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

Кафедра системного анализа и
автоматического управления

ИССЛЕДОВАНИЕ МАТЕМАТИЧЕСКОЙ МОДЕЛИ РАСПРЕДЕЛЕННОЙ ВЫЧИСЛИТЕЛЬНОЙ СИСТЕМЫ

АВТОРЕФЕРАТ МАГИСТЕРСКОЙ РАБОТЫ

студентки 2 курса 271 группы
направления 09.04.01 — Информатика и вычислительная техника
факультета КНиИТ
Постновой Оксаны Сергеевны

Научный руководитель
к. ф.-м. н., доцент _____ И. Е. Тананко

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

Саратов 2024

ВВЕДЕНИЕ

Актуальность темы. В настоящее время область распределенных вычислений охватывает все аспекты вычислений и доступа к информации через множество элементов обработки, соединенных любой формой коммуникационной сети, будь то локальная или глобальная сеть. В современом мире появляется все больше новых приложений, которым необходима распределенная обработки. Новые технологии в области сетевых и аппаратных технологий, снижение стоимости оборудования и более высокий уровень информируемости пользователей. Все это привело к преобразованию распределенных вычислительных систем в экономически эффективную, высокопроизводительную и отказоустойчивую действительность.

Высокая потребность в разработке и исследовании распределенных вычислительных систем и параллельной обработки объектов способствует более интенсивному развитию теории и методов анализа систем и сетей массового обслуживания с учетом деления и слияния требований.

Производительность вычислительных систем и систем хранения данных может быть улучшена за счет параллелизма. В вычислительных системах время выполнения программы может быть улучшено путем разбиения программы на подпрограммы, которые затем выполняются параллельно на разных процессорах. С точки зрения анализа производительности, такие параллельные ресурсы моделируются очередями fork-join [1], поскольку задание состоит из различных подзадач и затем соединяется (выходит) после завершения обслуживания всех этих подзадач.

Цель магистерской работы — разработка и исследование математической модели распределенной вычислительной системы.

Поставленная цель определила **следующие задачи**:

1. Обзор систем и сетей массового обслуживания с делением и слиянием требований и методов их анализа;
2. Разработка метода анализа для сети масового обслуживания с делением и слиянием требований с древовидной топологией;
3. Исследование метода анализа рассматриваемой сети массового обслуживания, используемой в качестве модели распределенной вычислительной системы.

Методологические основы исследования систем и сетей массового

обслуживания с делением и слиянием требований представлены в работах R. Nelson, О. А. Осипова, И. Е. Тананко, E. Varki, A. Kumar, А. В. Горбуновой.

Теоретическая значимость магистерской работы Разработан метод анализа для открытой сети массового обслуживания с делением и слиянием требовани, в которой структура сети представлена ориентированным ациклическим графом, который расширяет круг задач, решаемых в теории массового обслуживания, поскольку позволяет рассмотреть особенности структуры и функционирования сети массового обслуживания с делением и слиянием тренбований.

Практическая значимость магистерской работы Разработанный метод анализа для открытой сети массового обслуживания с делением и слиянием требовани и древовидной структурой может быть использован в качестве математических моделей распределенных вычислительных систем, систем передачи информации, гибких производственных систем, систем управления запасами.

Структура и объем работы. Магистерская работа состоит из введения, 4 разделов, заключения, списка использованных источников и одного приложения. Общий объем работы — 71 страница, из них 47 страниц — основное содержание работы, включая 15 рисунков и 8 таблиц, список использованных источников — 36 наименований.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Первый раздел «Обзор результатов исследований систем и сетей массового обслуживания с делением и слиянием требований» посвящен обзору основных опубликованных результатов исследований для классических систем [2] и сетей массового обслуживания с делением и слиянием требований [3, 4].

В *подразделе 1.1* представлены работы по системам массового обслуживания с делением и слиянием требований.

В *подразделе 1.2* представлены работы по сетям массового обслуживания с делением и слиянием требований.

Второй раздел «Однородные открытые сети обслуживания. Сеть Джексона» посвящен описанию однородных открытых сетей обслуживания [5, 6] и описанию сети Джексона.

Рассматривается ситема типа $M/M/m$, для которой основополагающим результатом является теорема Берке:

Теорема 1. Входящий пуассоновский поток, протекая через обслуживающий прибор с экспоненциальным распределением времени обслуживания, порождает выходящий пуассоновский поток.

Другими словами, если выходящий поток в стационарной системе массового обслуживания типа $M/M/m$ с пуассоновским входящим потоком с праметром λ и экспоненциальным распределением времени обслуживания с параметром μ, то в каждом из m обсуживающих приборов выходящий поток является пуассоновским с тем же самым параметром λ.

Джексон доказал, что каждая система ведет себя в сети так, как если бы она была независимой системой $M/M/m$ с входящим пуассоновским потоко с параметром λ_i. В общем случае полный входящий поток не является пуассоновским. Так же было доказано, что совместное распределение по всем системам разлагается в произведение маргинальных распределений, то есть

$$p(k_1, k_2, ..., k_N) = p_1(k_1)p_2(k_2)...p_N(k_N)$$

и что $p_i(k_i)$ представляет собой стационарные вероятности для классической системы $M/M/m$.

В *подразделе 2.1* описана классическая модель сети Джексона [7]. Модель является обобщением классической модели системы массового обслуживания типа $M/M/m$ на произвольную взаимосвязанную открытую сеть массового обслуживания с экспоненциальным распределением времени обслуживания и пуассоновским входящим потоком. В частности, имеется N систем массового обслуживания с номерами i, имеющие m обслуживающих приборов, дисциплина обслуживания «первый пришел - первый обслужен» и очередь ожидания неограниченной вместимости. Входящий поток в систему i является пуассоновским с интенсивностью λ_i, и предполагается, что входящие потоки независимы. Длительности обслуживания в системе i независимы и имеют экспоненциальное распределение с параметром μ_i, а также не зависит от поступления требований в систему i. Требование, покидающее систему i, немедленно и независимо переходит в систему j с вероятностью p_{ij},

требование покидает сеть с вероятностью

$$q_i = 1 - \sum_{j=1}^{N} p_{ij}.$$

В *подразделе 2.2* описана сеть Джексона с древовидной структурой [8].

Третий раздел «Математическпя модель распределенной вычислительной системы» посвящен описанию математической модели распределенной вычислительной системы. Рассматривается открытая экспоненциальная сеть массового обслуживания, состоящая из L одноприборных систем массового обслуживания S_i типа $M/M/1$ с интенсивностями обслуживания μ_i, $i = 1, ..., L$. Из источника S_0 в сеть обслуживания с интенсивностью λ_0 поступает пуассоновский поток требований одного класса. Предполагается, что сеть массового обслуживания — ациклическая, то есть структура сети представляется ориентированным ациклическим графом.

Каждое из поступающих требований состоит из двух фрагментов. В момент поступления требования в сеть, это требование разделяется на два фрагмента, называемых родственными, и оба фрагмента независимо друг от друга поступают в одну из систем обслуживания сети, с которыми связан источник. Фрагменты одного требования независимо друг от друга и независимо от фрагментов других требований переходят между системами сети обслуживания до тех пор, пока не завершится их обслуживание в сети. После этого фрагменты требований поступают в систему S_{L+1}, называемую системой сборки. Родственные фрагменты, ранее принадлежавшие одному требованию, собираются в единое требование, которое покидают систему сборки. Подробное описание функционирования системы сборки приводится далее.

Переходы фрагментов между системами сети и связь с источником и системой сборки производятся в соответствии с маршрутной матрицей $\Theta = (\theta_{ij})$, $i = 0, 1, ..., L$, $j = 1, ..., L + 1$, где θ_{ij} — вероятность того, что после пребывания в системе S_i фрагмент перейдет в систему S_j.

Сеть массового обслуживания с введенными ранее параметрами систем обслуживания, в которой обслуживаются фрагменты требований, без системы сборки обозначим через N. Элементы маршрутной матрицы $\bar{\Theta} = (\bar{\theta}_{ij})$, $i, j = 0, 1, ..., L$, сети N определяются как $\bar{\theta}_{ij} = \theta_{ij}$, если $j \neq 0$, и $\bar{\theta}_{i,0} = \theta_{i,L+1}$. Предполагается, что в сети N число смежных для S_i, $i = 0, 1, ..., L$, систем

обслуживания, в которые возможен переход фрагментов из S_i, намного больше двух (числа фрагментов требований), и вероятности $\bar{\theta}_{ij} > 0$, $j = 0, 1, ..., L$, сравнимы. В этом случае выходящий из сети N поток фрагментов требований можно считать пуассоновским с интенсивностью $2\lambda_0$.

Предполагается, что для сети N выполняется необходимое условие существования стационарного режима

$$\lambda_0 < \frac{1}{2} \min_{i=1,...,L} \frac{\mu_i \omega_0}{\omega_i} = a,$$

где $\omega = (\omega_j)$, $j = 0, 1, ..., L$, — вектор относительных интенсивностей потоков фрагментов требований в сети N — находится как решение уравнения $\omega = \omega\bar{\Theta}$ с условием нормировки $\sum \omega_j = 1$.

Система сборки S_{L+1} состоит из бесконечного числа обслуживающих приборов. Назначение этой системы — сбор требований из родственных фрагментов, поступающих из сети N. При поступлении в систему сборки одного из родственных фрагментов — первого по времени поступления в систему, называемого «первым» фрагментом требования, — он занимает свободный обслуживающий прибор. В момент поступления «второго» из родственных фрагментов мгновенно происходит объединение фрагментов в единое требование, которое покидает систему сборки, и освобождение обслуживающего прибора. Таким образом, длительность сборки требования из родственных фрагментов или длительность занятости обслуживающего прибора в системе сборки совпадает с длительностью пребывания в системе «первого» фрагмента или длительностью интервала времени между родственными фрагментами в выходящем из сети N потоке фрагментов.

Систему сборки будем представлять системой массового обслуживания типа $M/M/\infty$. В систему поступает пуассоновский поток фрагментов с интенсивностью λ_0, так как обслуживаются в системе только «первые» фрагменты. В системе бесконечное число обслуживающих приборов, в которых обслуживаются уникальные фрагменты требований. Длительность интервала времени сборки требования из фрагментов является экспоненциально распределенной случайной величиной с неизвестным параметром μ. Определим параметр μ. Из результатов имитационного моделирования сети массового обслуживания с древовидной структурой можно предположить, что пара-

метр μ является монотонно убывающей линейной функцией интенсивности λ_0, $\mu = \mu(\lambda_0)$, $\lambda_0 \in (0, a)$. Графиком функции $\mu(\lambda_0)$ является прямая, проходящая через две точки $(0, \nu)$ и $(a, 0)$, где ν — наибольшее значение μ, получаемое при $\lambda_0 \approx 0$ или при $\lambda_0 = \epsilon$, где $\epsilon > 0$ — малое число.

Обозначим τ_i — длительность пребывания в сети N фрагментов требований, поступивших в сеть через систему S_i, $i = 1, ..., L$, $g_i(u)$ — преобразование Лапласа плотности распределения случайной величины τ_i.

Плотность распределения длительности пребывания фрагментов требований в системе S_i типа $M/M/1$ имеет преобразование Лапласа

$$h_i(u) = \frac{\mu_i(1-\psi_i)}{u + \mu_i(1-\psi_i)}, \quad u \geq 0,$$

где $\psi_i = \lambda_i/\mu_i$, $\lambda_i = 2\lambda_0\omega_i/\omega_0$ — интенсивность потока фрагментов требований в систему S_i.

Математическое ожидание длительности пребывания в сети N фрагментов требований:

$$E\tau_0 = \frac{1}{\lambda_0}\sum_{i=1}^{L}\frac{\psi_i}{1-\psi_i}.$$

Обозначим: $P(n)$ — стационарная вероятность пребывания сети N в состоянии n, $n = 0, 1, 2, ...$, где n — число фрагментов в сети N; $p(d_1, d_2)$ — стационарная вероятность пребывания сети N в состоянии (d_1, d_2), где d_1, d_2 — число соответственно «первых» и «вторых» фрагментов в сети, $d_1, d_2 = 0, 1, 2, ...$, $d_2 \geq d_1$; $\hat{p}(k)$ — стационарная вероятность пребывания системы сборки в состоянии k, $k = 0, 1, 2, ...$, где k — число фрагментов («первых» фрагментов) в системе сборки. Очевидно, что $\hat{p}(d_2 - d_1) = p(d_1, d_2)$.

Значение ν:

$$\ln\nu + \frac{\lambda_0}{\nu} = \ln\frac{\lambda_0}{\hat{p}(1)}.$$

Решение уравнения можно найти численно или использовать приближенное равенство $\nu \approx \lambda_0/\hat{p}(1)$. Заметим, что $\hat{p}(1)/\lambda_0$ — математическое ожидание длительности реакции сети массового обслуживания, когда в ней находится не более одного требования.

Предполагая линейную зависимость μ от λ_0 и используя уравнение пря-

мой, проходящей через две точки $(0, \nu)$ и $(a, 0)$, получим

$$\mu = -\frac{\nu}{a}\lambda_0 + \nu, \qquad \lambda_0 \in (0, a).$$

Математическое ожидание числа требований в сети N^c

$$\bar{q}^c = \frac{1}{2}(\bar{q} + \lambda_0/\mu).$$

Математическое ожидание длительности реакции сети N^c

$$\bar{\tau}_0^c = \bar{\tau}_0 + 1/\mu,$$

где

$$\bar{\tau}_0 = E\tau_0 = \frac{\bar{q}}{2\lambda_0}$$

— есть математическое ожидание длительности реакции сети N.

В четвертом разделе «Анализ результатов моделирования распределенной вычислительной системы» приводятся сравнения результатов аналитического и имитационного моделирования рассматриваемой сети массового обслуживания для численного определения точности разработанного метода анализа.

Пример 1

Рассмотрим сеть массового обслуживания с $L = 10$ параллельнами системами обслуживания $S_i, i = 1, ..., 10$. Вектор интенсивностей обслуживания

$$\mu = (4.0 \;\; 4.0 \;\; 4.0 \;\; 4.0 \;\; 4.0 \;\; 4.0 \;\; 4.0 \;\; 4.0 \;\; 4.0 \;\; 4.0).$$

Каждое требование, поступающее в сеть, состоит из двух фрагментов. Интенсивность входящего потока в сеть $\lambda_0 = [0.0001, 1, 3, 5]$.

Маршрутная матрица имеет следующий вид:

$$\Theta = \begin{pmatrix} 0.0 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \\ 1.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 1.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 1.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 1.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 1.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \end{pmatrix}$$

Применяя аналитический метод, представленный в 3 главе, получаем относительную интенсивность потока требований из источника $\omega = 0.3329$. Время реакции сети при $\lambda_0 = 0.0001$ равно 0.5009, а математическое ожидание длительности сборки имеет значение 0.50086, интенсивность $\nu = 1.997$. Максимальное значение λ_0 в этой сети обслуживания равно 9.982.

Уравнение для вычисления интенсивности обслуживания в системе-сборке имеет вид:

$$\mu = -0.2 \cdot \lambda_0 + 1.997.$$

На рисунке 1 представлен график зависимости длительности пребывания фрагментов в сети от интенсивности входящего потока в сеть. На рисунке 2 график зависисмости интенсивноть обслуживания в системе сборки от интенсивности входящего потока, где снимний график — это результаты имитационного моделирования, а красный график — аналитического моделирования.

Пример 2

Рассмотрим сеть массового обслуживания с $L = 10$, с древовидной топологией сети. Интенсивности обслуживания μ_i, $i = 1, ..., 10$ являются неравномерными

$$\mu = (20.0 \;\; 10.0 \;\; 40.0 \;\; 4.0 \;\; 30.0 \;\; 45.0 \;\; 4.0 \;\; 4.0 \;\; 2.0 \;\; 40.0).$$

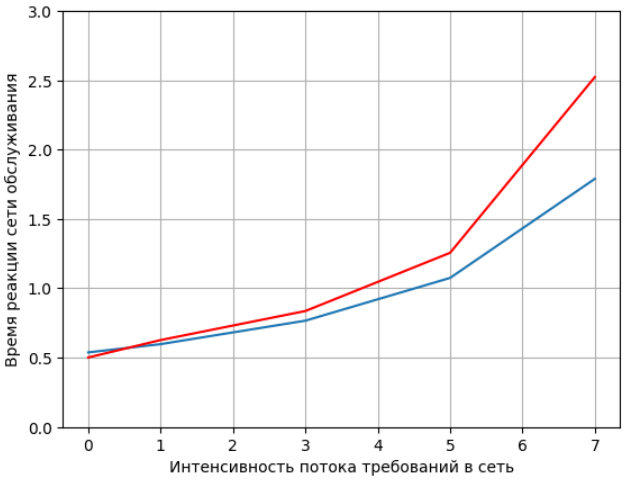

Рисунок 1 – Зависимость времени реакции сети от интенсивности входящего потока

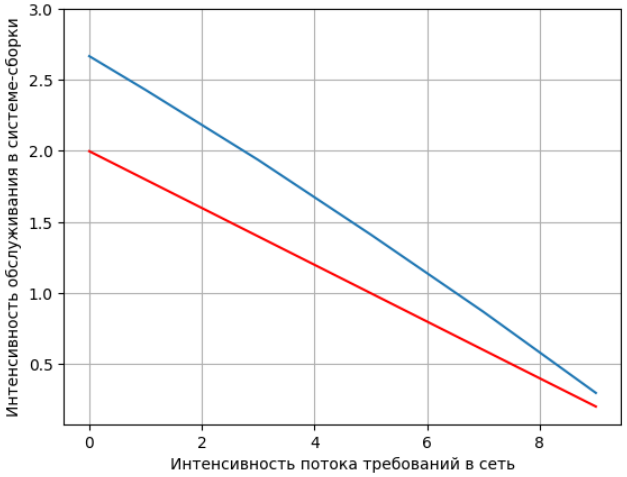

Рисунок 2 – Зависимость интенсивности обслуживания в системе-сборке от интенсивности входящего потока

Интенсивность входящего потока в сеть $\lambda_0 = [0, 1, 2, 3, 4, 4.5]$. Каждое требование разбивается на 2 фрагмента. Время реакции в данной сети обслуживания при интенсивности $\lambda_0 = 0$ равно 0.30166, а относительная интенсивность потока требований из источника $\omega = 0.3329$. Максимальное значение λ_0 в этой сети обслуживания меньше 4.991, а наибольшая интенсивность обслуживания $\nu = 3.315$.

Уравнение для вычисления интенсивности обслуживания в системе-сборке имеет вид:

$$\mu = -0.664 \cdot \lambda_0 + 3.315.$$

Маршрутная матрица имеет следующий вид:

$$\Theta = \begin{pmatrix} 0.0 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \\ 1.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 1.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 1.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 1.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 1.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \end{pmatrix}$$

На рисунке 3 представлен график зависимости длительности пребывания фрагментов в сети от интенсивности входящего потока в сеть. На рисунке 4 график зависисмости интенсивноть обслуживания в системе сборки от интенсивности входящего потока, где сниний график — это результаты имитационного моделирования, а красный график представляет результаты аналитического моделирования.

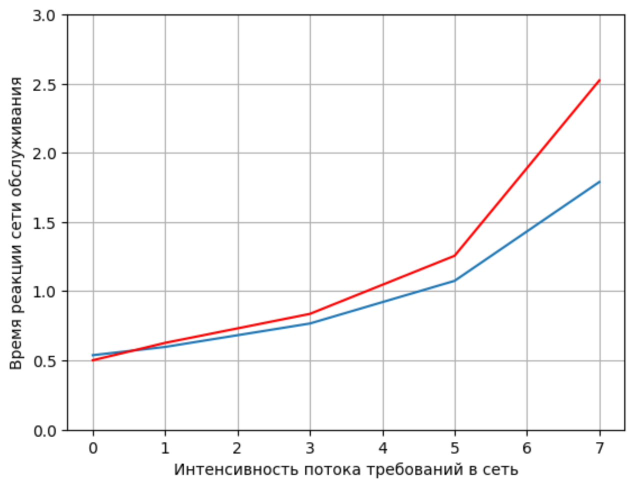

Рисунок 3 – Зависимость времени реакции сети от интенсивности входящего потока

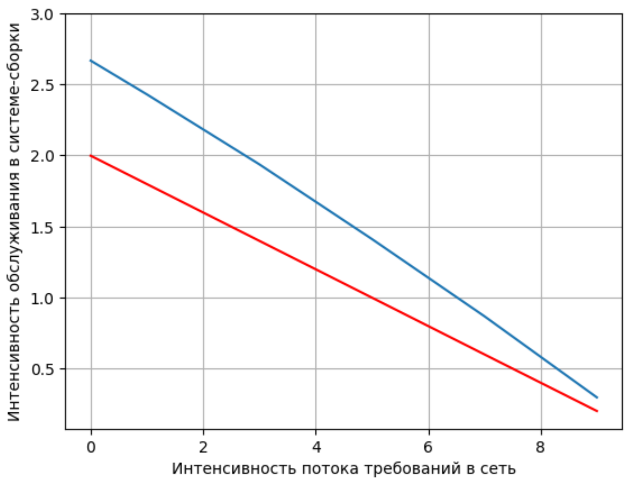

Рисунок 4 – Зависимость интенсивности обслуживания в системе-сборке от интенсивности входящего потока

ЗАКЛЮЧЕНИЕ

Распределенные вычисления касаются всех форм вычислений, доступа к информации и обмена информацией между несколькими вычислительными платформами, соединенными компьютерными сетями. Проектирование распределенных вычислительных систем — сложная задача. Она требует глубокого понимания проблем проектирования и глубокого понимания теоретических и практических аспектов их решения.

Разработанный метод анализа может быть использован в качестве математических моделей распределенных вычислительных систем, систем передачи информации, гибких производственных систем, систем управления запасами.

Отдельные части магистерской работы были представлены на конференции:

1. Карпенко, О. С. Использование моделей массового обслуживания при исследовании распределенных вычислительных систем студентами IT направлений / О. С. Карпенко, И. Е. Тананко // Информационные технологии в образовании (ИТО-Саратов-2022): Материалы XIV Всероссийской научно-практической конференции «Информационные технологии в образовании». — 2022. — С. 137–140.
2. Карпенко, О. С. Исследование имитационной модели открытой сети массового обслуживания с делением и слиянием требований / О. С. Карпенко, И. Е. Тананко, Е. С. Рогачко // Системы управления, информационные технологии и математическое моделирование (СУИТиММ-Омск-2023): Материалы V Всероссийской с международным участием научно- практической конференции. — 2023. — в печати.
3. Карпенко, О. С. Построение имитационной модели открытой сети массового обслуживания с делением и слиянием требований / О. С. Карпенко, И. Е. Тананко // Образование. Технологии. Качество (ОТК-Саратов-2023): Материалы XIV Всероссийской научно-практической конференции «Образование. Технологии. Качество». — 2023. — в печати.

Основные источники информации:

1. Thomasian A. Analysis of Fork/Join and Related Queueing Systems // ACM Computing Surveys. – 2014. – V. 47. – №. 2. – P. 1-71.
2. Горбунова, А. В. Обзор систем параллельной обработки заявок // Вест-

ник Российского университета дружбы народов. Серия: Математика. Информатика. Физика. — 2017. — Т. 25, № 4. — С. 350–362.

3. Nelson R. Approximate Analysis of Fork/Join Synchronization in Parallel Queues // IEEE Transactions on Computers. – 1988. – V. 38. – №. 6. – P. 335-346.
4. Flatto L., Hahn S. Two Parallel Queues Created by Arrivals with Two Demands I // SIAM Journal on Applied Mathematics. – 1984. – V. 44. – №. 5. – P. 1041–1053.
5. Митрофанов, Ю. И. Анализ систем массового обслуживания: учебно-методическое пособие / Ю. И. Митрофанов, Е. С. Рогачко, Н. П. Фокина. — Саратов: Изд-во «Научная книга», 2009. — 55 с.
6. Клейнрок, Л. Теория массового обслуживания : Учебник — М. : Машиностроение, 1979. — 432 с.
7. Jackson, J. -R. Networks of waiting lines // Operations Res. – 1957. – V. 5. – №. 7. – P. 518–521.
8. Lemoine, J. State of the Art Networks of Queues A Survey of Equilibrium Analysis // Management Science. – 1977. – V. 4. – №. 24. – P. 464-481.