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

Кафедра системного анализа и
автоматического управления

МЕТОДЫ ОПТИМАЛЬНОГО УПРАВЛЕНИЯ В SPLIT-MERGE СИСТЕМАХ МАССОВОГО ОБСЛУЖИВАНИЯ

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студентки 4 курса 481 группы
направления 27.03.03 — Системный анализ и управление
факультета КНиИТ
Гурковой Виктории Марковны

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

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

Саратов 2021

ВВЕДЕНИЕ

Актуальность темы. Системы массового обслуживания (СМО) с делением и слиянием требований (*Fork-Join Queueing Systems*) [1-5] представляют собой модели реальных систем с параллельной обработкой (многопроцессорные системы, GRID-системы, MapReduce и т.д.). Ключевой особенностью таких систем является деление поступающих требований на части — фрагменты, которые обслуживаются параллельно на приборах системы обслуживания, и последующее объединение обслуженных фрагментов в исходные требования.

Системы массового обслуживания с делением и слиянием требований также являются математическими моделями, используемыми для анализа дискретных стохастических систем с параллельным и распределенным принципами функционирования.

В зависимости от возможных вариантов деления требований при поступлении и слияния фрагментов после завершения обслуживания выделяют три основных класса систем обслуживания с делением и слиянием требований [2]:

— с центральным делением без синхронизирующей очереди (centralized splitting model without synchronization queue, split-merge model),
— с центральным делением и синхронизирующей очередью (centralized splitting model with synchronization queue),
— с распределённым делением и синхронизирующей очередью (distributed splitting model with synchronization queue).

Структурная и функциональная специфика таких систем требует разработки новых эффективных моделей и методов для решения задач анализа, синтеза и оптимизации. Для описания и анализа таких систем используются разнообразные математические абстракции: модель акторов, сети Петри, потоки работ, модели теории массового обслуживания [6, 7].

В большинстве работ [8, 9], посвященных сетям обслуживания с делением и слиянием требований рассматриваются сети, состоящие из множества параллельных систем массового обслуживания, а основным результатом является определение длительности пребывания требований в сети обслуживания. Также в настоящее время в литературе изложены результаты для одного класса требований [2] в случае, когда очередь имеет неограниченную вмести-

мость.

Таким образом, актуальной является задача, связанная с построением математических моделей систем с произвольным числом классов поступающих требований и разработкой методов анализа и оптимизации данных систем обслуживания.

Цель бакалаврской работы — описание и исследование системы массового обслуживания с делением и слиянием требований двух классов типа *split-merge*.

Поставленная цель определила **следующие задачи**:

1. Изучить основные результаты исследования систем массового обслуживания с делением и слиянием требований;
2. Описать систему массового обслуживания типа *split-merge* с делением и слиянием требований без управления;
3. Описать систему массового обслуживания типа *split-merge* с делением и слиянием требований с управлением;
4. Разработать комплекс программ имитационного моделирования и численного анализа системы обслуживания;
5. Изучить и применить следующие методы оптимизации: метод полного перебора, итерационный метод и метод машинного обучения «Q- обучение».

Методологические основы исследования систем массового обслуживания с делением и слиянием требований представлены в работах A. Thomasian, Y. Narahari, P. Sundarrajan, R. Nelson, A. N. Tantawi, A. Rizk, F. Baccelli.

Теоретическая значимость бакалаврской работы. Рассмотренная модель системы массового обслуживания расширяет круг задач, решаемых в теории массового обслуживания, поскольку позволяют рассмотреть особенности структуры и функционирования систем обслуживания с делением и слиянием требований нескольких классов, связанные с усложнением способов маршрутизации требований разных классов в системе, а также наличием зависимостей между фрагментами одного требования. Найдено стационарное распределение вероятностей состояний исследуемой системы, а также получены формулы для её основных стационарных характеристик.

Практическая значимость бакалаврской работы. Представленные в работе результаты могут быть применены для математического модели-

рования стохастических систем с параллельным и распределённым принципами функционирования, для решения задач анализа, оптимизации и синтеза указанных систем (задачи планирования и распределения ресурсов, исследование алгоритмов балансировки в современных распределённых системах и т.д.).

Структура и объем работы. Бакалаврская работа состоит из введения, 5 разделов, заключения, списка использованных источников и цифрового носителя в качестве приложения. Общий объем работы — 70 страниц, из них 62 страницы — основное содержание, включая 19 рисунков и 6 таблиц, список использованных источников информации — 47 наименований.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Первый раздел «Обзор основных результатов исследования систем массового обслуживания с делением и слиянием требований» посвящен обзору основных опубликованных результатов исследований систем массового обслуживания с делением и слиянием требований.

В подразделе 1.1 приведено описание результатов работ, посвящённых сетям массового обслуживания, состоящим из параллельных систем обслуживания.

Подраздел 1.2 посвящен обзору результатов исследования сетей обслуживания с произвольной топологией.

В подразделе 1.3 рассматриваются работы, в которых описаны примеры использования систем массового обслуживания с делением и слиянием требований для моделирования реальных систем.

Второй раздел «Система массового обслуживания с делением и слиянием требований без управления» посвящен описанию системы массового обслуживания с делением и слиянием требований двух классов типа *split-merge* без управления.

Подраздел 2.1 посвящен математическому описанию системы массового обслуживания с делением и слиянием требований двух классов типа *split-merge*. Рассматривается система массового обслуживания типа *split-merge*, состоящая из M обслуживающих приборов, с двумя очередями ограниченной вместимости с дисциплиной обслуживания FCFS (Рисунок 1). В систему обслуживания поступают два класса требований, время между интервалами

поступления которых распределено по экспоненциальному закону с параметрами λ_1 и λ_2 соответственно. Требования первого класса поступают в первую очередь, требования второго класса — во вторую очередь соответственно.

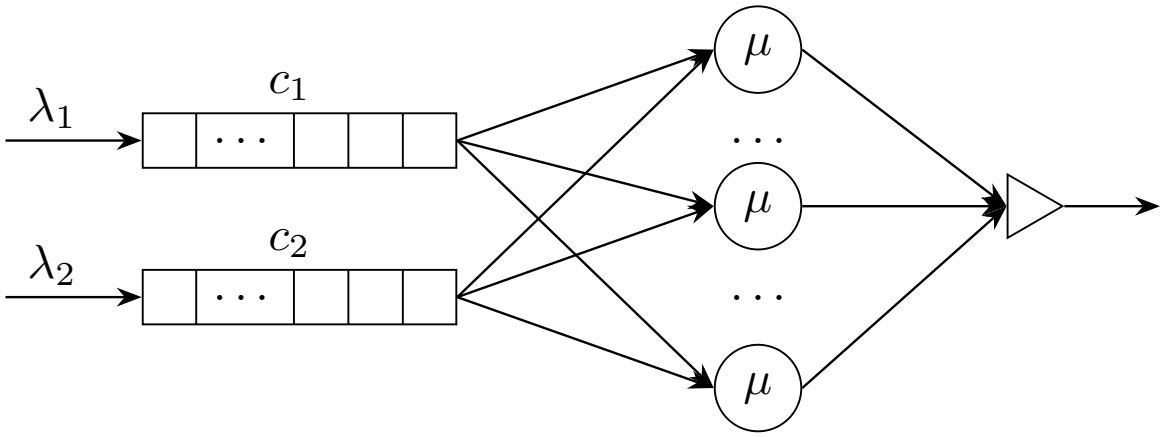

Рисунок 1 – Split-merge система массового обслуживания с двумя классами требований

Требование, поступающее на обслуживание состоит из заданного числа фрагментов. Пусть d_1 — количество фрагментов в первом классе требований, а d_2 — количество фрагментов во втором классе требований. В том случае, когда число свободных приборов достаточно для обслуживания всех фрагментов требования, фрагменты одновременно занимают свободные приборы и начинают обслуживаться. Если приборов недостаточно для обслуживания всех фрагментов требований первого или второго класса, требования поступают на ожидание в очередь. Максимальное число *требований*, которое может находиться в первой и второй очереди, равно c_1 и c_2 соответственно. В связи с ограниченной вместимостью очередей, требования, не заставшие свободных мест в первой или второй очереди, теряются, то есть возвращаются обратно в источник.

Длительность обслуживания фрагмента любого класса на приборе есть экспоненциально распределённая случайная величина с параметром μ.

Ключевая особенность данной системы заключается в том, что фрагмент, обслуживание которого завершено, не освобождает обслуживающий прибор, а занимает (блокирует) его до момента, когда все родственные ему фрагменты, т.е. фрагменты одного и того же требования, не завершат своё обслуживание. Только после завершения обслуживания последнего фрагмента требование освобождает приборы и покидает систему.

В данном подразделе также вводятся следующие обозначения:

— $\boldsymbol{s} = ((q_1, q_2), \mathcal{A}, \mathcal{B})$ — вектор, описывающий состояние рассматриваемой системы обслуживания;

— $x(\boldsymbol{s}) = |\mathcal{A}|d_1 + |\mathcal{B}|d_2$ — число приборов, *занятых* требованиями первого или второго класса;

— $J(\boldsymbol{s}) = M - x(\boldsymbol{s})$ — число *свободных* приборов в системе.

В подразделе 2.2 определяется инфинитезимальный оператор $\boldsymbol{Q}$ и находится стационарное распределение $\boldsymbol{\pi}$ вероятностей состояний рассматриваемой системы как решение уравнения глобального баланса вместе с условием нормировки

$$\boldsymbol{\pi Q} = \mathbf{0}, \quad \boldsymbol{\pi 1} = \mathbf{1}.$$

В подразделе 2.3 получены следующие формулы для вычисления основных стационарных характеристик рассматриваемой системы обслуживания:

1. Математическое ожидание Q_1, Q_2 числа требований в очередях для каждого класса:
$$Q_i = \sum_{\boldsymbol{s} \in \mathcal{S}} q_i \pi(\boldsymbol{s}), \quad i = 1, 2.$$

2. Вероятность отказа, т.е. вероятность того, что требование застанет очередь полностью заполненной, для требований первого и второго класса:
$$p_{fi} = \sum_{(q_1, q_2, \mathcal{A}, \mathcal{B}) \in \mathcal{S}, q_i = c_i} \pi((q_1, q_2, \mathcal{A}, \mathcal{B})), \quad i = 1, 2.$$

3. Вероятность того, что обе очереди заполнены до предельных значений c_1 и c_2:
$$p_f = \sum_{\boldsymbol{s} \in \mathcal{S}, q_1 = c_1, q_2 = c_2} \pi((q_1, q_2, \mathcal{A}, \mathcal{B})).$$

4. Математическое ожидание W_i длительности пребывания требования в очереди для каждого класса:
$$W_i = \frac{Q_i}{\lambda_i (1 - p_{fi})}, \quad i = 1, 2.$$

5. Математическое ожидание G числа свободных приборов в системе:
$$G = \sum_{\boldsymbol{s} \in \mathcal{S}} (M - x(\boldsymbol{s})) \pi(\boldsymbol{s}).$$

6. Математическое ожидание S_1, S_2 числа требований на приборах для каждого класса:
$$S_i = \sum_{\boldsymbol{s}\in\mathcal{S}} \frac{x_i(\boldsymbol{s})\pi(\boldsymbol{s})}{d_i}, \quad i = 1, 2,$$
где $x_i(\boldsymbol{s}), i = 1, 2$ — число приборов, занятых требованиями первого или второго класса соответственно ($x_1(\boldsymbol{s}) = |\mathcal{A}|d_1, x_2(\boldsymbol{s}) = |\mathcal{B}|d_1$).
7. Математическое ожидание S числа требований на приборах:
$$S = S_1 + S_2.$$
8. Математическое ожидание длительности пребывания требований каждого класса в системе:
$$T_i = W_i + H_{d_i}\frac{1}{\mu}, \quad i = 1, 2,$$
где H_k есть k-ая частная сумма гармонического ряда, $H_k = \sum_{i=1}^{k} \frac{1}{i}$.
9. Математическое ожидание длительности пребывания требований в системе:
$$T = \frac{Q_1 + Q_2 + S_1 + S_2}{\lambda_1(1 - p_{f1}) + \lambda_2(1 - p_{f2})}.$$

Третий раздел «Система массового обслуживания с делением и слиянием требований с управлением» посвящен описанию системы массового обслуживания с делением и слиянием требований двух классов типа *split-merge* с управлением.

В подразделе 3.1 определяется множество управлений, которые могут быть применены в рассматриваемой системе обслуживания. Управление осуществляется в моменты перехода требований из очереди на обслуживающие приборы и состоит в том, чтобы определить номер очереди, из которой необходимо взять требование на обслуживание исходя из текущего состояния системы. Таким образом, множество управлений $\mathcal{Y}$ содержит два элемента, то есть $\mathcal{Y} = \{0, 1\}$, где управление $y = 0$ означает, что требование для обслуживания выбирается из очереди для требований первого класса, а управление $y = 1$ означает, что требование для обслуживания выбирается из очереди для требований второго класса.

В подразделе также описываются уравнения состояний рассматрива-

емой системы. Пусть $\alpha(\boldsymbol{s}, \boldsymbol{s}')$ — интенсивность перехода системы из состояния $\boldsymbol{s}$ в состояние $\boldsymbol{s}'$. Тогда, согласно имеющемуся множеству управлений $\mathcal{Y} = \{0, 1\}$, имеем:

1. *Если $q_1, q_2 \geq 1$ и систему покидает требование 1-го класса при $k_2 > 0$:*
   а) при управлении $y = 0$:

$$\alpha((q_1, q_2, \{a_1, \ldots, a_{n-1}, 1\}, \mathcal{B}), (q_1 - 1, q_2, \mathcal{A} - \{1\} + \{d_1\}, \mathcal{B})) = \mu|\mathcal{A}|_1;$$

   б) при управлении $y = 1$:

$$\alpha((q_1, q_2, \{a_1, \ldots, a_{n-1}, 1\}, \mathcal{B}), (q_1, q_2 - k_2, \mathcal{A} - \{1\}, \mathcal{B} + \{d_2\} \times k_2)) = \mu|\mathcal{A}|_1.$$

2. *Если $q_1, q_2 \geq 1$ и систему покидает требование 2-го класса при $k_1 > 0$:*
   а) при управлении $y = 0$:

$$\alpha((q_1, q_2, \mathcal{A}, \{b_1, \ldots, b_{m-1}, 1\}), (q_1 - k_1, q_2, \mathcal{A} + \{d_1\} \times k_1, \mathcal{B} - \{1\})) = \mu|\mathcal{B}|_1;$$

   б) при управлении $y = 1$:

$$\alpha((q_1, q_2, \mathcal{A}, \{b_1, \ldots, b_{m-1}, 1\}), (q_1, q_2 - 1, \mathcal{A}, \mathcal{B} - \{1\} + \{d_2\})) = \mu|\mathcal{B}|_1.$$

3. *Отсутствие управления:*
   а) если $a_n = 1, q_1 = 0, q_2 \geq 1$:

$$\alpha((0, q_2, \{a_1, \ldots, a_{n-1}, 1\}, \mathcal{B}), (0, q_2 - k_2, \mathcal{A} - \{1\}, \mathcal{B} + \{d_2\} \times k_2)) = \mu|\mathcal{A}|_1;$$

   б) если $a_n = 1, q_1 \geq 1, q_2 = 0$:

$$\alpha((q_1, 0, \{a_1, \ldots, a_{n-1}, 1\}, \mathcal{B}), (q_1 - 1, 0, \mathcal{A} - \{1\} + \{d_1\}, \mathcal{B})) = \mu|\mathcal{A}|_1;$$

   в) если $b_m = 1, q_1 = 0, q_2 \geq 1$:

$$\alpha((0, q_2, \mathcal{A}, \{b_1, \ldots, b_{m-1}, 1\}), (0, q_2 - 1, \mathcal{A}, \mathcal{B} - \{1\} + \{d_2\})) = \mu|\mathcal{B}|_1;$$

   г) если $b_m = 1, q_1 \geq 1, q_2 = 0$:

$$\alpha((q_1, 0, \mathcal{A}, \{b_1, \ldots, b_{m-1}, 1\}), (q_1 - k_1, 0, \mathcal{A} + \{d_1\} \times k_1, \mathcal{B} - \{1\})) = \mu|\mathcal{B}|_1.$$

В данном подразделе также определяется множество всех возможных стратегий управления $\boldsymbol{\Delta} = \{\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_i\}, i = 1, \ldots, |\mathcal{Y}|^\eta$. Элементы множества $\boldsymbol{\Delta}$ для каждого состояния $\boldsymbol{s}$ определяются следующим образом:

$$\boldsymbol{\delta}_i = (y_1, y_2, \ldots, y_\eta),$$

где y_j — элемент множества управлений $\mathcal{Y} = \{0, 1\}$, η — количество состояний системы, в которых возможно управление.

В подразделе 3.2 приведено описание методов оптимизации рассматриваемой системы, а именно: метода полного перебора, итерационного метода (для марковских процессов) и метода «Q-обучение» (машинное обучение). Посредством использования метода полного перебора были найдены оптимальные стратегии управления для минимизации математического ожидания длительности пребывания требований каждого класса в системе.

Четвертый раздел «Комплекс программ имитационного моделирования и численного анализа системы» посвящен описанию разработанного программного комплекса для исследования, анализа и оптимизации рассматриваемой системы обслуживания.

Подраздел 4.1 посвящен описанию программного модуля *model properties*, в котором определены основные параметры системы обслуживания (количество обслуживающих приборов, размерности очередей, количество фрагментов в каждом классе требований, интенсивности поступления требований, интенсивность обслуживания требований приборами системы);

Подраздел 4.2 посвящен описанию программного модуля *states*, содержащего классы и методы для генерации состояний системы, их обработки и представления.

Подраздел 4.3 содержит описание программного модуля *analytical calculations*, в котором реализован численный анализ системы (генерация инфинитезимального оператора системы, вычисление стационарных вероятностей состояний системы, расчёт стационарных характеристик и др.).

Подраздел 4.4 содержит описание программного модуля *simulation model*, в котором реализовано имитационное моделирование системы (обработчики событий; часы модельного времени для дискретно-событийного имитационного моделирования; требования, фрагменты, обслуживающие устройства как объекты для системы обслуживания).

Подраздел 4.5 посвящен описанию программного модуля *policy*, в котором реализованы классы и методы для внедрения управления в аналитическую и имитационную модели (генерация состояний, в которых возможно управление, и смежных с ними состояний; определение всех возможных векторов стратегий управления и т.д.).

Подраздел 4.6 посвящен описанию программного модуля *optimization*, содержащего методы оптимизации исследуемой системы обслуживания, такие как итерационный метод принятия решений и метод Q-обучения.

Пятый раздел «Численные примеры» посвящен описанию численных экспериментов над рассматриваемой системой обслуживания.

В подразделе 5.1 описаны результаты расчёта основных стационарных характеристик системы массового обслуживания без управления в зависимости от изменения интенсивности входящих потоков требований. На рисунках 2, 3 представлены графики зависимости м.о. длительности пребывания требований в системе T_1, T_2 и вероятностей отказа p_{f1}, p_{f2} для первого и второго классов соответственно от интенсивности входящего потока.

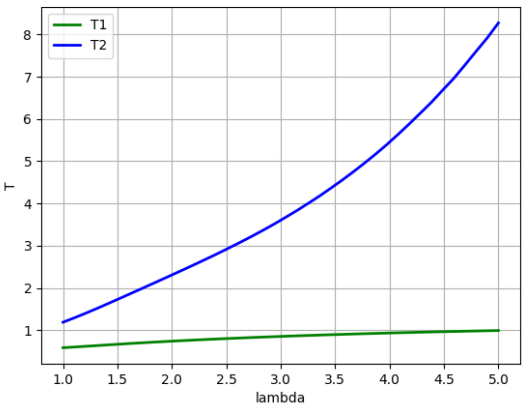

Рисунок 2 – Зависимость T_1, T_2 от интенсивности входящего потока

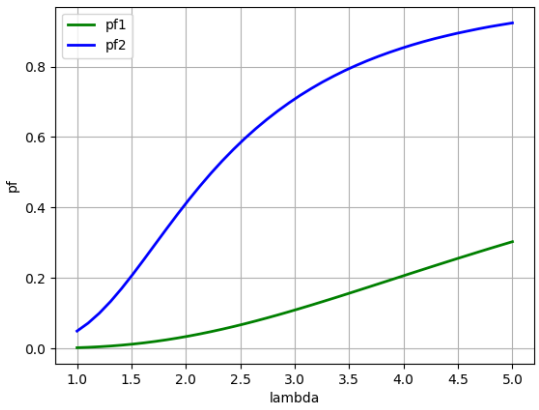

Рисунок 3 – Зависимость p_{f1}, p_{f2} от интенсивности входящего потока

В подразделе 5.2 описаны результаты расчёта основных стационарных характеристик системы массового обслуживания с управлением в зависимости от изменения интенсивности входящих потоков требований. На рисунках 4, 5 представлены графики зависимости м.о. длительности пребывания требований в системе T_1, T_2 и вероятностей отказа p_{f1}, p_{f2} для первого и вто-

рого классов соответственно от интенсивности входящего потока.

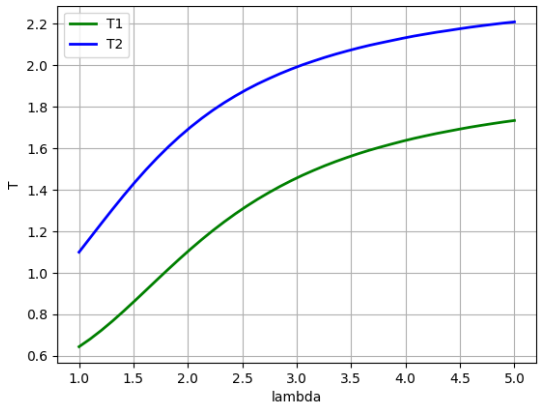

Рисунок 4 – Зависимость T_1, T_2 от интенсивности входящего потока

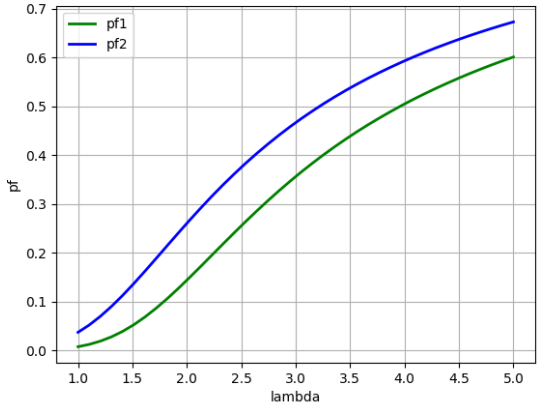

Рисунок 5 – Зависимость p_{f1}, p_{f2} от интенсивности входящего потока

В подразделе 5.3 описаны результаты расчёта стационарных характеристик системы обслуживания при каждой стратегии управления для заданных параметров системы. Получены оптимальные стратегии управления для минимизации значений стационарных характеристик системы.

В подразделе 5.4 описаны результаты применения итерационного метода. Было определено, что если доход за обслуживание требований первого класса превышает доход за обслуживание требований второго класса, то стратегия управления всегда соответствует выбору требований из первой очереди. Если же разница между доходом за обслуживание требований второго класса и доходом за обслуживание требований первого класса превышает 0.6 единиц, то стратегия управления всегда соответствует выбору требований из второй очереди.

В подразделе 5.5 описаны результаты применения метода Q-обучения. Построена Q-таблица, согласно которой была определена оптимальная стратегия управления, увеличивающая среднюю ожидаемую прибыль за единицу времени.

ЗАКЛЮЧЕНИЕ

В данной работе были получены следующие основные результаты:

1. Дано формальное описание системы массового обслуживания типа $split-merge$ с делением и слиянием требований двух классов без управления.
2. Описаны уравнения состояний исследуемой системы обслуживания, вычислено стационарное распределение вероятностей состояний системы, а также определены формулы для её основных стационарных характеристик.
3. Введено управление для исследуемой системы обслуживания. Приведено описание уравнений состояний системы с управлением и определено множество всех возможных стратегий управления.
4. Описаны методы оптимизации, используемые для поиска оптимальных стратегий управления, а именно: метод полного перебора, итерационный метод и метод Q-обучения.
5. Разработан комплекс программ имитационного моделирования и численного анализа системы массового обслуживания типа $split-merge$ с делением и слиянием требований двух классов.
6. Вычислены основные стационарные характеристики системы и найдена оптимальная стратегия управления посредством использования итерационного метода и метода Q-обучения.

Отдельные части бакалаврской работы были представлены на конференции:

1. Гуркова В. М., Заварзин А. С., Осипов О. А. Задача распределения нагрузки в сети массового обслуживания с делением и слиянием требований. – Информационные технологии и математическое моделирование (ИТММ-2019): Материалы XVIII Международной конференции имени А.Ф. Терпугова. Саратов, СГУ, 26-30 июня 2019 г. Томск: Изд-во НТЛ, 2019, часть 2, 149-152.
2. Гуркова В. М., Осипов О. А. Исследование split-merge системы с двумя классами требований и потерями. – Математическое и программное обеспечение информационных, технических и экономических систем (МПОИТЭС-2020): Материалы Международной научной конференции. Томск, 28-30 мая 2020 г. Томск: Изд-во ТГУ, 2020, 254-259.

Основные источники информации:

1. Thomasian A. Analysis of Fork/Join and Related Queueing Systems // ACM Computing Surveys. – New York, 2014. – Vol. 47, № 2. – 17:1–17:71.
2. Narahari Y., Sundarrajan P. Performability Analysis of Fork-join Queueing Systems // Journal of the Operational Research Society. – 1995. – Vol. 46, № 10. – P. 1237–1249.
3. Nelson R., Tantawi A. N. Approximate analysis of fork/join synchronization in parallel queues // IEEE Transactions on Computers. – 1988. – Vol. 37, № 6. – P. 739–743.
4. Rizk A., Poloczek F., Ciucu F. Computable Bounds in Fork-Join Queueing Systems // ACM SIGMETRICS Performance Evaluation Review. – New York, 2015. – Vol. 43, № 1. – P. 335–346.
5. Baccelli F., Massey W. A., Towsley D. Acyclic fork-join queuing networks // Journal of the ACM. – 1989. – Vol. 36, № 3. – P. 615–642.
6. Boxma O. J., Koole G., Liu Z. Queueing-Theoretic Solution Methods for Models of Parallel and Distributed Systems: tech. rep. / Centrum voor Wiskunde en Informatica. – 1994. – 25 p.
7. Plateau B., Fourneau J.-M. A methodology for solving Markov models of parallel systems // Journal of parallel and distributed computing. – 1991. – Vol. 12, № 4. – P. 370-387.
8. Flatto L., Hahn S. Two Parallel Queues Created by Arrivals with Two Demands I // SIAM Journal on Applied Mathematics. – 1984. – Vol. 44, № 5. – P. 1041–1053.
9. Varki E., Dowdy L. W. Analysis of balanced fork-join queueing networks // ACM SIGMETRICS Performance Evaluation Review. – 1996. – Vol. 24, № 1. – P. 232–241.