обработка

Общая философия параллельных вычислений состоит в разделении больших задач на более мелкие подзадачи с возможностью параллельного выполнения. Параллельное выполнение подзадач позволяет сократить время выполнения всей задачи. В облачной системе обработка функции кодирования канала является наиболее ресурсоемкой и, кроме того, имеет недетерминированное поведение [7]. Далее мы рассмотрим параллельную обработку функции кодирования канала, которая может выполняться либо для каждого UEs (т.е. для транспортных блоков (TBs)), либо для блока кода (CBs) параллельно.

Для конкретики, мы используем тот факт, что в LTE, когда размер транспортного блока (TB) слишком велик, перед обработкой он сегментируется, с помощью подфункции кодирования / декодирования, на более короткие блоки данных, называемые CBs. CB представляет собой наименьший блок обработки, который может выполняться параллельно.

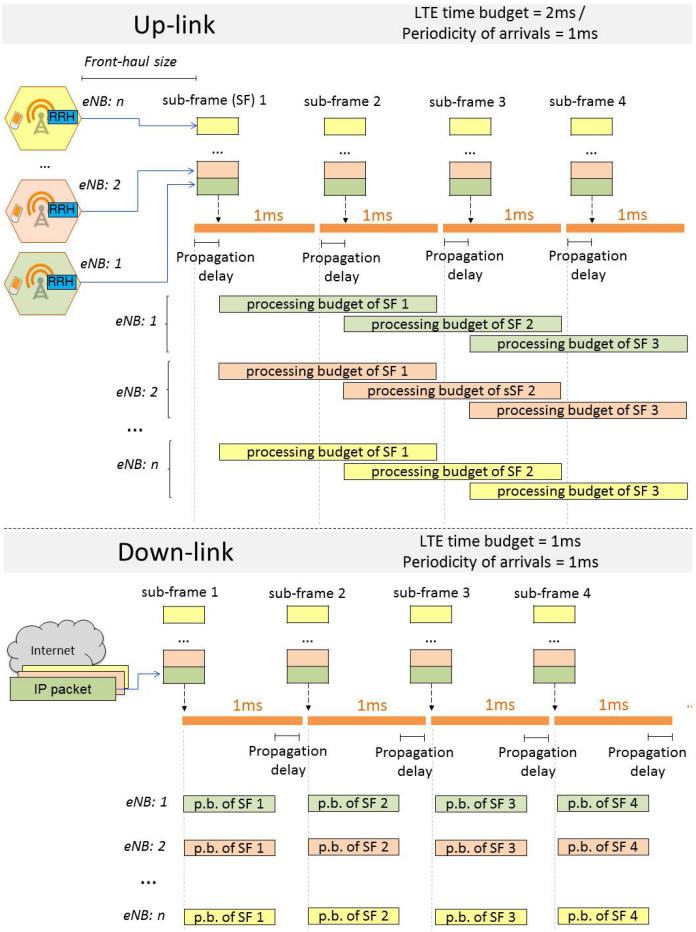

Рисунок 3 – Процесс обработки подкадров

Стоит отметить, что передача радиосигнала между UE и eNB генерирует TB каждую миллисекунду [14]. Время обработки TB зависит от условий радиоканала, загрузки данных на UE и объема трафика в ячейке. Планировщик радиосвязи выделяет количество блоков физических ресурсов (RBs) для каждого UE в зависимости от объема трафика в соте и определяет схему модуляции и кодирования (MCS) на основе качества радиоканала. Как количество блоков ресурсов (NRB), так и MCS определяют размер транспортного блока (TBs), т.е. полезные данные [15].

В работе [3] используется параллельная обработка в строгом смысле этого слова, чтобы задания выполнялись одновременно на отдельных ядрах, таким образом избегая процессоров с разделением времени. При параллельных вычислениях каждое задание выполняется на одном ядре и только на одном в любой момент. И наоборот, параллельные вычисления позволяют одновременное выполнение заданий на одном ядре за счет перекрытия периодов времени; это приводит к моделям PS с общим использованием процессоров [16]. Одна-

ко недостаток совместного использования процессора заключается в том, что многозадачность на одном ядре требует переключения контекста и разделения памяти, что может заметно увеличить задержку.

При выполнении параллельной обработки для каждого UE количество заданий параллельного кодирования определяется количеством UEs, запланированного в подкадре LTE. При применении параллельной обработки для каждого CBs количество заданий параллельного кодирования определяется произведением количества запланированных UEs и количества CBs для каждого из них. Количество CBs на UE задается большим целым числом TBS, деленным на размер используемого блока кода (CBS), т.е. $N_{CB} = [TBS/(6144 - 24)]$ [8], где мы используем тот факт, что LTE определяет минимальный и максимальный размер блока кода, равный 40 и 6144 бита соответственно. Последние 24 бита каждого CB соответствуют циклической проверке избыточности (CRC) [6, 12].

2 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 [17–20] авторы сосредоточены на оптимизации некоторых параметров $C-RAN$.

В работе [18] 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 4), весь входящий трафик UEs сначала обрабатывается диспетчером. Предполагается, что интенсивность

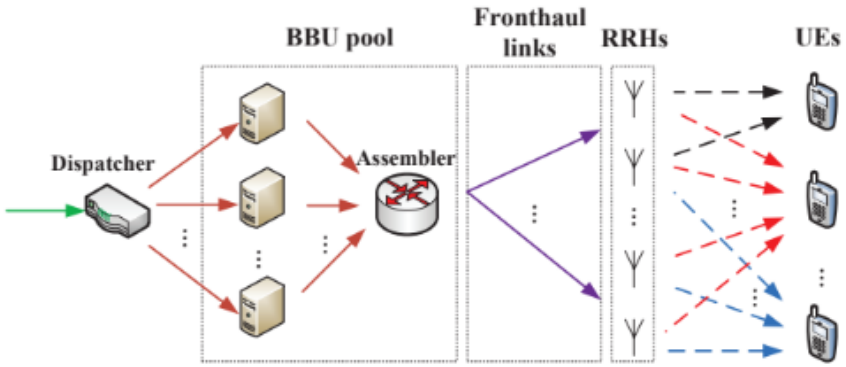

Рисунок 4 – Схема $C-RAN$

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 5. В частности, в пуле BBU транс-

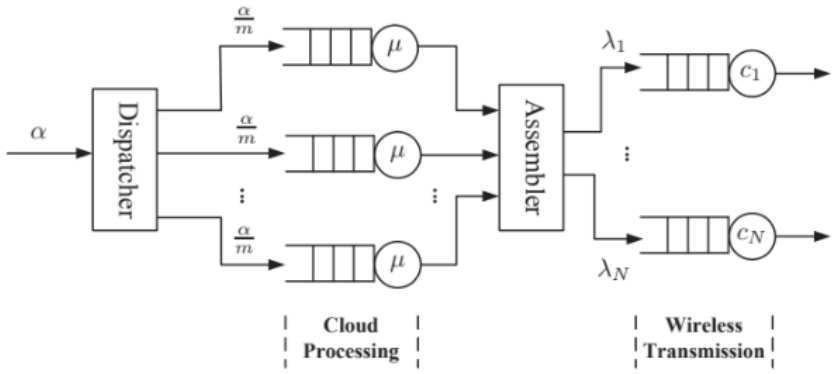

Рисунок 5 – Модель сети массового обслуживания, представляющая облачную обработку $C-RAN$ и беспроводную передачу данных

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:

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

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

Дальнейшая 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 [19] 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 [17] 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. Кро-

ме 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 $VBSs$, энергопотребление сети разделяется на две части: (i) энергопотребление RRH и транс-

портной 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 [17] 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 [2, 3, 8, 9, 12, 21–23] посвещаны решению задач анализа моделей C-RAN.

В работе [22] 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 [22] 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 центров 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 - пря-

мая 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 [23] 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 [3] 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 [8,9]. 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 [10], 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 [24]. 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:

$$\rho = \frac{\lambda}{\mu C} < 1. \tag{2.1}$$

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 экспо-

ненциально 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 [13]. 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].

ЗАКЛЮЧЕНИЕ

В данной работе рассмотрены основные приципы функционирования облачной сети радиодоступа. 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 Network Functions Virtualisation (NFV); Architectural Framework [Электронный ресурс]. — 2013. — URL: https://www.etsi.org/deliver/etsi_gs/nfv/001_099/002/01.01.01_60/gs_nfv002v010101p.pdf (дата обращения 09.01.2023). Загл. с экр. Яз. англ.

2 *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.

3 *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.

4 Cloud-ran architecture for 5g [Электронный ресурс]. — 2019. — URL: http://www.hit.bme.hu/~jakab/edu/litr/5G/WhitePaper_C-RAN_for_5G_-Telefonica_Ericsson.PDF (дата обращения 09.01.2023). Загл. с экр. Яз. англ.

5 Cloud-ran [Электронный ресурс]. — 2015. — URL: http://www.hit.bme.hu/~jakab/edu/litr/5G/Ericsson_wp-cloud-ran.pdf (дата обращения 09.01.2023). Загл. с экр. Яз. англ.

6 LTE; Evolved Universal Terrestrial Radio Access (E-UTRA); Physical layer procedures [Электронный ресурс]. — 2017. — URL: https://www.etsi.org/deliver/etsi_ts/136200_136299/136213/14.02.00_60/ts_136213v140200p.pdf (дата обращения 09.01.2023). Загл. с экр. Яз. англ.

7 *Nikaein, N.* Processing Radio Access Network Functions in the Cloud / N. Nikaein // Proceedings of the 6th International Workshop on Mobile Cloud Computing and Services. — Paris: ACM, 2015. — Pp. 36–43.

8 *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.

9 *Rodriguez, V. Q.* Towards the deployment of a fully centralized Cloud-RAN architecture / V. Q. Rodriguez, F. Guillemin // 13th International Wireless Com-

munications and Mobile Computing Conference (IWCMC). — Valencia: IEEE, 2017. — Pp. 1055–1060.

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

11 *Barik, R. K.* Performance analysis of virtual machines and containers in cloud computing / R. K. Barik, R. K. Lenka, K. R. Rao, D. Ghose // International Conference on Computing, Communication and Automation (ICCCA). — Greater Noida: IEEE, 2016. — Pp. 1204–1210.

12 *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.

13 *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.

14 *Ашурметова, Н. З.* Исследование архитектуры облачной системы C-RAN / Н. З. Ашурметова // *Наукосфера*. — 2021. — Vol. 7, no. 1. — Pp. 125–133.

15 Cloud-RAN, the next-generation mobile network architecture [Электронный ресурс]. — 2017. — URL: `https://www-file.huawei.com/-/media/corporate/pdf/mbb/cloud-ran-the-next-generation-mobile-network-architecture.pdf?la=en` (дата обращения 09.01.2023). Загл. с экр. Яз. англ.

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

17 *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.

18 *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.

19 *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.

20 *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.

21 *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.

22 *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.

23 *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.

24 *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.