На правах рукописи

КАМИЛЬ Висам Абдуладим Камиль

УПРАВЛЕНИЕ ПРОЦЕССАМИ ОБРАБОТКИ ДАННЫХ В МЕГАСЕТЯХ НА ОСНОВЕ ГРАФОВЫХ МОДЕЛЕЙ И СИСТЕМЫ РАЗРАБОТКИ ПОТОКОВЫХ ПРИЛОЖЕНИЙ

Специальность: 2.3.5. Математическое и программное обеспечение вычислительных систем, комплексов и компьютерных сетей

АВТОРЕФЕРАТ

диссертации на соискание ученой степени
кандидата технических наук

Воронеж – 2025

Работа выполнена в ФГБОУ ВО «Воронежский государственный технический университет».

Научный руководитель: **Мутин Денис Игоревич**, доктор технических наук, доцент

Официальные оппоненты: **Перепёлкин Дмитрий Александрович**, доктор технических наук, профессор, ФГБОУ ВО «Рязанский государственный радиотехнический университет им. В.Ф. Уткина», декан факультета вычислительной техники

Корнеев Андрей Мастиславович, доктор технических наук, профессор, ФГБОУ ВО "Липецкий государственный технический университет", профессор кафедры общей механики

Ведущая организация: **ФГБОУ ВО “Самарский государственный технический университет»**

Защита состоится «23» мая 2025 года в 14^{00} часов в конференц-зале на заседании диссертационного совета 24.2.286.04, созданного на базе ФГБОУ ВО «Воронежский государственный технический университет», по адресу: г. Воронеж, Московский просп., д. 14, ауд. 216.

С диссертацией можно ознакомиться в научно-технической библиотеке ФГБОУ ВО «Воронежский государственный технический университет» и на сайте https://cchgeu.ru.

Автореферат разослан «28» марта 2025 г.

Ученый секретарь диссертационного совета  Гусев Константин Юрьевич

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы. Информационно-вычислительные сети являются основой современного цифровизированного общества. В этой связи особенно важной является задача обеспечения согласованного функционирования структурных элементов таких принципиально распределенных систем. При этом одна из подзадач связана с необходимостью реализации устойчивости образованных мегасистем к штатным или несанкционированным внешним возмущениям. В конечном счете взаимозависимость между устойчивостью сети и ее структурой позволяет рационализировать интегрированную производительность именно через структуру мегасети, учитывая при этом множество параметров. Большой вклад в развитие методов создания и управления большими сетями внесли Зиганурова Р.А., Ковалев И.В., Кравец О.Я., Олескин А.В., Antsaklis P.J., Jovanovic M.R., Murray R.M. В теоретическом плане ряд методов создания мегасетей сводится к оптимизации выбора главных кластеров и мостов каждого кластера. Разумной идеей кажется модифицировать архитектуру системы управления облачными средами, расширив ее на случай модели объединенной сети, в которой узлы управляются зашумленной динамикой согласования, а веса ребер могут быть положительными или отрицательными.

С развитием таких технологий, как эффективные алгоритмы оптимизации сети, агрегирование данных и маршрутизация стали ключевыми подходами к оптимизации. Информация для сенсорных узлов, широко распространенных в современных информационно-вычислительных системах, становится высокодоступной. Каждый раз вместо того, чтобы принимать во внимание отдельные сенсорные узлы, система может распределять свои ограниченные ресурсы в определенных кластерах для оптимизации затрат и ресурсов. Важные знания могут быть извлечены с использованием соответствующих и эффективных алгоритмов интеллектуального анализа данных. Необходимы алгоритмы кластеризации групп сенсорных узлов мегасети, обеспечивающие инвариантность к большому объему сетевых данных и динамичности сетевого местоположения узлов.

За последние несколько лет резко выросла значимость технологий и приложений, требующим обработки больших объемов данных, порождаемых мегасетями. Сложность извлечения такой информации обычно связана с объемом, скоростью и разнообразием анализируемых данных. Конкретные сценарии работы с большими данными могут обладать одной или несколькими из этих характеристик. В частности, для тех сценариев, в которых объем и скорость имеют первостепенное значение, в игру вступают распределенные потоковые платформы. Они позволяют разрабатывать потоковые приложения, т.е. приложения, которые обрабатывают потенциально бесконечные потоки данных, непрерывно и быстро обновляя полезные выходные данные (статистику, прогнозы и т.д.). Отсюда понятна необходимость создания архитектуры системы разработки потоковых приложений для обработки данных на основе моделей потоков данных

Таким образом, актуальность темы диссертационного исследования

продиктована необходимостью дальнейшего развития средств математического и программного обеспечения управления процессами обработки данных в мегасетях на основе графовых моделей и системы разработки потоковых приложений.

Тематика диссертационной работы соответствует научному направлению ФГБОУ ВО «Воронежский государственный технический университет» «Вычислительные комплексы и проблемно-ориентированные системы управления».

Целью работы является разработка математического и программного обеспечения управления процессами обработки данных в мегасетях на основе графовых моделей и системы разработки потоковых приложений для обработки данных с использованием специальной метрики согласованности и интеграции логики приложений.

Задачи исследования. Для достижения поставленной цели необходимо решить следующие задачи:

1. Провести анализ проблем создания математического и программного обеспечения управления процессами обработки данных в мегасетях на основе графовых моделей и системы проектирования потоковых приложений с использованием специальной метрики согласованности и интеграции логики приложений.

2. Создать графовую модель синтеза мегасети из набора непересекающихся подграфов с согласованностью в виде нормы H_2, обеспечивающую соединение подграфов с обеспечением оптимальной согласованности финальной мегасети.

3. Сформулировать оптимизационную задачу выбора мостов подграфов мегасети с построением соединительных ребер, обеспечивающую получение оценок минимальной и максимальной согласованности с весами ребер на положительной полуоси.

4. Разработать алгоритм кластеризации данных групп сенсорных узлов мегасети, обеспечивающий инвариантность к большому объему сетевых данных и динамичности сетевого местоположения узлов.

5. Создать архитектуру системы разработки потоковых приложений для обработки данных на основе моделей потоков данных, обеспечивающую интеграцию в среду IDE поверх UML.

Объект исследования: процессы обработки данных в мегасетях с потоковыми приложениями в их составе.

Предмет исследования: средства математического и программного управления процессами обработки данных в мегасетях на основе алгоритмов кластеризации данных и архитектуры системы разработки потоковых приложений для обработки данных.

Методы исследования. При решении поставленных в диссертации задач использовались методы теории графов, теории вероятностей, теории принятия решений, а также методы объектно-ориентированного программирования.

Тематика работы соответствует следующим пунктам паспорта специ-

альности 2.3.5 «Математическое и программное обеспечение вычислительных систем, комплексов и компьютерных сетей»: п.3 «Модели, методы, архитектуры, алгоритмы, языки и программные инструменты организации взаимодействия программ и программных систем»; п.9 «Модели, методы, алгоритмы, облачные технологии и программная инфраструктура организации глобально распределенной обработки данных».

Научная новизна работы. В диссертации получены следующие результаты, характеризующиеся научной новизной:

1. Графовая модель синтеза мегасети из набора непересекающихся подграфов с согласованностью в виде нормы H_2, отличающаяся непрерывностью весов ребер на всей числовой оси и зашумленной динамикой согласования узлов, реализующая соединение подграфов с обеспечением оптимальной согласованности финальной мегасети.

2. Оптимизационная задача выбора мостов подграфов мегасети с построением соединительных ребер, отличающаяся параллелизмом процесса решения и обеспечивающая получение оценок минимальной и максимальной согласованности с весами ребер на положительной полуоси.

3. Алгоритм кластеризации данных групп сенсорных узлов мегасети, отличающийся использованием нечеткой логики для учета гетерогенных параметров сенсорной сети и гетерогенного управления сетью, обеспечивающий инвариантность к большому объему сетевых данных и динамичности сетевого местоположения узлов.

4. Архитектура системы разработки потоковых приложений для обработки данных на основе моделей потоков данных, отличающаяся интеграцией логики приложения в модель и обеспечивающая интеграцию в среду IDE поверх UML.

Теоретическая и практическая значимость исследования заключается в разработке специальных средств математического и программного обеспечения управления процессами обработки данных в мегасетях на основе графовых моделей и системы проектирования потоковых приложений с использованием специальной метрики согласованности и интеграции логики приложений.

Теоретические результаты работы могут быть использованы в проектных и научно-исследовательских организациях, занимающихся разработкой платформенно-инвариантных систем управления мегасетями с потоковыми приложениями обработки данных.

Положения, выносимые на защиту

1. Графовая модель синтеза мегасети из набора непересекающихся подграфов с согласованностью в виде нормы H_2 реализует соединение подграфов с обеспечением оптимальной согласованности финальной мегасети.

2. Оптимизационная задача выбора мостов подграфов мегасети с построением соединительных ребер обеспечивает получение оценок минимальной и максимальной согласованности с весами ребер на положительной полуоси.

3. Алгоритм кластеризации данных групп сенсорных узлов мегасети обеспечивает инвариантность к большому объему сетевых данных и динамичности сетевого местоположения узлов.

4. Архитектура системы разработки потоковых приложений для обработки данных на основе моделей потоков данных обеспечивает интеграцию в среду IDE поверх UML.

Результаты внедрения. Основные результаты внедрены в Научно-исследовательском институте вычислительных комплексов им. М. А. Карцева» (г. Москва) при проектировании распределенной информационно-вычислительной системы, в учебный процесс Воронежского государственного технического университета в рамках дисциплин: «Вычислительные машины, системы и сети», «Информационные сети и телекоммуникационные технологии», а также в рамках курсового и дипломного проектирования.

Апробация работы. Основные положения диссертационной работы докладывались и обсуждались на следующих конференциях: XXIX International Open Science Conference «Modern informatization problems in the technological and telecommunication systems analysis and synthesis» (Yelm, WA,, USA, 2024); VII Международной НПК «Наука и технологии: перспективы развития и применения» (Петрозаводск, 2024); VI Всероссийской НПК «Информационные технологии в экономике и управлении» (Махачкала, 2024); XXX International Open Science Conference «Modern informatization problems in the technological and telecommunication systems analysis and synthesis» (Yelm, WA,, USA, 2025), а также на научных семинарах кафедры автоматизированных и вычислительных систем ВГТУ (2023-2025 гг.).

Достоверность результатов обусловлена корректным использованием теоретических методов исследования и подтверждена результатами сравнительного анализа данных вычислительных и натурных экспериментов.

Публикации. По результатам диссертационного исследования опубликовано 12 научных работ (2 – без соавторов), в том числе 5 – в изданиях, рекомендованных ВАК РФ (из них 1 – в издании Wos и одно свидетельство о регистрации программы для ЭВМ). В работах, опубликованных в соавторстве и приведенных в конце автореферата, лично автором получены следующие результаты: [1, 7] - графовая модель синтеза мегасети из набора непересекающихся подграфов с согласованностью в виде нормы H_2; [2, 8] - оптимизационная задача выбора мостов подграфов мегасети с построением соединительных ребер; [4, 9, 10] - алгоритм кластеризации данных групп сенсорных узлов мегасети с использованием нечеткой логики для учета гетерогенных параметров сенсорной сети и гетерогенного управления сетью; [12] - архитектура системы разработки потоковых приложений для обработки данных на основе моделей потоков данных с интеграцией в среду IDE поверх UML; [3, 5] - информационное и программное обеспечение для экспериментальной оценки качества разработанных методов и алгоритмов.

Структура и объем работы. Диссертационная работа состоит из введения, четырех глав, заключения, списка литературы из 159 наименований. Работа изложена на 150 страницах.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во введении обоснована актуальность исследования, сформулированы его цель и задачи, научная новизна и практическая значимость полученных результатов, приведены сведения об апробации и внедрении работы.

В первой главе исследуются особенности обработки данных в мегасетях на основе графовых моделей и системы разработки потоковых приложений для обработки данных с использованием специальной метрики согласованности и интеграции логики приложений. Отмечено, что повысить эффективность управления облачными средами можно путем разработки графовой модели синтеза мегасети из набора непересекающихся подграфов с согласованностью в виде нормы H_2, постановки оптимизационная задача выбора мостов подграфов мегасети с построением соединительных ребер, разработки алгоритма кластеризации данных групп сенсорных узлов мегасети и создания архитектуры системы разработки потоковых приложений для обработки данных на основе моделей потоков данных.

Исследована проблема оптимизации согласованности сети сетей (КоС) с произвольным знаком весов ребер. Решается задача выбора узлов-мостов между всеми подсистемами КоС, узлы-мосты связываются межсистемными ребрами. Проблема оптимизации решается с учетом связи согласованности и эффективного сопротивления. Минимизация осуществляется путем поиска узлов-мостов в графах ограниченного сопротивления. Эти узлы-мосты могут быть идентифицированы независимо от других подграфов и соединительной топологии. Доказано, что для КоС с древовидным опорным графом решение является оптимальным даже в рамках более общей модели, допускающей наличие нескольких узлов моста на подграф. Представлены границы согласованности, а также аналитические примеры оптимальных КоС. Представлены численные результаты, иллюстрирующие производительность различных топологий КоС.

Далее необходимо провести исследование и разработку алгоритма кластеризации данных групп сенсорных узлов мегасети, отличающегося использованием нечеткой логики для учета гетерогенных параметров сенсорной сети и гетерогенного управления сетью, обеспечивающего инвариантность к большому объему сетевых данных и динамичности сетевого местоположения узлов.

Результат анализа потребовал формализации данных задач, а также алгоритмизации их решения с учетом особенностей, отраженных на рис. 1. Сформулирована цель и задачи исследования.

Вторая глава посвящена разработке графовой модели синтеза мегасети из набора непересекающихся подграфов с согласованностью в качестве нормы H_2, обеспечивающую соединение подграфов с обеспечения оптимальной согласованности финальной мегасети.

Проведен анализ проблематики оптимального проектирования сети в объединении сетей, графе, состоящем из набора непересекающихся подграфов и набора добавленных ребер между ними. Узлы подчиняются зашумлен-

ной согласовательной динамике, разрешены ребра с весами на всей числовой оси. Вводится понятие согласованности в качестве нормы H_2. Суть нормы состоит в том, что она представляет собой оценку расстояния до консенсуса, которое рассчитывается в виде дисперсии.

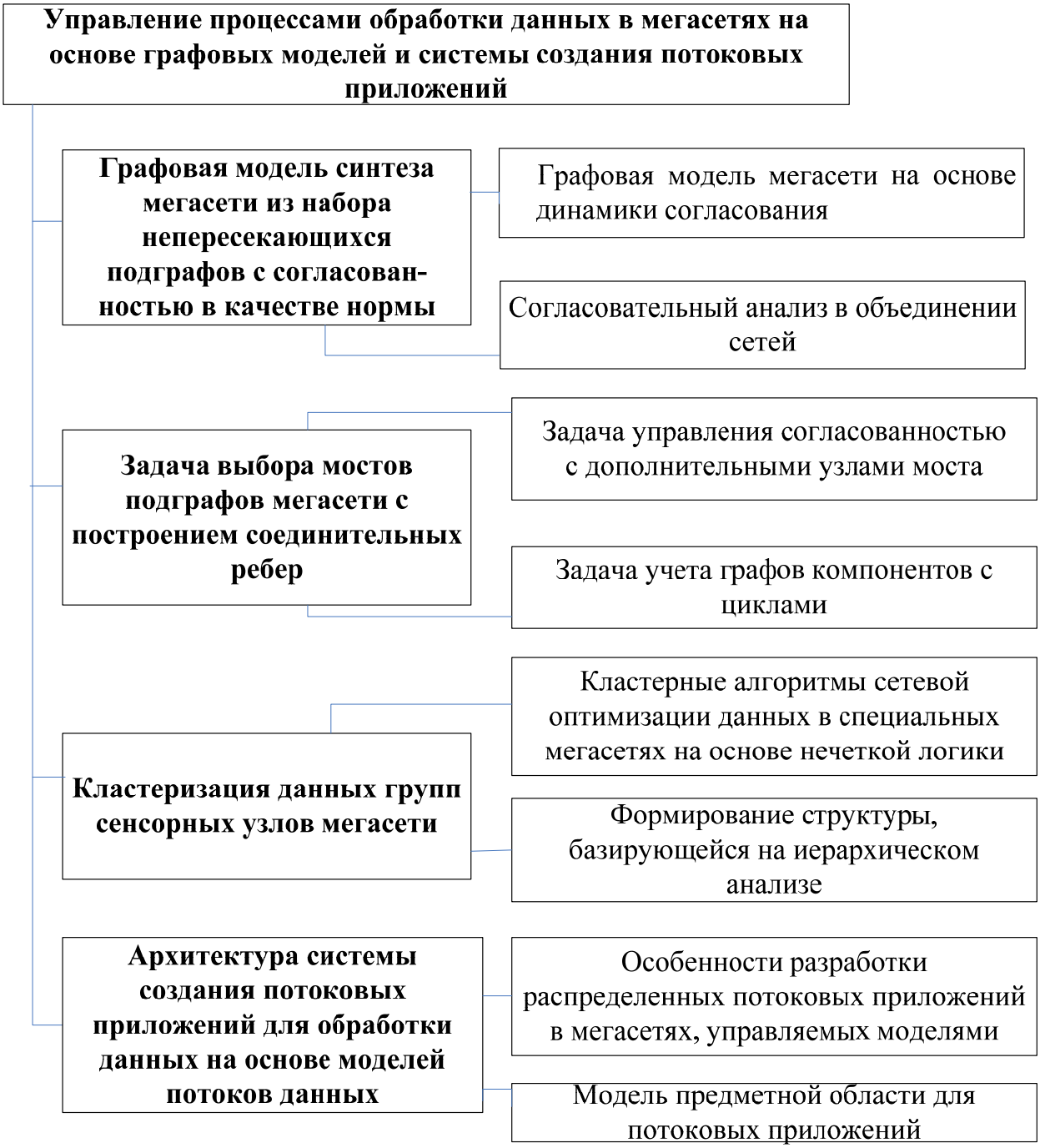

Рис. 1. Дизайн исследования

Ставится задача о том, как соединить подграфы, выбрав один соединительный узел в каждом подграфе, чтобы полученная объединенная сеть имела оптимальную согласованность. Затем показано, что эту проблему можно решить, идентифицируя соединительный узел в каждом подграфе независимо от других узлов и подграфов. Таким образом, задача может быть решена за полиномиальное по размеру наибольшего подграфа время. Если даже в подграфе существует не единственный соединительный узел, то полученное

для древовидной топологии решение также оптимально.

Исследуется мегасеть как множество непересекающихся подсетей $\{G_1,...,G_p\}$. Каждая подсеть $G_i=(V_i,E_i,w_i)$ представляет собой связный неориентированный граф, где n_i - количество узлов, V_i - набор узлов, E_i - набор ребер, $w_i: E_i \to R$; $w_i \in \{-,+\}$ - функция знака веса ребер в E_i.

Примем единственность узла-моста $l_i \in V_i$ в любой подсети G_i. КоС проектируется путем интеграции подсетей через поиск узлов моста и соединение их между собой посредством ребер. Цель – соединить ребрами узлы моста $l_i \in V_i$ с узлами $l_j \in V_j$ подсети G_j. На рис. 1 показан пример пары узлов моста, соединяющих два подграфа для формирования Сети сетей (КоС или NoN в англоязычной нотации). Магистральный граф $B=(V_B, E_B, w_B)$ - это граф, определяемый узлами моста, $V_B=\{l_1,. . . ,l_p\}$ - ребра между ними, $E_B \subseteq \{(l_i,l_j) \mid l_i,l_j \in V_B\}$ и весовой функции $w_B:E_B \to R$. Граф B может содержать как положительные, так и отрицательные ребра. Считаем, что магистральный граф B в КоС G есть взвешенный связный неориентированным граф.

Объединенная сеть формально определяется как

$$G=(V,E,\omega),\ |V| = N = \sum_{i=1}^{p} n_i ,$$

где $V = V_1 \cup ... \cup V_p$; $E = E_1 \cup ... \cup E_p \cup E_B$ б $w(j,k) = \begin{cases} w_i(j,k), (j,k) \in E_i \\ w_B(j,k), (j,k) \in E_B \\ 0, \text{иначе} \end{cases}$.

Для компактности далее будем записывать w(j,k) как w_{jk}.

Каждый узел $j \in V$ имеет скалярное состояние x_j:

$$\dot{x}_j(t)=u_j(t)+\nu_j(t) \qquad (1)$$

где u_j - управляющий вход, а ν_j - белый шум с нулевым средним значением и единичной дисперсией.

Рассмотрим динамику согласования, когда каждый узел обновляет свое состояние на основе относительных состояний своих соседей в КоС. Управляющий вход задается так:

$$u_j(t) = -\sum_{k \in N_i} w_{jk} \left(x_j(t) - x_k(t) \right) \qquad (2)$$

где N_j - множество соседей узла j. Пусть x будет вектором состояний узла. Динамику сети G можно записать как

$$\dot{x}(t)=-Lx(t)+\nu(t) \qquad (3)$$

где ν - вектор возмущений, L - лапласиан группы G:

$$L_{jk} = \begin{cases} -w_{jk}, (j,k) \in E \\ \sum_{i \in N_j} w_{ji} \\ 0, \text{иначе} \end{cases}$$

Оценим количественно производительность сети по ее согласованно-

сти, которая определяется следующим образом:

$$H_C(G) = \lim_{t\to\infty} \sum_{j=1}^{N} \mathrm{var}\left(x_j(t) - \frac{1}{N} \sum_{k=1}^{N} x_k(t) \right)$$

Согласованность представляет собой общую дисперсию отклонений от среднего текущего состояния узла. Эта величина ограничена при условии, что КоС является связной. Малая величина дисперсии свидетельствует о сильной согласованности мегасети. Если дисперсия велика, то это указывает на значительную «энтропию» в системе.

Определим согласованность как следующее выражение с использованием нормы H_2: $H_2(G) = \mathrm{tr}\left(\int_0^\infty e^{-Lt} P e^{-Lt} dt \right)$, где $P = I - \frac{1}{N} \mathbb{Z}\mathbb{Z}^T$, $\mathbb{Z}$ - единичный вектор. Для псевдообратной матрицы L^+ к положительно полуопределенной матрице L с нулевым простым собственным значением имеет место $H_2(G) = \frac{1}{2} \mathrm{trace}\left(L^+ \right)$.

Для любого связного графа величина $H_2(G)$ конечна. Эта величина зависит от топологии мегасети. Если минимизировать $H_2(G)$, то в полученной КоС будет минимальная «сетевая энтропия» и наилучшая согласованность. Рассмотрим эту оптимизационную задачу в ситуации, когда множество подграфов задано. Пусть топология магистрального графа определена априори, т.е. заранее определено, какие подграфы должны быть связаны с какими другими подграфами. Задача состоит в поиске для каждой подсети (подграфе) единственного узла-моста, через который будет происходить взаимодействие с другими подсистемами. Оптимизационная задача сводится к оптимизации согласованности финальной мегасети. Представим лапласиан $L(l_1,...,l_p)$ финальной мегасети как функцию от набора оптимальных узлов-мостов. Таким образом, запишем эту матрицу Лапласа как $L(l_1,...,l_p)$. Задача оптимизации выбора узлов моста может быть формализована следующим образом:

$$\min_{l_i \in V_i,\, i=1,...,p} H_2(G) = \frac{1}{2} \mathrm{trace}\left(L\left(l_1,...l_p \right)^+ \right) \qquad (4)$$

Таким образом, в работе предложена графовая модель синтеза мегасети из набора непересекающихся подграфов с согласованностью в виде нормы H_2, отличающаяся непрерывностью весов ребер на всей числовой оси и зашумленной динамикой согласования узлов, реализующая соединение подграфов с обеспечением оптимальной согласованности финальной мегасети.

Отдельно рассмотрены интегрированные сети с весами ребер на положительной полуоси. Для таких сетей установлены оценки минимальной и максимальной согласованности. Решения даны в аналитическом виде и проиллюстрированы примерами. Последние представляют расширение изучения величины нормы H_2 для интегрированных сетей.

Существует взаимозависимость между согласованностью и РПС

(Resistance distance – расстояние по сопротивлению) в электротехнике. Применим эту взаимозависимость при решении задачи (4).

Пусть $G_g=(V_g,E_g,w_g)$ - произвольный взвешенный граф, $|V_g|=n$, E_g^+ - подмножество ребер E_g с положительными весами, E_g^- - подмножество ребер с отрицательными весами, $G_g^+=(V_g,\ E_g^+,w_g^+)$, $G_g^-=(V_g,\ E_g^-,w_g^-)$, $G_g^+\cup G_g^-=G_g$, w_g^+, w_g^- - веса для редукции сужением w_g на ребра с соответствующим знаком веса, L_g - взвешенная матрица Лапласа группы G_g.

Для ненаправленных графов G_g с положительными весами ребер исследуем электрическую сеть с топологией G, $1/w_g(u,v)$ - сопротивление любого ребра. Потенциальная близость $r(u,v)$ для ребер u и v при проходящем токе в 1 А ($u, v\in V_g$) определяется через L_g^+:

$$r(u,v)=\left(L_g^+\right)_{uu}-\left(L_g^+\right)_{vv}+2\left(L_g^+\right)_{uv} \qquad (6)$$

где $\left(L_g^+\right)_{ij}$ - (i,j)-й элемент L_g^+.

Тогда полное эффективное сопротивление (ЭфС) определяется как

$$\Omega_{G_g}=\frac{1}{2}\sum_{u,v\in V_g} r(u,v)=\frac{1}{2}\sum_{u<v\in V_g} r(u,u),\ \text{или}\ \Omega_{G_g}=n\sum_{u\in V_g}\left(L_g^+\right)_{uu}-\mathbb{Z}^T L_g^+\mathbb{Z}=n\bullet\mathrm{trace}\left(L_g^+\right).$$

Концепция ЭфС распространена на сети с отрицательными весами ребер. Назовем графы с неположительными весами, в которых РПС всегда конечны, графами ограниченного сопротивления (ОГС).

Определение 1. $G_g=(V_g,E_g,w_g)$ будем называть ограниченным графом сопротивлений (ОГС), если: G_g - связен; $\forall(u,v)\in E_g^-$ ребро в составе цикла; в G_g нет циклов с более чем одним $(u,v)\in E_g^-$; $\forall(u,v)\in E_g^-:\ \left|w_g(u,v)\right|<\dfrac{1}{r_{G_g^+}(u,v)}$,

$r_{G_g^+}(u,v)$ РПС для u и v в G_g^+.

Пример ОГС: граф с положительными весами ребер. Доказано

Предположение 2. Пусть граф G_g - ОГС. Пусть можно разделить на подграфы A и B с положительными весами ребер. При этом существует единственная вершина x, принадлежащая подграфам A и B. РПС между любыми $a\in A$ и $b\in B$ вычисляется как:

$$r(a,b)=r(a,x)+r(x,b). \qquad (7)$$

Определение 2. Рассмотрим граф. Индекс сопротивления узла $v\in V$ графа $G=(V,E,w)$ определим как

$$C(v)=\sum_{u\in V,\ u\neq v} r(u,v) \qquad (8)$$

Обозначим индекс сопротивления для узлов в подграфе G_i как $C_i(v_i)$

(узел v_i подграфа G_i). Доказаны следующие утверждения:

Утверждение 1. Пусть граф КоС G из p подграфов $G_i=(V_i,E_i,w_i)$, i=1,...,p, сформирован с использованием заданной топологии опорного графа с одним мостовым узлом $l_i \in V_i$ на подграф. Предположим, что опорный граф и все его подграфы являются ОГС. Тогда условие $l_i = \underset{v \in V_i}{\operatorname{argmin}} C_i(v) \ \forall G_i$ достаточно для минимальности согласованности G. Доказаны

Утверждение 2. Пусть G=(V,E,w) - КоС как ОГС. Для узлов $u,v \in V$, $u \in V_i$, $v \in V_j$, $i \neq j$, РПС равно $r(u,v)= r(u,l_i)+r(l_i, l_j)+ r(l_j,v)$.

Утверждение 3. Пусть G=(V,E,w) - КоС как ОГС. Тогда выражение (13) определяет согласованность G:

$$H_2(G) = \frac{1}{2N}\left(\sum_{i=1}^{p} 2n_i H_2(G_i) + \sum_{i=1}^{p}\sum_{j=i+1}^{p} |V_i||V_j| r(l_i,l_j) + \sum_{i=1}^{p} |V - V_i| C_i(l_i) \right) \qquad (9)$$

Следствие 1. Для топологий КоС G с единичными весами ребер, $|V_i|=n$; $i=1,...p$, $\min H_2(G)$ дается формулой:

$$H_2(G) \geq \frac{1}{2N}(2n^2(p-1) + p(n-1) + 2p(p-1)(n-1)) \qquad (10)$$

Следствие 2. Верхняя граница согласованности любого КоС G со всеми весами ребер, равными 1, состоящего из p подграфов с $|V_i| =n$ для i=1,...,p:

$$H_2(G) \leq \frac{1}{2N}(2n^2(p-1) + p(n-1) + 2p(p-1)(n-1))$$

Чтобы сравнить КоС со связными и несвязными опорными графами, введем концепцию компонента графа. Пусть G - КоС, образованный из подграфов $G_1,...,G_p$, с опорным графом B. Компонентный граф графа G определяется как $C=(V_C,E_C)$, где $V_C=\{1,...,p\}$ с каждым узлом, соответствующим подграфу в G, и $E_C=\{(i_1, j_1)...(i_R, j_R)\}$ - множество компонентов рёбер графа с $|E_C|=R$. Ребро существует тогда и только тогда, когда $(u,v) \in E_B$, где $u \in V_{i_r}$ и $v \in V_{j_r}$. Существует ребро $(u,v) \in E_B$, где $u \in V_{i_r}$ и $v \in V_{j_r}$.

Исследовано влияние не одного, а нескольких узлов моста в одном подграфе на согласованность КоС. Рассмотрен вопрос о том, улучшит ли согласованность использование нескольких узлов моста на подграф.

На рис. 2 представлены два КоС на базе одного графа компонентов. Рис. 2 слева - КоС G_A и G_B, из одних и тех же трех подграфов. Магистральный граф для G обозначен штриховыми линиями, а магистральный граф для G_B обозначен пунктирными линиями; рис. 2 справа - граф компонентов для G_A и G_B.

Рассмотрим модифицированный вариант задачи (4), в котором задан граф компонентов, т.е. известно, какие подграфы должны быть связаны друг с другом. Цель состоит в том, чтобы определить, какой узел (узлы) моста в каждом подграфе следует использовать для соединения подграфов, чтобы оптимизировать согласованность G. Необходимо, чтобы каждое ребро в графе компонентов соответствовало ровно одному ребру в КоС. В этом случае

матрица Лапласа КоС G, которую необходимо построить, может быть определена как функция узлов моста, соответствующих R ребрам в опорном графе, т.е. $L = L\left(\left(l_i^l, l_j^l\right),...,\left(l_i^R, l_j^R\right)\right)$.

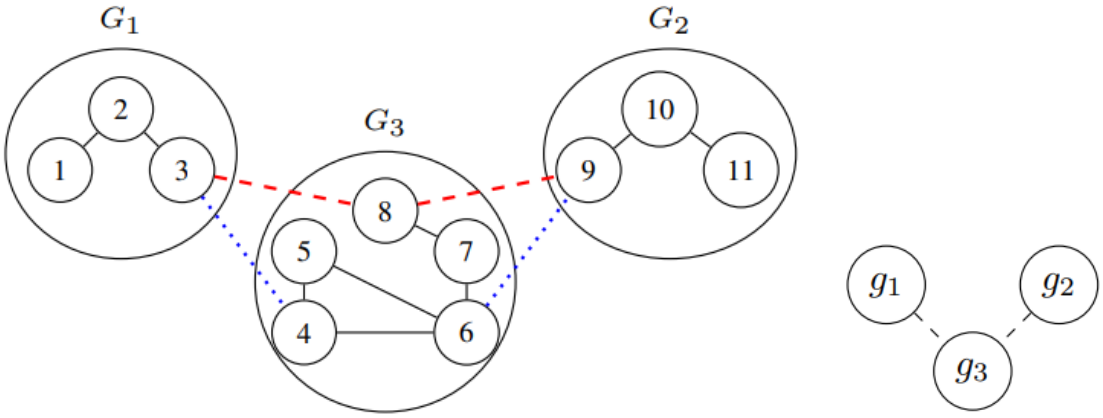

Рис. 2. Различные варианты формирования КоС

Оптимизируемое выражение в случае нескольких узлов моста в каждом подграфе:

$$H_C(G_m) = \frac{1}{2N}\left(\sum_{i=1}^{p} 2n_i H_C(G_i) + \sum_{i=1}^{p}\sum_{j=i+1}^{p}\left(\left|V_j\right| C_i\left(l_{i\to j}\right) + \right.\right.$$
$$\left.\left. + \left|V_i\right|\left|V_j\right| r(l_{i\to j}, l_{j\to i}) + \left|V_i\right| C_j\left(l_{j\to i}\right)\right)\right) \qquad (11)$$

где G_m - КоС с опорным графом. Показано, что для древовидной структуры нахождение единственного оптимального узла-моста для каждого подграфа дает оптимальную производительность.

Некомбинаторное решение оптимизационной задачи основано на распараллеливании на множество задач нахождения оптимального узла-моста.

Таким образом, осуществлена постановка оптимизационнойзадачи выбора мостов подграфов мегасети с построением соединительных ребер, отличающаяся параллелизмом процесса решения и обеспечивающая получение оценок минимальной и максимальной согласованности с весами ребер на положительной полуоси.

Третья глава посвящена разработке алгоритма кластеризации данных групп сенсорных узлов мегасети, обеспечивающего инвариантность к большому объему сетевых данных и динамичности сетевого местоположения узлов.

Оптимизация мегасети играет ключевую роль в планировании и проектировании сети. Хорошо спроектированная сеть может улучшить эффективность сети, и таким образом лучше выполнять агрегацию. Кластеризация представляет собой методику, направленную на группирование сенсорных узлов датчиков в несколько небольших групп - узлы кластера. Эти узлы кластера отвечают за агрегацию данных от отдельных узлов и их маршрутизацию в желаемое место. Основные параметры кластеризации в отношении эффективности алгоритма: количество кластеров; внутрикластерное взаимо-

действие; мобильность узлов; типы узлов в части энергопотребления; выбор главного кластера; наложение: (разделение сенсорной сети на многочисленные перекрывающиеся кластеры с определенной средней точностью перекрытия).

Однако некоторые сетевые атрибуты не могут быть непосредственно измерены точно. С учетом многомерных данных классические схемы кластеризации могут работать неэффективно. Необходимо учитывать как огромный объем сетевых данных, так и динамичность сетевого местоположения. В укрупненном виде алгоритм кластеризации таков:

- кластеризуемая структура создается на основе иерархического анализа учитываемых параметров;
- осуществляется определение лингвистических переменных для оценки кластеров сенсорных узлов, которые затем преобразуются в трапециевидные нечеткие числа, чтобы сделать их простыми и точными для дальнейшего рассмотрения;
- применяется методика кластеризации сенсорного узла для многомерной сети на основе нечеткой логики.

Укрупненный алгоритм создания структуры сети приведен на рис. 3.

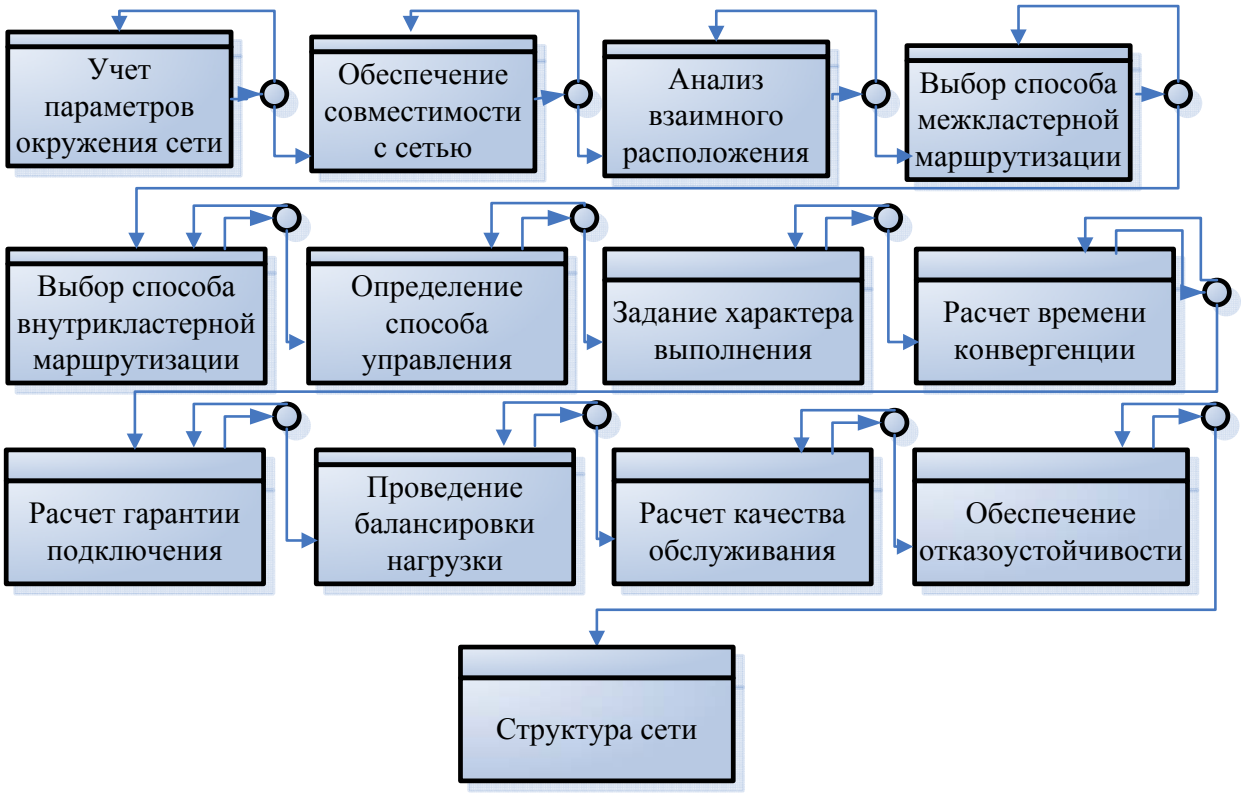

Рис. 3. Укрупненный алгоритм создания кластеризуемой структуры

Определение лингвистических переменных и их преобразование проводится с применением нечеткой логики для преобразования данных в трапециевидные нечеткие числа $\Theta=(a, b, c, d)$. Таким образом, функция принадлежности может быть рассчитана с использованием трапециевидного нечеткого числа:

$$f(x)=\begin{cases}0, x \le a \\ \dfrac{x-a}{b-a}, a \le x < b \\ 1, b \le x < c \\ \dfrac{d-x}{d-c}, c \le x < d \\ 0, x \ge d\end{cases} \quad (12)$$

где параметры a, b, c, d - действительные числа. Из-за произвольной природы проблемы кластеризации в крупномасштабной многомерной сенсорной сети параметры используются в качестве лингвистических переменных.

В табл. 1 приведены их численные характеристики, выбранные на основе тематических исследований ученых.

Таблица 1

Лингвистические термины	Трапециевидное число (параметры)
Идеально	1, 0.98, 0.95, 0.92
Очень хорошо	0.95, 0.92, 0.86, 0.82
Хорошо	0.85, 0.75, 0.67, 0.63
Удовлетворительно	0.63, 0.60, 0.58, 0.55
Плохо	0.58, 0.55, 0.52, 0.49
Очень плохо	0.52, 0.45, 0.42, 0.39
Крайне плохо	0.40, 0.37, 0.33, 0.30

Точность кластеризации может быть выражена так:

$$CP_{\delta}=\frac{cX(c-1)\sum_{p}^{c}\sum_{x\in P_p}\sum_{m\subseteq P_p}\left(\rho_m(x)-o_p^m\right)^2}{2xp_e x\sum_{p}^{c}\sum_{x\in P_p}\sum_{m\subseteq P_p}\left(o_p^m-o_k^m\right)^2} \quad (13)$$

Эвристический алгоритм состоит из следующих шагов.

1. Для каждого узла проводится расчет степени ассоциации, дезагрегированная от подкритериев к основным критериям.

2. Расчет функции энтропии ассоциации и функции коэффициента ассоциации для выборочного набора узлов в сети.

3. Сопоставление подкритерия оценки с основным критерием.

4. Выражение оценки для подкритерия и оценочного значения для основного критерия в виде трапециевидных чисел.

5. Определение степени ассоциации узла.

6. Расчет индекса оценки, как отношение энтропийной функции ассоциации и функции коэффициента ассоциации по отношению к результату.

7. (цикл) Для каждой пары атрибутов сравниваем индекс оценки и исключаем атрибут с наибольшим индексом.

8. На основе нечеткой матрицы отношений формируем выборки, которые сгруппированы в один или несколько кластеров.

9. Если диагональные значения нечеткой матрицы отношений не равны, переход на шаг 7, иначе начальный кластер сформирован.

10. Для каждого начального кластера вычисляется взвешенная степень ассоциации (4) и точность кластеризации (13).

Чем меньше значение точности кластеризации, тем выше будет эффективность кластеризации сети.

Сравнение известных алгоритмов с предложенным приведено в табл. 2.

Таблица 2

Алгоритмы кластеризации	**Количество кластеров**	**Точность кластеризации**	**Время алгоритма**	**Возможность обработки больших данных**
UCLF	6	3.25	$O(N^2)$	Да
Иерархическая агломеративная кластеризация, основанная на доверии	4	3.73	O(N)	Да
AVC	7	2.83	$O(N^2)$	Да
CLARA	Нет	Нет	$(O(K(40+K^2)+K(N-K))^+$	Нет
DBSCAN	Нет	Нет	O(N log N)	Нет
BIRCH	Нет	Нет	O(N)	Да
K-means подход к кластеризации	Зависит от подхода и количества точек данных	Нет	O(NKd)	Нет
Нечеткий c-means	Зависит от подхода и количества точек данных	Нет	O(N)	Нет
Иерархическая кластеризация	Зависит от подхода и количества точек данных	Нет	$O(N^2)$	Нет
ROCK	21	Нет	O(N log N)	Да
Предложенный подход	4	2.41 (наименьшая)	O(N)	Да

Таким образом, в работе предложен эвристический алгоритм кластеризации данных групп сенсорных узлов мегасети, отличающийся использованием нечеткой логики для учета гетерогенных параметров сенсорной сети и гетерогенного управления сетью, обеспечивающий инвариантность к большому объему сетевых данных и динамичности сетевого местоположения узлов.

В главе 4 представлена архитектура системы разработки потоковых приложений для обработки данных на основе моделей потоков данных, обеспечивающая интеграцию в среду IDE поверх UML.

Рассмотрена архитектура системы, управляемой моделями, которая опирается на концептуальные сходства между различными потоковыми

платформами для упрощения и ускорения проектирования, разработки и эксплуатации распределенных потоковых приложений, одновременно преодолевая проблему привязки запуска к конкретному исполнению движка. Система основана на предметно-ориентированный язык моделирования в форме профиля UML, который позволяет разработчикам моделировать потоковое приложение в виде графа потока данных, используя стандартное моделирование UML и, в частности, составные структурные диаграммы.

На рис. 4 представлен обзор рабочего процесса с точки зрения разработчика приложения.

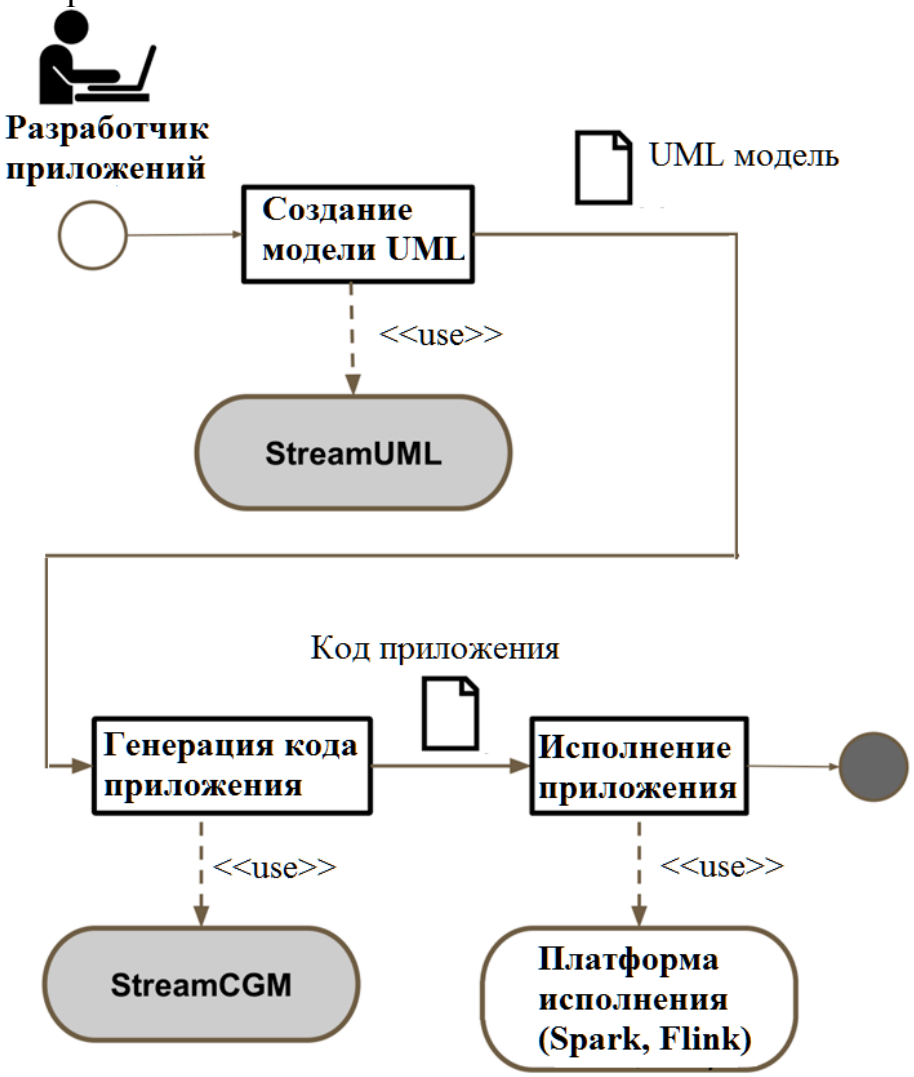

Рис. 4. Обзор рабочего процесса использования системы разработки

Белые и серые круги обозначают начало и конец рабочего процесса соответственно. Квадратные прямоугольники обозначают действия, а сплошные стрелки соединяют действия для создания рабочего процесса. Символы документа, обозначенные сплошными стрелками, представляют собой артефакты, которые создаются в конце определенного действия и передаются в качестве входных данных для следующего. Пунктирные стрелки связывают действия с инструментами, используемыми разработчиком приложения для их выполнения. Инструменты представлены в виде округлых рамок (те из

них, которые являются оригинальными результатами данного исследования, отмечены серым цветом).

Основная часть модели предметной области показана на рис. 5.

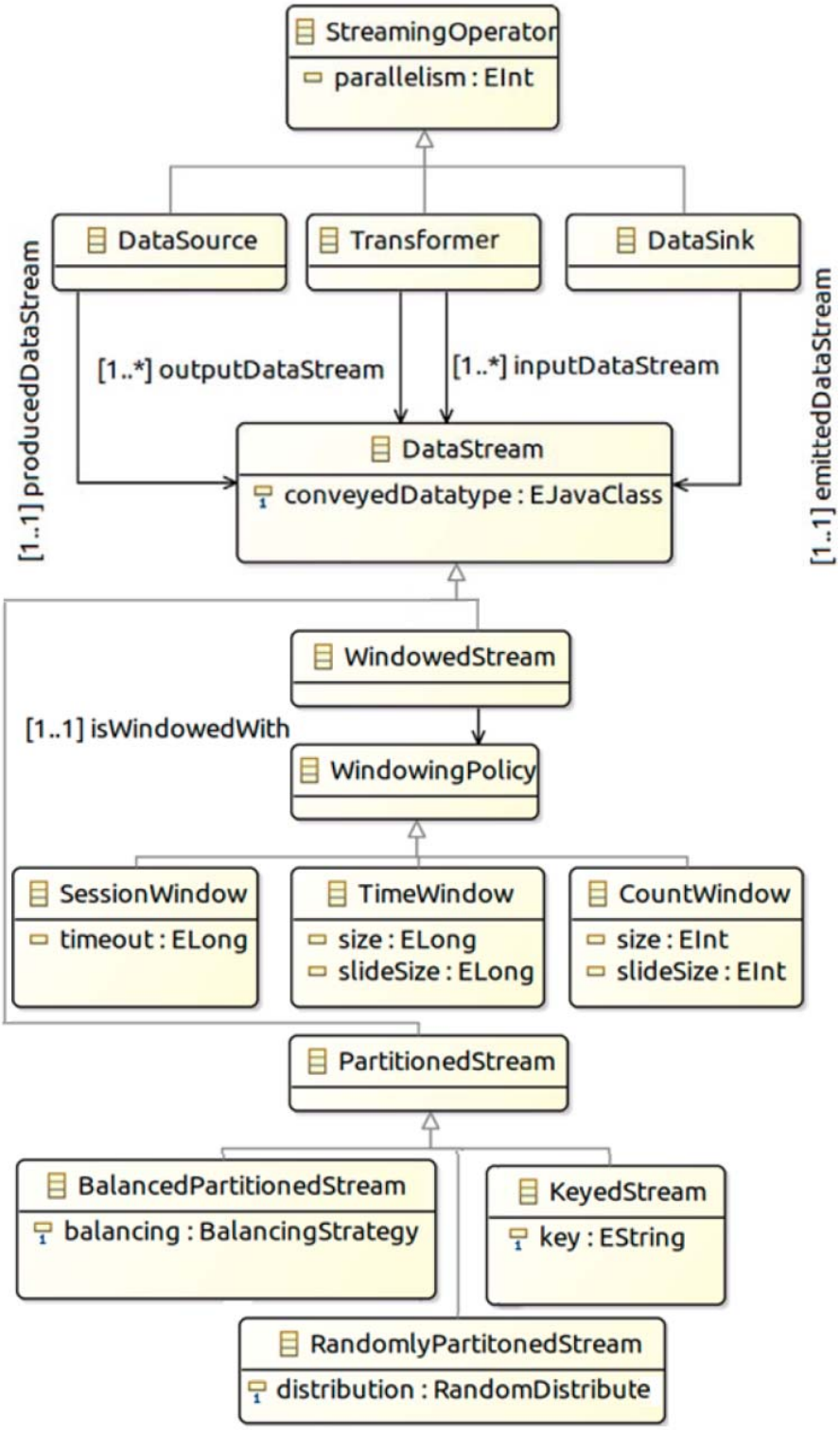

Рис. 5. Модель предметной области для потоковых приложений

Язык предоставляет основные абстракции, необходимые для моделирования распределенного потокового приложения, тем самым освобождая разработчиков от необходимости знать подробности о конкретных целевых платформах, но при этом позволяя варьировать настройки для конкретной платформы.

Основная идея заключается в преобразовании приложений в топологии или прямые ациклические графы (DAG), в которых операторы являются компонентами, которые можно развертывать по отдельности, и обеспечивают механизмы отказоустойчивости. Кроме того, операторы обычно распараллеливаются, что означает наличие нескольких запущенных экземпляров оператора, каждый из которых можно рассматривать как отдельный процесс, который потенциально может выполняться в своей собственной среде.

Согласно этим стратегиям развертывания, одно потоковое приложение на самом деле представляет собой совокупность независимых и индивидуально развертываемых компонентов. Платформа распределенной потоковой передачи отвечает за принятие решений о том, как распределять параллельные вычисления, с конкретными указаниями, в соответствии с которыми должны быть разделены входные данные, в то время как планировщик затем приступает к планированию заданий и обеспечению их выполнения.

Разработчик использует UML-моделирование и, в частности, составные структурные диаграммы, обогащенные профилем StreamUML (реализованным в Eclipse Papyrus), для создания модели своего потокового приложения независимо от платформы, как с точки зрения графика вычислений потока данных, так и с точки зрения более продвинутых аспектов, таких как распараллеливание и временная семантика. Во время этого процесса моделирования разработчик поддерживается механизмами проверки, определенными в StreamUML, с помощью ограничений OCL. Затем разработчик запускает генерацию кода с помощью StreamCGM (реализованного в Eclipse Acceleo), который принимает в качестве входных данных созданную UML-модель и создает код приложения. В завершение можно выполнить сгенерированный код, используя выбранную целевую платформу выполнения (Spark или Flink).

Определено использование диаграмм классов UML и их пакета управления информационным потоком в сочетании с составными структурными диаграммами как оптимальное решение, поскольку это позволяет:

1) точно моделировать потоковые приложения в соответствии с семантикой UML;

2) сохранять сложность подхода к моделированию приемлемой.

При таком подходе диаграммы классов используются для определения компонентов приложения и обмена информацией на более высоком уровне абстракции, а диаграммы составных структур используются для представления модели выполнения приложения, при этом экземпляры ранее определенных компонентов и передачи данных моделируются более подробно.

В табл. 3 представлены наиболее релевантные стереотипы, которые отражают классы моделей предметной области. Для каждого стереотипа описаны его тип, обобщенные стереотипы и расширенные метаклассы UML.

Генерация кода состоит из трех этапов, интегрированных в среду IDE:

1) **Генерация типов данных**: на этом этапе генератор кода создает новый пакет, который заполняется необходимыми классами для всех типов данных, используемых в модели. В частности, при генерации выполняется поиск пакетов UML, помеченных стереотипом «StreamDatatypes», и для каж-

дого вложенного типа данных создается соответствующий простой старый объект Java (POJO). Этот этап не зависит от выбранной целевой платформы.

Таблица 3

Некоторые стереотипы, представленные профилем UML

Стереотип	**Тип расширения**	**Стереотип расширения**	**Метакласс**
«DistributedStreamingApplication»	Абстрактный	None	UML::StructuredClassifier
«StreamingOperator»	Абстрактный	None	UML::StructuredClassifier
«DataStream»	Абстрактный	None	UML::Connector
«StreamDataTypes»	Конкретный	None	UML::Package
«FlinkApplication»	Конкретный	«DistributedStreamingApplication»	Нет
«SparkApplication»	Конкретный	«DistributedStreamingApplication»	Нет
«DataSource»	Абстрактный	«StreamingOperator»	Нет
«DataSink»	Абстрактный	«StreamingOperator»	Нет
«Transformer»	Абстрактный	«StreamingOperator»	Нет
«PartitionedStream»	Абстрактный	«DataStream»	Нет
«WindowedStream»	Конкретный	«DataStream»	Нет
«KeyedStream»	Конкретный	«PartitionedStream»	Нет
«RandomlyPartitionedStream»	Конкретный	«PartitionedStream»	Нет
«BroadcastedStream»	Конкретный	«PartitionedStream»	Нет
«NonParallelStream»	Конкретный	«PartitionedStream»	Нет
«KafkaSource»	Конкретный	«DataSource»	Нет
«SocketSource»	Конкретный	«DataSource»	Нет
«TextFileSink»	Конкретный	«DataSink»	Нет
«SocketSink»	Конкретный	«DataSink»	Нет
«MapTransformer»	Конкретный	«Transformer»	Нет
«FlatmapTransformer»	Конкретный	«Transformer»	Нет
«SumTransformer»	Конкретный	«Transformer»	Нет
«CountTransformer»	Конкретный	«Transformer»	Нет
«ReduceTransformer»	Конкретный	«Transformer»	Нет
«FilterTransformer»	Конкретный	«Transformer»	Нет
«JoinTransformer»	Конкретный	«Transformer»	Нет
«FlattenTransformer»	Конкретный	«Transformer»	Нет
«NMapTransformer»	Конкретный	«Transformer»	Нет
«NFlatmapTransformer»	Конкретный	«Transformer»	Нет
«WindowTransformer»	Конкретный	«Transformer»	Нет

2) **Генерация потоковых функций**: на этом этапе генератор кода создает новый пакет, который заполняется необходимыми классами, реализующими различные пользовательские операторы, присутствующие в модели. В частности, генератор кода ищет мета-типизированные объекты «StructuredClassifier», помеченные стереотипом, который требует от пользователя указать логику преобразования (например, «MapTransformer», «WindowTransformer» и т.д.) и для каждого из них создает новый класс Java, который реа-

лизует определенные интерфейсы, предоставляемые выбранной целевой платформой.

3) **Генерация потокового задания**: на этом этапе генератор кода создает новый пакет, содержащий фактическое приложение, которое должно быть выполнено, т.е. класс Java, который создает и выполняет граф потока данных, определенный в модели. В этом классе различные функции, созданные на втором этапе (например, LineSplitter, WordCounter), создаются и подключаются с использованием потоков данных (например, текста, токенов, подсчетов), которые передают кортежи различных типов данных, определенных на первом этапе (например, WordToken, WordCount).

Разработана структура программного прототипа системы разработки потоковых приложений для обработки данных на основе моделей потоков данных (рис. 6). Элементы программной реализации прошли государственную регистрацию в ФИПС.

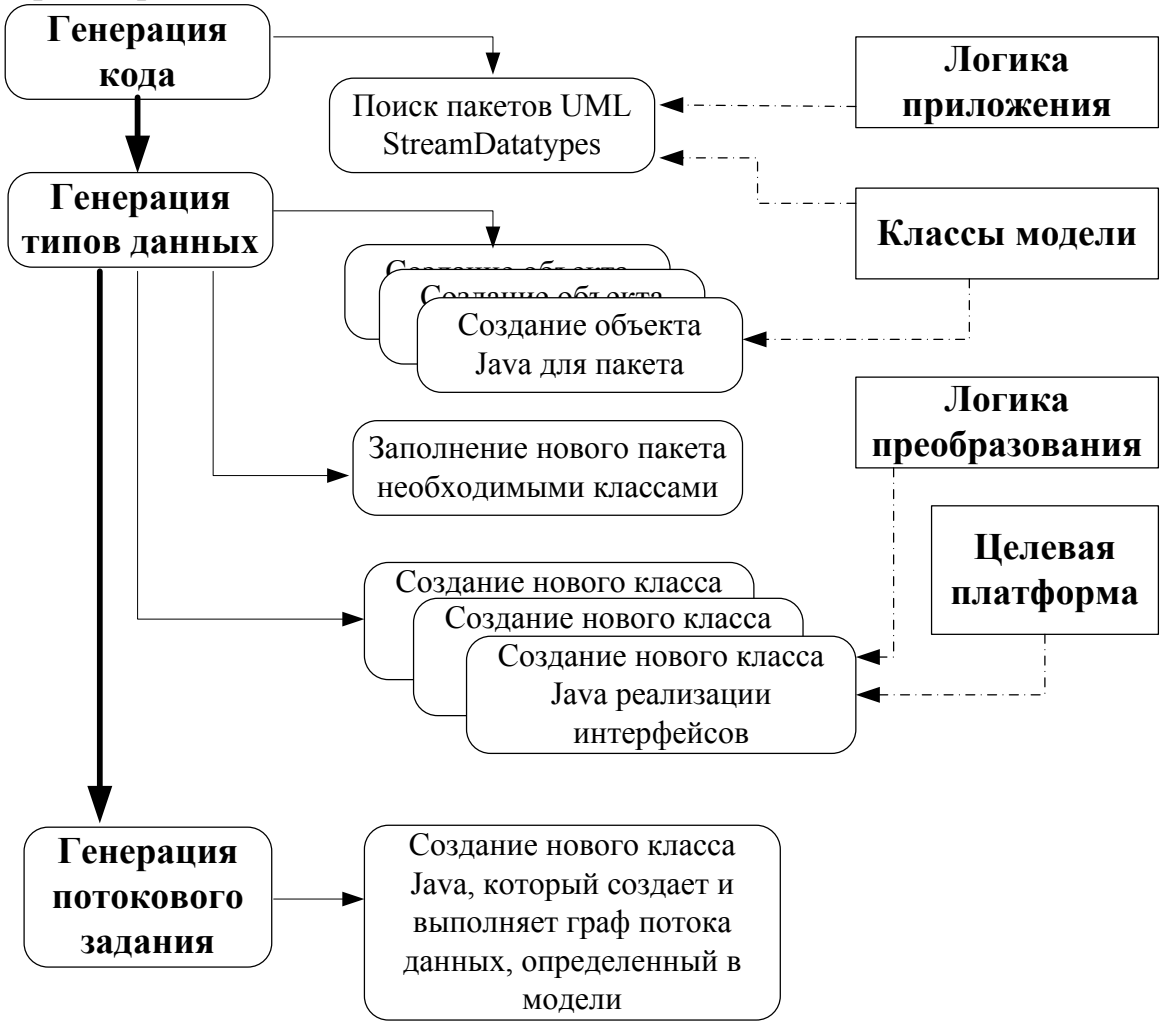

Рис. 6. Структура программного прототипа системы разработки потоковых приложений для обработки данных на основе моделей потоков данных

Таким образом, создана архитектура системы разработки потоковых приложений для обработки данных на основе моделей потоков данных, отличающаяся интеграцией логики приложения в модель и обеспечивающая интеграцию в среду IDE поверх UML.

Заключение

В процессе выполнения диссертационного исследования получены следующие основные результаты:

1. Предложена графовая модель синтеза мегасети из набора непересекающихся подграфов с согласованностью в виде нормы H_2, учитавающая непрерывность весов ребер на всей числовой оси и зашумленную динамику согласования узлов, реализующая соединение подграфов с обеспечением оптимальной согласованности финальной мегасети.

2. Предложена оптимизационная задача выбора мостов подграфов мегасети с построением соединительных ребер, использующая параллелизмом при решении и обеспечивающая получение оценок минимальной и максимальной согласованности с весами ребер на положительной полуоси.

3. Разработан алгоритм кластеризации данных групп сенсорных узлов мегасети на основе использования нечеткой логики для учета гетерогенных параметров сенсорной сети и гетерогенного управления сетью, обеспечивающий инвариантность к большому объему сетевых данных и динамичности сетевого местоположения узлов. Трудоемкость разработанного алгоритма составляет O(N), точность кластеризации по сравнению с известными прототипами лучше на 4%.

4. Предложена архитектура системы разработки потоковых приложений для обработки данных на основе моделей потоков данных с включением логики приложения в модель, и обеспечивающая интеграцию в среду IDE поверх UML.

5. Создана структура программного прототипа системы разработки потоковых приложений для обработки данных на основе моделей потоков данных. Элементы программного обеспечения зарегистрированы в ФИПС.

Рекомендации и перспективы дальнейшей разработки темы

1. Результаты исследования рекомендуются к применению в задачах проектирования вычислительных систем с разреженной архитектурой на основе потоковых приложений.

2. Дальнейшая разработка темы будет направлена на практическую реализацию теоретических и алгоритмических результатов, интеграцию в проекты наиболее распространенных распределенных систем. Развитие результатов будет направлено на автоматизацию разработки потоковых приложений.

Основные результаты диссертации опубликованы в следующих работах:

Публикации в изданиях списка ВАК

1. Камиль В.А.К., Мутин Д.И., Атласов И.В. Теоретические основы управления согласованностью подсистем при обработке данных в объединении компьютерных сетей// Системы управления и информационные технологии, №4(94), 2023. С. 35-40.

2. Камиль В.А.К., Кочегаров М.В., Мутин Д.И. Аналитическое моделирование многокластерной системы специального назначения на основе нескольких сценариев мониторинга// Моделирование, оптимизация и информа-

ционные технологии. 2024; 12(4). URL: https://moitvivt.ru/ru/journal/pdf?id=1713.

3. Камиль В.А.К., Мутин Д.И., Божко Л.М. Разработка распределенных потоковых приложений в мегасетях на основе UML-моделей// Системы управления и информационные технологии, №4(98), 2024. С. 72-77.

Публикация в издании, индексируемом в WoS

4. Kravets O.Ja., Mutina E.I., Zaslavskaya O.Yu., Redkin Yu.V., Rahman P.A., Aksenov I.A., Kamil Wisam Abduladheem Kamil. Improving the efficiency of using the resources of virtual data centers with a heterogeneous structure by re-positioning virtual machines// International Journal on Information Technologies and Security, vol.17, no.1, 2025. Pp. 3-12. https://ijits-bg.com/2025.v17.i1.01.

Свидетельство о государственной регистрации программы для ЭВМ

5. Программа интерактивного управления очередью системы мониторинга/ Е.В. Сидоренко, М.В. Кочегаров, В.А.К. Камиль, К.А.Ж. Амоа, О.А. Ющенко. Свидетельство о регистрации программы для ЭВМ № 2025615513 от 12.02.2025. М.: ФИПС, 2025.

Статьи и материалы конференций

6. Камиль В.А.К. Некомбинаторное решение задачи оптимизации согласованности для построения сообщества сетей// Информационные технологии моделирования и управления, №4(134), 2023. – С. 290-302.

7. Kamil W.A.K., Mutin D.I. Analytical examples of optimal network design in network integration based on the NoN-model// Modern informatization problems in the technological and telecommunication systems analysis and synthesis (MIP-2024'AS): Proc. of the XXIX-th Int. Open Science Conf. - Yelm, WA, USA: Science Book Publishing House, 2024. Pp. 153-158.

8. Камиль В.А.К., Мутин Д.И. Методы кластеризации сенсорных узлов на основе нечетких множеств// Информационные технологии моделирования и управления, №1(135), 2024. – С. 37-46.

9. Камиль В.А.К., Мутин Д.И., Голиков А.А. Алгоритмизация определения количества обслуживаемых кластеров и количества сенсорных узлов, сопоставленных с каждым кластером// Экономика и менеджмент систем управления, №2(52), 2024. – С. 55-65.

10. Камиль Висам Абдуладим Камиль, Мутин Д.И. Алгоритм кластеризации сенсорных узлов в несколько кластеров с использованием ассоциации нечеткого понятия // Наука и технологии: перспективы развития и применения: сб. статей VII Междунар. НПК. - Петрозаводск: МЦНП «НОВАЯ НАУКА», 2024. - С. 7-13.

11. Камиль Висам Абдуладим Камиль. Задача оптимизации пропускной способности облака SAAS с учетом штрафных санкций// Сб. тр. VI Всеросс. НПК «Информационные технологии в экономике и управлении». – Махачкала, 2024. С. 83-89.

12. Kamil W.A.K., Mutin D.I. Technological features of creating distributed streaming applications in mega-networks based on UML models// Modern informatization problems in the technological and telecommunication systems analysis

and synthesis (MIP-2025'AS): Proc. of the XXX-th Int. Open Science Conf. - Yelm, WA, USA: Science Book Publishing House, 2025. – pp. 163-173.

Подписано в печать 21.03.25
Формат 60x84/16. Бумага для множительных аппаратов.
Усл. печ. л. 1,0. Тираж 80 экз. Заказ №147.
ФГБОУ ВО «Воронежский государственный технический университет»
394026 Воронеж, Московский просп., 14