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

Кафедра системного анализа и
автоматического управления

ОПТИМАЛЬНОЕ РАЗМЕЩЕНИЕ СТАНЦИЙ В БЕСПРОВОДНОЙ СЕТИ СВЯЗИ

АВТОРЕФЕРАТ МАГИСТЕРСКОЙ РАБОТЫ

Студента 2 курса 271 группы
направления 09.04.01 — Информатика и вычислительная техника
факультета КНиИТ
Агапова Максима Леонидовича

Научный руководитель
доцент к. ф.-м. н. _____ О. А. Осипов

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

Саратов 2021

ВВЕДЕНИЕ

Актуальность темы. Мобильные приложения становятся все более ресурсоемкими, в то время как вычислительная емкость мобильных устройств ограничена из-за их портативных размеров. Эффективным подходом к повышению производительности мобильных приложений является сброс некоторых из своих задач в удаленное облако, где приложение состоит из нескольких задач. Существующие исследования по разгрузке мобильных задач в основном рассматривали облако как удаленное место разгрузки, в связи с его обилием вычислительных ресурсов. Однако, удаленное облако обычно расположено далеко от своих пользователей, и задержка сети, связанная с передачей данных между пользователями и облаком, может быть очень дорогостоящей. Это особенно нежелательно в приложениях с дополненной реальностью и мобильных многопользовательских игровых системах, где время отклика имеет решающее значение для удобства работы пользователя.

В ходе последних исследований было предложено использовать кластеры компьютеров, называемые cloudlets (локальные облака), в качестве дополнения к удаленному облаку для разгрузки [1]. Локальные облака обычно размещаются в точке доступа в сети и могут быть доступны для пользователей через беспроводное соединение. Ключевым преимуществом локального облака перед удаленным является то, что непосредственная физическая близость между локальными облаками и пользователями обеспечивает более короткое время задержки связи, тем самым улучшая пользовательский опыт работы с интерактивными приложениями [2].

Цель данной работы — описание задачи размещения локального облака в беспроводной сети связи, изучение эффективности алгоритмов распределения локальных облаков по точкам доступа, а также их сравнение.

В соответствии с поставленной целью определены **следующие задачи**:

1. Ознакомиться с основными понятиями и определениями, связанными с беспроводными сетями;
2. Изучить понятие сетей массового обслуживания, их параметры и характеристики;
3. Изучить используемую модель системы WMAN, сформулировать основную проблему;

4. Изучить алгоритмы распределения локальных облаков по точкам доступа;
5. Реализовать программу для анализа сети WMAN;
6. Исследовать эффективность работы алгоритмов, а также их производительность в зависимости от изменения параметров сети массового обслуживания.

Методологические основы оптимального размещения станций в беспроводной сети связи представлены в работах M. Satyanarayanan [1], S. Clinch [2], B.-G. Chun [3], H. Hong [4], S. Kosta [5], Y. Zhang [6].

Практическая значимость магистерской работы. В ходе выполнения выпускной квалификационной работы была разработана программу для анализа сети WMAN, с помощью которой можно исследовать эффективность работы алгоритмов, а также их производительность в зависимости от изменения параметров сети массового обслуживания.

Структура и объем работы. Магистерская работа состоит из введения, 6 разделов, заключения, списка использованных источников и одного приложения. Общий объем работы — 65 страниц, из них 53 страницы — основное содержание, включая 12 рисунков и 1 таблицу, список использованных источников информации — 20 наименований.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Первый раздел «Виды беспроводных сетей» посвящен описанию основных понятий и определений, связанных с беспроводными сетями на основе имеющихся исследований.

В подразделе 1.1 приведено описание локального облака и его отличия от удаленного облака. Также затрагивается проблема размещения локального облака и распределения пользователей по ним в WMAN.

Подраздел 1.2 посвящен возможности использования локального облака для разгрузки пользовательских задач. Для разгрузки задач мобильные пользователи инкапсулируют задачи в виртуальную машину, которая затем загружается в локальное облако для выполнения. Затем пользователь может выгрузить дополнительные задачи, используя операции выгрузки на своей виртуальной машине в локальное облако. Как только задача на виртуальной машине в локальном облаке выполнена, результат возвращается пользователю.

Второй раздел «Необходимые сведения из теории массового обслуживания» посвящен описанию необходимых сведений из теории массового обслуживания. В частности описаны M/M/1 и M/M/c системы.

В подразделе 2.1 описываются основные обозначения из теории массового обслуживания. Теория массового обслуживания включает три раздела: элементарную теорию массового обслуживания, промежуточную теорию и общую теорию. Для обозначения различных типов СМО, которые рассматриваются здесь, применяются довольно простые сокращения. Они содержат три позиции и имеют вид A/B/c; таким образом обозначается СМО с c обслуживающими приборами, а A и B указывают соответственно на распределение времени между соседними требованиями и распределение времени обслуживания. A и B принимают значения из следующего набора символов, которые указывают соответствующее распределение:

M - *показательное распределение*

E_r - *распределение Эрланга порядка* r

H_R - *гиперпоказательное распределение порядка* R

D - *постоянная величина*

G - *произвольное распределение*

Подраздел 2.2 посвящен общим результатам.

В подразделе 2.3 описываются марковский и пуассоновский процессы, а также процессы размножения и гибели.

Подраздел 2.4 посвящен M/M/1 системе. Система M/M/1 является простейшей. Она представляет собой классический пример, для рассмотрения которого требуется лишь элементарный математический аппарат. На вход такой системы поступает пуассоновский поток (с интенсивностью λ) и система совершает единичные (ординарные) переходы (обслуживание и поступление одного требования).

Подраздел 2.5 посвящен M/M/c системе.Рассмотрим обобщение на случай c обслуживающих приборов. Перед совокупностью c обслуживающих приборов образуется одна очередь, и требование из головы очереди поступает в первый доступный прибор. Как обычно, λ— интенсивность входящего потока, а $1/\mu$ — среднее время обслуживания, причем $\rho = \lambda/c\mu$. Стационарная вероятность застать в системе k требований определяется равенствами

$$p_k = \begin{cases} p_0 \frac{(c\rho)^k}{k!}, & k \le c, \\ p_0 \frac{\rho^k c^c}{c!}, & k \ge c, \end{cases} \tag{1}$$

где

$$p_0 = \left[\sum_{k=0}^{c-1} \frac{(c\rho)^k}{k!} + \frac{(c\rho)^c}{c!(1-\rho)} \right]^{-1}. \tag{2}$$

В начале текущего столетия эту СМО рассматривал в качестве одной из моделей работы телефонной сети основатель теории массового обслуживания А. К. Эрланг. Его именем названа C-формула Эрланга, которая определяет вероятность того, что поступившее требование должно ждать обслуживания. Эта формула определяет $\sum_{k=c}^{\infty} p_k$ из равенства (1).

Эрланг рассматривал модель телефонной системы, которая во всем совпадает с СМО типа M/M/c, но не допускает возможности ожидания; иначе говоря, это система с потерями, в которой в любой момент времени могут находиться не более c требований. В этом случае вероятность застать k требований в системе дается равенством

$$p_k = \frac{(\lambda/\mu)^k/k!}{\sum_{i=0}^{c} (\lambda/\mu)^i/i!}. \tag{3}$$

справедливым для $0 \le k \le c$. Важной величиной здесь является вероятность того, что требование в момент поступления в систему не застает ни одного свободного прибора и поэтому будет потеряно; она задается B-формулой Эрланга или формулой потерь Эрланга, которая определяет p_c из равенства (3).

Третий раздел «Модель системы WMAN» посвящен описанию модели системы WMAN, представленной набором точек доступа $P = \{p_1, ...p_m\}$, соединенных между собой интернетом, и набором пользователей мобильной связи $U = \{u_1, ..., u_n\}$, которые могут получить доступ к сети через точки доступа. Воспользуемся неориентированным графом $G(V, E)$ для представления соединений между пользователями и точками доступа в WMAN, где $V = P \cup U$. Существует два типа ребер в G. Ребро между пользователем u_i и точкой доступа $p_k(u_i, p_k) \in E$ указывает на то, что u_i соединен с точкой p_k беспроводным способом (изображено на рис. 1 пунктиром). Ребро между

двумя точками доступа p_i и p_k $(p_i, p_k) \in E$ указывает на то, что они имеют расстояние в один шаг (изображено на рис. 1 сплошными линиями).

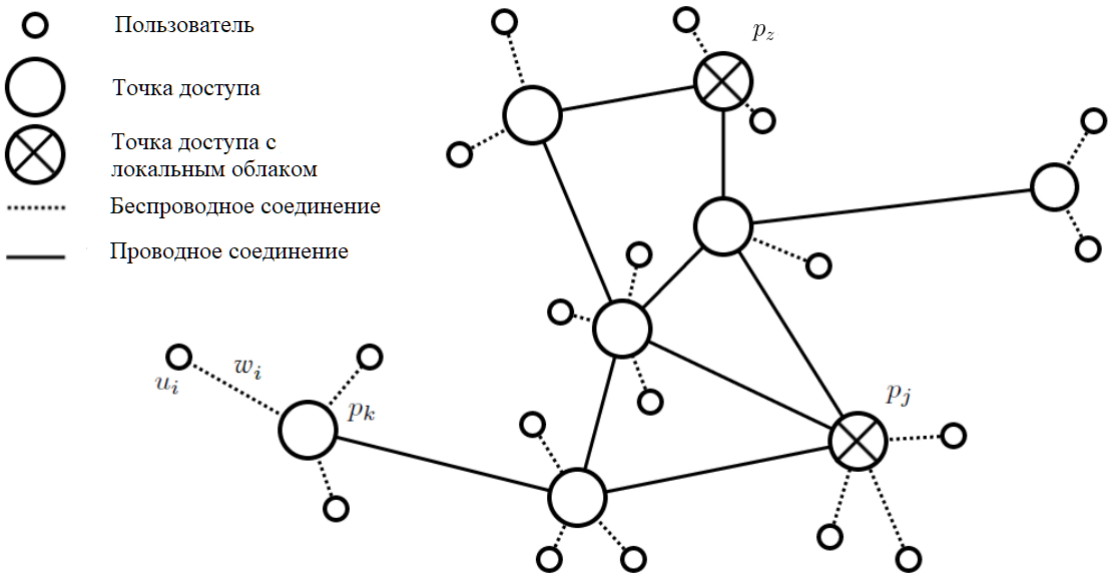

Рисунок 1 - Пример WMAN.

Предположим, что граф G связанный, что подразумевает, что все точки доступа доступны друг другу через высокоскоростное подключение к интернету. Кроме того, существует удаленное облако, доступ к которому можно получить с каждой точки доступа через интернет. Задачи каждого мобильного пользователя могут непредсказуемо меняться, особенно когда пользователь запускает несколько приложений одновременно.

Каждый пользователь u_i генерирует поток задач с некоторой интенсивностью λ_i. Предполагается, что задачи образуют пуассоновский поток. Часть задач, которая пересылается для обработки, продолжит выполняться на устройстве. Данный тип задач не учитывается в рассматриваемой модели.

Чтобы выгрузить свои задачи в локальные облака для обработки, пользователю необходимо передать свои задачи через сеть G. Возьмем в качестве примера рис. 2, пусть w_i будет обозначать задержку беспроводной связи между u_i и p_k, к которой u_i подключен. Если выгруженная задача u_i будет обрабатываться в локальном облаке, расположенном по адресу p_j, то задача должна быть перенаправлена с p_k на p_j.

Предполагается, что все задачи пользователей имеют единый размер пакета данных, поэтому задержки при передаче любой задачи по сети между парой точек доступа идентичны. Для моделирования такой сетевой задерж-

ки в WMAN следует обозначить $D \in \mathbb{R}^{m \times m}$ матрицей сетевой задержки, где $D_{k,j}$ представляет собой задержку связи при перенаправлении задачи между p_k и p_j.

В подразделе 3.1 описывается модель распределения многопользовательских мобильных задач. В качестве модели будет использоваться сеть массового обслуживания. Предполагается, что в G есть K-локальных облаков, которые должны быть размещены. Задачи пользователей могут быть выполнены как на одном из K-локальных облаков, так и на удаленном облаке. Первоначально задачи пользователя направляются на некоторое локальное облако, если при этом оно перегружено, то часть задач будет перенаправлена на удаленное облако. Локальные облака моделируются как $M/M/c$ системы обслуживания, где каждое локальное облако состоит из c однородных серверов с интенсивностью обслуживания μ.

Подраздел 3.2 посвящен формулировке проблемы размещения локальных облаков по точкам доступа. Проблема размещения K локальных облаков в WMAN G определяется следующим образом. Учитывая целое число $K \geq 1$ и параметры системной модели $(G, \Lambda, W, D, T_{net}, \lambda_{max}, B, \mu, c)$, задача состоит в том, чтобы найти X (размещение локальных облаков среди точек доступа) и Y (назначение пользователей на локальные облака) таким образом, чтобы время отклика системы t было минимальным.

Четвертый раздел «Размещение облака для минимизации времени отклика» посвящен описанию алгоритмов для решения проблемы размещения K локальных облаков.

Подраздел 4.1 посвящен алгоритму распределения локальных облаков по точкам доступа на основе пользовательских нагрузок. Для каждого пользователя, подключенного к точке доступа p_k, необходимо найти локальное облако, к которому p_k имеет наименьшую сетевую задержку $D_{k,j}$, и назначить u_i для этого локального облака. Это минимизирует сетевую задержку между пользователем и его локальным облаком.

Подраздел 4.2 посвящен алгоритму распределения локальных облаков по плотности пользователей. Для размещения локальных облаков в точках доступа сперва выбирается точка доступа p_j с наибольшей суммарной нагрузкой от всех кандидатов $\lambda_{T_{net}}(j)$, после чего размещается локальное облако в точке доступа p_j. Затем удаляется набор пользователей, непосредственно

подключенных к p_j из сети G, производится перерасчет кандидатов в каждой точке доступа в обновленной сети, и находится следующая точка доступа с наибольшей нагрузкой кандидата. Этот процесс повторяется K раз до тех пор, пока все K локальные облака не будут размещены.

Подраздел 4.3 посвящен описанию метода распределения локальных облаков по точкам доступа для динамических сетей WMAN. **Пятый раздел «Описание программы для анализа сети WMAN»** посвящен описанию программы, реализованной на языке программирования *python* для анализа сети WMAN.

Шестой раздел «Результаты исследования эффективности работы алгоритмов» посвящен описанию полученных результатов исследований эффективности работы алгоритмов, а также приведен пример применения используемой модели сети WMAN на практике.

Подраздел 6.1 посвящен использованию карты метрополитена Гонконга в качестве шаблона для сети WMAN, где проводные соединения между каждым районным узлом используются для представления границ проводных соединений между узлами WMAN.

Подраздел 6.2 посвящен описанию случайно сгенерированной сети для тестирования и оценки предложенных алгоритмов. При моделировании сети применяется библиотека *networkx*, предназначенная для создания, манипуляции и изучения структуры, динамики и функционирования сложных сетевых структур.

Подраздел 6.3 посвящен исследованию эффективности алгоритмов размещения локальных облаков. Рассмотрим график зависимости времени отклика системы от увеличения количества локальных облаков в сети. Как видно на рис. 2, алгоритм DBC уменьшает свое время отклика системы с $K = 1$ до $K = 15$, пока, в конечном счете, не выровняется. При $K = 1$ большая часть поступающих задач направляется дальше в удаленное облако. При $K = 5$ время отклика системы уменьшилось на 29% и продолжает уменьшаться вплоть до $K = 10$. Однако после $K = 10$ в сеть попадает такое количество локальных облаков, что каждый из них способен выполнить все принимаемые задачи. Поскольку пользователи могут передавать все свои задачи в назначенные им локальные облака, время отклика системы ограничивается снизу беспроводной связью и сетевыми задержками между пользо-

вателями и локальными облаками, что приводит к выравниванию времени отклика системы. Это говорит о том, что после размещения определенного количества локальных облаков, размещение дополнительных локальных облаков будет соответствовать закону убывающей доходности с точки зрения времени отклика системы.

В свою очередь, алгоритм HAF обладает недостаточной производительностью со средним значением времени отклика системы на 10% больше, чем у алгоритма DBC. А при K в диапазоне $10 \leq K \leq 15$ разница в производительности около 41%.

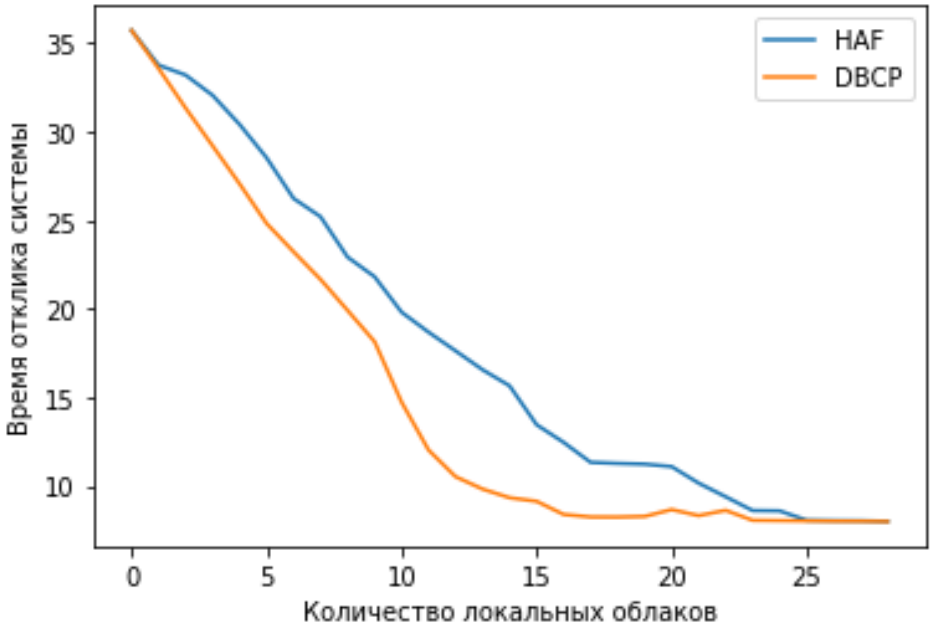

Рисунок 2 - Время отклика системы в зависимости от алгоритмов размещения локальных облаков.

Для дальнейшей оценки эффективности алгоритмов размещения облаков было изучено влияние параметров системы на производительность предложенных алгоритмов.

ЗАКЛЮЧЕНИЕ

В результате выполнения данной выпускной квалификационной работы разработана программа для анализа сети WMAN, реализованы алгоритмы размещения локальных облаков, а также проведен анализ эффективности имеющихся алгоритмов.

Все цели, поставленные в ходе магистерской работы были выполнены.

Полученные в работе результаты могут быть использованы для создания усовершенствованной модели системы, которая будет способна отслеживать и прогнозировать перемещение пользователей в сети.

В ходе выполнения данной работы было принято участие в студенческой научной конференции факультета КНиИТ со статьей «Задача размещения станций в беспроводной сети связи».

Основные источники информации:

1. Satyanarayanan, M. The case for vm-based cloudlets in mobile computing [Электронный ресурс] / M. Satyanarayanan, P. Bahl, R. Caceres, N. Davies. URL: https://ieeexplore.ieee.org/document/5280678/authorsauthors (дата обращения 22.12.2019). Загл. с экрана. - Яз. англ.
2. Clinch, S. How close is close enough understanding the role of cloudletsin supporting display appropriation by mobile users [Электронныйресурс] / S. Clinch, J. Harkes, A. Friday, N. Davies, M. Satyanarayanan. URL: https://ieeexplore.ieee.org/document/6199858 (дата обращения 22.12.2019). Загл. с экрана. - Яз. англ.
3. Chun, B.-G. Clonecloud: Elastic execution between mobile device and cloud // Proceedings of the Sixth Conference on Computer Systems, Salzburg, Austria, April, 2011. ACM, 2011. P. 301–314. https://doi.org/10.1145/1966445.1966473.
4. Hong, H. Placing virtual machines to optimize cloud gaming experi-ence [Электронный ресурс] / H. Hong, D. Chen, C. Huang, K. Chen, C. Hsu. URL: https://www.semanticscholar.org/paper/Placing-Virtual-Machines-to-Optimize-Cloud-Gaming-Hong-51 Chen/ba8324733337de404afdf3e5def4f645823cd3b4 (дата обращения 13.03.2020). Загл. с экрана. - Яз. англ.
5. Kosta, S. Thinkair: Dynamic resource allocation and parallel execution in thecloud for mobile code offloading // Proceedings IEEE INFOCOM, Orlando, FL, USA, March, 2012. IEEE, 2012. Pp. 945–953.
6. Zhang, Y. To offload or not to offload: an efficient code partition algorithm for mobile cloud computing [Электронный ресурс] / Y. Zhang, H. Liu, L. Jiao, X. Fu. URL: http://citeseerx.ist.psu.edu/viewdoc/ download?doi=10.1.1.568.8803rep=rep1type=pdf (дата обращения 22.12.2019). Загл. с экрана. - Яз. англ.