DOI: 10.25728/mlsd.2023.1403

МНОГОКРИТЕРИАЛЬНОЕ УПРАВЛЕНИЕ ОБРАБОТКОЙ ДАННЫХ В ГЕОГРАФИЧЕСКИ РАСПРЕДЕЛЕННЫХ ВЫЧИСЛИТЕЛЬНЫХ СИСТЕМАХ С ТРЕБОВАНИЯМИ К КАЧЕСТВУ ОБСЛУЖИВАНИЯ

Клименко А.Б.
Институт информационных наук и технологий безопасности РГГУ, Москва, Россия
Anna_klimenko@mail.ru

Аннотация. В статье рассмотрен вопрос управления обработкой данных в географически распределенных вычислительных системах на базе туманных и краевых вычислений. В данной работе представлена общая постановка задачи оптимизации и предложен метод ее решения применительно к управлению обработкой данных. Проведены вычислительные эксперименты, на основе результатов которых сделаны выводы о достижимости решения сформулированной проблемы.

Ключевые слова: обработка данных, туманные вычисления, распределение нагрузки.

Введение

В настоящее время для широкого круга систем процедуры обработки больших объемов данных зачастую проводятся в условиях ограничения на время, что весьма характерно для систем дополненной реальности, киберфизических систем, групп БЛА и т. д. Многие пользовательские приложения также имеют требования к обеспечению требований к уровню QoS.

Наличие ограничений по времени порождает проблему недостаточности облачных вычислений в аспекте обеспечения должного уровня QoS, что в итоге, начиная с 2012г., стало причиной для разработки и интенсивного использования концепций туманных и краевых вычислений. Следует отметить, что у этих концепций есть как достоинства, так и недостатки: поскольку туманный и краевой уровни сети динамичны с точки зрения топологии и вычислительной нагрузки, порождаемой пользовательскими приложениями, существует необходимость в относительно частых процедурах перепланирования и переноса вычислительных задач.

В данной работе под управлением обработкой данных в географически распределенных средах будем понимать выработку решений по поводу закрепления вычислительных задач по гетерогенному коллективу решающих устройств. Следовательно, необходимо решить, как задачу распределения нагрузки, так и обозначить метод распределения, который бы обеспечивал наискорейшую выработку решения.

К настоящему времени сложился достаточно широкий круг исследовательских работ, ведущихся в данном направлении – а именно, в направлении распределения задач по узлам, которые можно условно разделить на два основных класса:

- Распределение рабочей нагрузки с однокритериальной оценкой эффективности, например, [1,2]. Как правило, целевыми функциями являются время выполнения задач или затраты энергии;
- Распределение рабочей нагрузки с оценкой эффективности по некоторому набору критериев [2-7], которые, как правило, включают время, стоимость энергии, нагрузку на каналы передачи данных.

Однако отсутствуют такие постановки задач, которые бы учитывали неоднородность критериев туманного и краевого слоев сети, а также модели, рассматривающие процесс перепланирования и перераспределения задач в комплексе с критериями эффективности распределения нагрузки.

Основные задачи, решенные в рамках данной статьи, заключаются в следующем:

- Сформулирована обобщенная модель задачи многокритериального распределения рабочей нагрузки, отличающаяся от ранее предложенных моделей использованием гетерогенных критериев и интеграцией процесса реконфигурации в описание задачи с учетом основного временного ограничения QoS;
- Предложен основной метод решения проблемы с учетом дополнительных затрат на транзит данных.

1. Базовая модель реализации туманных вычислений

Рассмотрим некоторые основы концепции туманных вычислений, которые весьма важны для дальнейшей постановки задачи.

В современных системах, функционирующих по этой парадигме, предполагается наличие планировщика, брокера туманного слоя [8,9]. В целом данная сущность выполняет следующие шаги:

- Получает запрос пользователя на обработку какого-то объема данных;

- Пытается распределить обработку данных между туманными узлами по соседству или, если нет узлов, обладающих необходимыми ресурсами, брокер перенаправляет запрос в облако;
- Интегрирует результаты вычислений и отправляет их пользователю

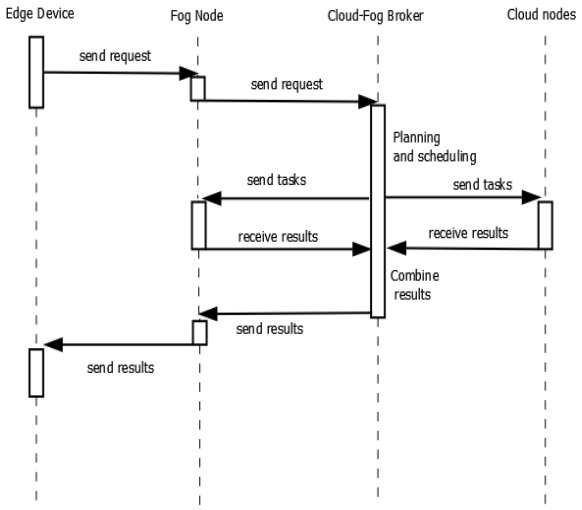

Рис. 1. Схема работы брокера туманного слоя

Кроме того, говоря о туманных вычислениях, необходимо учитывать географическое распределение туманных узлов, поскольку не гарантировано распределение нагрузки только по соседним узлам, расположенным на расстоянии одного транзитного участка сети. В случае, когда обработка данных располагается на географически распределенных узлах, каждый транзитный участок сети порождает накладные временные и ресурсные расходы при обработке данных, что также необходимо учитывать.

И, наконец, при наличии требований к качеству обслуживания, процедуру перепланирования необходимо рассматривать с точки зрения времени, которое она может занять, а также, возможно, с точки зрения других критериев распределения задач, на которые может повлиять перепланирование и собственно перенос нагрузки.

2. Формализация задачи

Будем считать систему распределенных вычислений эффективной в том случае, когда она реализует пользовательские операции оптимальным (субоптимальным) образом с точки зрения критериев оценки и при этом удовлетворяет существующим ограничениям.

Для распределения нагрузки по фрагменту географически распределенной сети используются следующие входные данные.

- Граф задач, предназначенных к решению: $G_1 = \{<g_i, r_i>, R\}$, где g_i – вычислительная сложность задачи, r_i – требования к ресурсам устройства, на котором будет происходить размещение, включая: требования к объему памяти, требования к пропускной способности канала, к производительности и т.д. Данный граф – ациклический и направленный, где ребра R взвешены объемом передаваемых данных между задачами.
- Граф сети представляется произвольным направленным мультиграфом, где вершины взвешены характеристиками узлов сети (производительность, объем памяти, энергопотребление и т.д.), ребра взвешены скоростями передачи данных по каналам связи. То есть: $G_2=\{M,C\}$, $M = \{m_j\}$ – ресурсы, которыми располагает узел, $C =\{c_j\}$ – каналы связи.
- Имеются общие критерии оценивания качества распределения задач $S_0 = \{s_k\}$.
- Гетерогенность сети и специфика используемых устройств продуцирует индивидуальные критерии качества распределения, специфичные отдельным узлам, и составляет множество: $P_0=\{p_l\}$.
- Имеются общие ограничения $constr = \{constr_k\}$.
- Процедура управления характеризуется параметрами $<g_r, r_r, t_r>$ где
- G_r – вычислительная сложность задачи перепланирования, r_r – требования к ресурсам узла, где выполняется расчет нового закрепления, t_r – время выполнения перепланирования (перемещение данных).

Таким образом, управление системой РВ будет заключаться в решении следующей задачи: необходимо для графов G_1 и G_2 найти такие закрепление задач за устройствами и $<g_r, r_r, t_r>$, чтобы при имеющихся ограничениях: $r_i <= m_j$, $constr_{,j,} r_r <= m_j$ обеспечить $S_0 \to max, P_0 \to max$.

3. Общий метод решения

Рассмотрим задачу, сформулированную ранее. В общем случае решением этой проблемы является распределение задач по узлам, что описывается матрицей:

$$A = \begin{bmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & \ddots & \vdots \\ a_{im} & \cdots & a_{nm} \end{bmatrix} \quad (1)$$

где $a_{ij} = \begin{cases} 1, if\ task\ i\ is\ assigned\ to\ the\ node\ j; \\ 0, otherwise. \end{cases}$

Рассмотрим общий случай, когда время реконфигурации системы не влияет на какие-либо конкретные ограничения или целевые функции. Тогда будет только ограничение по времени для пользовательской операции Т, на которое может повлиять неподходящий метод реконфигурации, который может занять слишком много времени. Итак, можно сформировать следующие основные требования к процедуре реконфигурации:

- метод решения задач распределения задач должен быть таким, чтобы занимать как можно меньше времени;
- перераспределенные задачи должны быть распределены по узлам таким образом, чтобы минимизировать время передачи данных;
- результирующее распределение задач должно быть максимально эффективным.

Из этих требований можно выделить следующее:

- поскольку решаемая задача является *np*-сложной, мы должны рассмотреть те методы оптимизации, которые гарантируют некоторые неоптимальные решения за ограниченное время;
- нам нужно добавить некоторые ограничения – или дополнительную целевую функцию, чтобы минимизировать перенос задач после получения нового распределения.

Первое утверждение делает выгодным использование некоторых метаэвристических итерационных методов, таких как, например, имитация отжига, случайный поиск или генетические алгоритмы. В качестве основной цели выбора метода, позволяющего за минимальное время добиться наилучшего решения, вполне рационально выбрать метод с минимально возможной вычислительной сложностью, которую можно оценить путем экспериментального исследования и дальнейшей систематизации.

Второе утверждение можно реализовать следующим образом, т.е. формируется новая целевая функция, добавляемая к общим целевым функциям.

Таким образом, общий метод решения сформулированной задачи состоит из следующих шагов.

- Сформировать добавочную ЦФ следующим образом: рассатривая матрицу назначений А как финальный вариант распределения, матрицу А’ как предшествующее перепланированию назначение, сформировать следующую ЦФ: $\forall i, j\ |a'_{ij} - a_{ij}| \to min$.
- Добавить полученную ЦФ как компонент вектора общих ЦФ.
- Выбрать представление многокритериальной функции: посредством скаляризации или как многокритериальную задачу.
- Выбрать метод решения, опираясь на экспериментальные и экспертные оценки эффективности алгоритмов решения.
- Решить задачу оптимизации.

4. Экспериментальное исследование

Рассмотрим входные данные, представленные в виде графа задач указанной трудоемкости и графа фрагмента гомогенной сети (см. рис.2,3). Целевыми функциями являются индивидуальные критерии надежности узлов, а также энергозатраты на решение заданного комплекса задач.

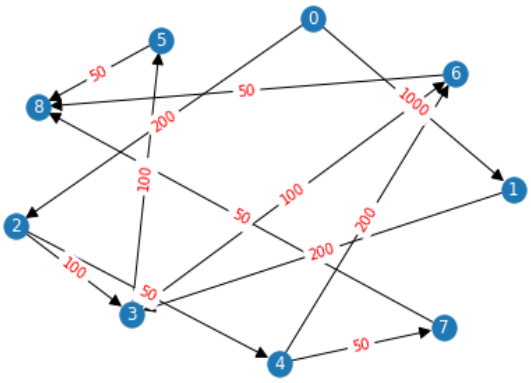

Рис. 2. Граф задач

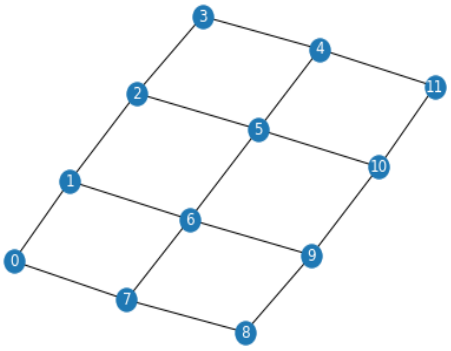

Рис. 3. Граф сети

Для решения задачи распределения нагрузки с получением фронта Парето был использован генетический алгоритм NGSAII. Полученные недоминируемые решения представлены на рис. 4. время решения для получения решений схожего качества заняло от 60 до 3600с. В табл.1 приведены решения совместно со значениями мультипликативной свертки на основе индивидуальных критериев узлов и критерия энергетических затрат.

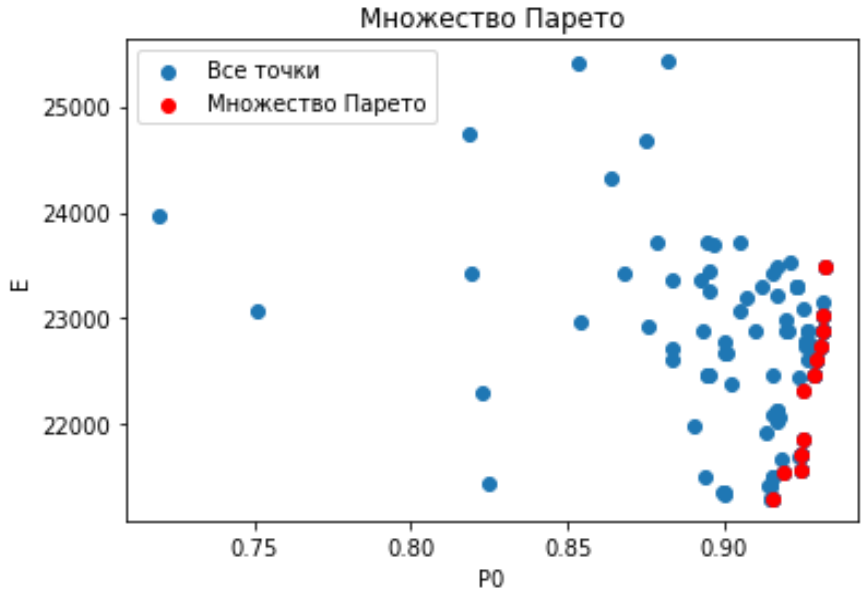

Рис. 4. Точки фронта Парето при решении двухкритериальной задачи оптимизации

Таблица 1. Недоминируемые решения фронта Парето при решении двухкритериальной задачи распределения нагрузки

	Стоимостная функция 1	**Энергозатраты**	**Значение свертки**
1	0.9291244649698247	22620.0	41.07535
2	0.9316258934321543	23040.0	40.43515
3	0.9317725129005804	23490.0	39.66677
4	0.9291244649698247	22620.0	41.07535
5	0.9287161491824625	22470.0	41.33138
6	0.9246400968895908	21720.0	42.57091
7	0.9291244649698247	22620.0	41.07535
8	0.9246400968895908	21720.0	42.57091
9	0.9189280746747525	21555.0	42.63178
10	0.9242337518158983	21570.0	42.84811
11	0.9246400968895908	21720.0	42.57091
12	0.9287161491824625	22470.0	41.33138
13	0.915601551625009	21300	42.98599
14	0.915601551625009	21300.0	42.98599
15	0.9312543652382311	22890.0	40.6839
16	0.9246400968895908	21720.0	42.57091
17	0.9287161491824625	22470.0	41.33138
18	0.9316258934321543	23040.0	40.43515
19	0.9312543652382311	22890.0	40.6839
20	0.9316258934321543	23040.0	40.43515
21	0.9291244649698247	22620.0	41.07535
22	0.9312543652382311	22890.0	40.6839

Также были выполнены распределения задач при помощи алгоритмов случайного поиска (10 тыс. итераций) и имитации отжига с температурной схемой тушения.

Таблица 2. Результаты работы алгоритмов случайного поиска и имитации отжига

	Случайный поиск	Имитация отжига
Стоимостная функция 1	0.9512	0.9184
Энергозатраты	20730.0	21180.0
Свертка	45.88739	43.3649
Время получения решения	t=5c-300c	t<1c

По результатам проведенных процедур поиска видно, что, опираясь на экспериментальные исследования и экспертные мнения имеется возможность выбора наиболее эффективного в плане

времени/трудоемкости алгоритма. Исходя из этого, могут быть применены технологии искусственного интеллекта на основании данных по описанию задач и фрагмента сети.

5. Заключение

Современные вычислительные задачи имеют высокую сложность и имеют строгие ограничения по времени на выполнение, а также туманность и краевые уровни сети представляют собой гетерогенную и динамичную среду. Несмотря на широкий спектр публикаций в этой области, точные литературные исследования выявили отсутствие моделей и методов распределения рабочей нагрузки с учетом неоднородности критериев и учета времени перепланирования.

Основными вкладами этой статьи являются:

- общая модель задачи многокритериального распределения рабочей нагрузки;
- общий метод решения проблемы с учетом дополнительных затрат на транзит данных.
- Было проведено несколько экспериментальных исследований со следующими результатами:
- Использование мультипликативной свертки с учетом транзитной передачи данных позволяет формировать сообщества устройств с меньшими затратами вычислительных ресурсов.
- Моделирование отжига с температурной схемой закалки позволяет получить неоптимальные решения хорошего качества.

В совокупности эти результаты показывают преимущества их комплексного использования для решения задачи оптимизации с неоднородными критериями.

Литература

1. *Huaiying S., Huiqun Y., Guisheng F.* Contract-Based Resource Sharing for Time Effective Task Scheduling in Fog-Cloud Environment//IEEE Trans. on Network and Service Management. PP. 1-1. 10.1109/TNSM.2020.2977843.G. 2020.
2. *Siqi L., Xu C., Zhi. Z and Shuai Y.* Fog-Enabled Joint Computation, Communication and Caching Resource Sharing for Energy-Efficient IoT Data Stream Processing// IEEE Trans. on Vehicular Technology. 70. 3715-3730. 10.1109/TVT.2021.3062664.
3. *Duong N., Long L., Vijay B. A* Market-Based Framework for Multi-Resource Allocation in Fog Computing// IEEE/ACM Transactions on Networking. PP. 1-14. 10.1109/TNET.2019.2912077.
4. *Branka M., Kostic-Ljubisavljevic A., Perakovic D., Cvitić I.* Deadline-Aware Task Offloading and Resource Allocation in a Secure Fog-Cloud Environment. Mobile Networks and Applications. 1-14. 10.1007/s11036-023-02120-y.
5. *Ali H., Sridevi, R.* Mobility and Security Aware Real-Time Task Scheduling in Fog-Cloud Computing for IoT Devices: A Fuzzy-Logic Approach. The Computer Journal. 10.1093/comjnl/bxad019.
6. *Morkevicius N., Liutkevicius A., Venčkauskas A*. Multi-Objective Path Optimization in Fog Architectures Using the Particle Swarm Optimization Approach. Sensors. 23. 3110. 10.3390/s23063110.
7. *Morkevicius N., Venčkauskas A., Šatkauskas N., Toldinas J.* Method for Dynamic Service Orchestration in Fog Computing. Electronics. 10. 1796. 10.3390/electronics10151796.
8. *Masarweh M., Alwadan T., Afandi W.* Fog Computing, Cloud Computing and IoT Environment: Advanced Broker Management System. Journal of Sensor and Actuator Networks. 11. 10.3390/jsan11040084.
9. *Hayat B., Lee S., Kim K. H.* Resource allocation through logistic regression and multicriteria decision making method in IoT fog computing. 2022. Transactions on Emerging Telecommunications Technologies. 33. 10.1002/ett.3824.