ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ И ТЕЛЕКОММУНИКАЦИИ
СПбГУТ))) Март 2017 Том 5 № 1 ISSN 2307-1303 http://www.itt.sut.ru
TELECOM IT
March 2017 Vol. 5 Iss. 1 ISSN 2307-1303 http://www.itt.sut.ru

К ПОСТРОЕНИЮ МОДЕЛИ РАЗДЕЛЕНИЯ НАГРУЗКИ В СИСТЕМЕ ТУМАННЫХ ВЫЧИСЛЕНИЙ[1]

К. Е. Самуйлов[1, 2]

[1] РУДН, Москва, 117198, Российская Федерация
[2] ФИЦ ИУ РАН, Москва, 119333, Российская Федерация
Адрес для переписки: samuylov_ke@rudn.university

Аннотация
В статье исследован компромисс между энергозатратами и задержками передачи в системе туманно-облачных вычислений. Формализована задача разделения нагрузки, которая определяет оптимальное разделение нагрузки между туманом и облаком.
Ключевые слова
облачные вычисления, туманные вычисления, оптимизация, компромисс потребления энергии и задержек, разделение нагрузки.

Информация о статье
УДК 621.391:004.02
Язык статьи – русский.
Поступила в редакцию 01.02.17, принята к печати 28.02.17.

Ссылка для цитирования: Самуйлов К. Е. К построению модели разделения нагрузки в системе туманных вычислений // Информационные технологии и телекоммуникации. 2017. Том 5. № 1. С. 8–14.

ON THE CONSTRRUCTION OF WORKLOAD ALLOCATION MODEL IN FOG COMPUTING SYSTEM

K. Samuylov[1, 2]

[1] RUDN University, Moscow, 117198, Russian Federation
[2] IPI RAN, Moscow, 119333, Russian Federation
Corresponding author: samuylov_ke@rudn.university

[1] Исследование выполнено при финансовой поддержке РНФ в рамках научного проекта № 16-11-10227.

Abstract—The article explores the trade-off between energy costs and transmission delays in the system of foggy cloud computing. The problem of workload allocation has been formalized, which determines the optimal load sharing between fog and cloud subsystems.

Keywords—Cloud computing, fog computing, optimization, power consumption-delay tradeoff, workload allocation.

Article info
Article in Russian.
Received 01.02.17, accepted 28.02.17.

For citation: Samuylov K.: On the Construction of Workload Allocation Model in Fog Computing System // Telecom IT. 2017. Vol. 5. Iss. 1. pp. 8–14 (in Russian).

Введение

Интернет переживает сдвиг в сторону структуры, основанной на облачных вычислениях. С увеличением мобильного траффика передача невероятно огромного объема данных в облако не только была тяжелой задачей для пропускной способности канала связи, но и стала причиной задержек передачи и снижения качества услуг для конечного пользователя. В дополнение к этому, с ростом роли мобильного трафика, не менее важна поддержка мобильности и георасределения. По этой причине, становление облачных вычислений в качестве всеобъемлющего подхода для централизованного хранения, получения и управления информацией, успешная интеграция облачного вычисления и мобильных приложений, является важной задачей. Для решения таких задач компания Cisco представила концепцию туманных вычислений, предназначенную для локальной обработки части заданий на туманных устройствах. Слой тумана состоит из георасределенных серверов, которые развернуты на сетевой периферии. Каждый туманный сервер представляет собой облегченную версию облачного сервера, и оборудованный хранилищем данных большого объема и способностью к вычислениям и беспроводной передачи. Задача состоит в моделировании функции потребления энергии и задержки каждой части туманно-облачной системы и формализации задачи разделения нагрузки. В итоге с помощью численных примеров показано, что туманные вычисления могут существенно улучшить систему облачных вычислений по критерию снижения задержек передачи. Данная статья фактически является кратким обзором известных работ, в том числе приведенной литературы [1, 2].

Постановка задачи

Для нахождения компромисса между энергозатратами и задержкой, с одной стороны, необходимо минимизировать совокупные энергозатраты всех туманных устройств и облачных серверов. Функция энергозатрат в туманно-облачной вычислительной среде определена как

$$P^{sys} \triangleq \sum_{i \in N} P_i^{fog} + \sum_{j \in M} P_j^{cloud} .$$

С другой стороны, необходимо гарантировать качество обслуживания, т. е. требования к задержкам запросов конечных пользователей. Задержка, которую ощущает конечный пользователь, состоит из вычислительной задержки, которая включает в себя время ожидания начала обслуживания, и задержки передачи по сети (коммуникационные задержки). Таким образом, функция задержек в туманно-облачной вычислительной системе имеет вид:

$$D^{sys} \triangleq \sum_{i \in N} D_i^{fog} + \sum_{j \in M} D_j^{cloud} + \sum_{i \in N} \sum_{j \in M} D_{ij}^{comm} .$$

Мы рассматриваем задачу минимизации энергопотребления туманно-облачной вычислительной системы и обеспечение нижней границы задержки $\bar{D}$ для конечных пользователей. Целевые переменные – это рабочая нагрузка x_i i-го туманного устройства, рабочая нагрузка y_j j-го облачного сервера, поток трафика λ_{ij}, отправленный из i-го туманного устройства на j-й облачный сервер, а также частота процессора f_i, количество машин n_i и индикатор включённого/выключенного состояния σ_j на j-м облачном сервере. Основной задачей распределения рабочей нагрузки на туманно-облачной вычислительной системе – это компромисс между системным энергопотреблением и задержками, ощущаемыми конечным пользователем.

Декомпозиция задачи и подход к решению

Для решения поставленной задачи в [1] был разработан подход, заключающийся в ее декомпозиции на три подзадачи, которые могут быть решены с помощью существующих методов оптимизации.

1. Компромисс энергозатрат и задержек для туманного вычисления (рис. 1), в и состоит первая подзадача, которая может быть сформулирована следующим образом:

$$\min_{x_i} \sum_{i \in N} (a_i x_i^2 + b_i x_i + c_i + \frac{\eta_i}{\nu_i - x_i}) ,$$

$$\begin{cases} \sum_{i \in N} x_i = X \\ 0 \le x_i \le \min\{x_i^{\max}, l_i\}, \forall i \in N \end{cases} .$$

Здесь параметр n_i является долевым коэффициентом в компромиссе между энергозатратами и вычислительной задержкой на i-м туманном устройстве. При условии, что рабочая нагрузка X распределена относительно туманной подсистемы, первая подзадача является выпуклой задачей с линейными ограничениями. Задача с легкостью решается методом выпуклой оптимизации, например, методом внутренних переменных. После нахождения оптимальной рабочей нагрузки x_i для i-го туманного устройства, можем посчитать энергопотребление и вычислительную задержку в туманной вычислительной подсистеме соответственно:

$$\begin{cases} P^{fog}(X) = \sum_{i \in N} [a_i (x_i^*)^2 + b_i x_i^* + c_i] \\ D^{fog}(X) = \sum_{i \in N} \dfrac{1}{\nu_i - x_i^*} \end{cases}.$$

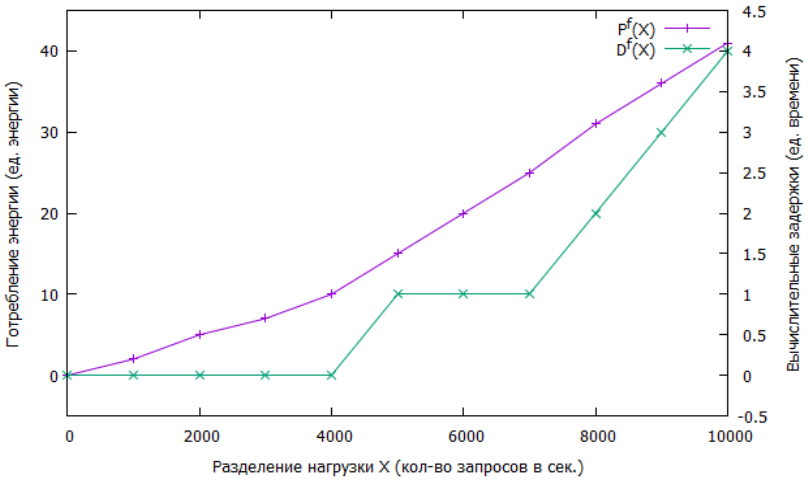

Рис. 1. Компромисс для подсистемы туманного вычисления

2. Компромисс энергозатрат и задержек для системы облачного вычисления (рис. 2). Мы предполагаем, что задержка должна быть меньше, чем регулируемый параметр $\bar{D}_j$, который можно считать порогом задержки, определяющей диапазон дохода/штрафа на j-м облачном сервере, т. е. $D_j^{cloud} \le \bar{D}_j$. Тогда вторая подзадача может быть сформулирована следующим образом:

$$\min_{y_i, f_j, n_j, \sigma_j} \sum_{j \in M} \sigma_j n_j (A_j f_j^p + B_j),$$

$$\begin{cases} \sum_{j \in M} y_j = Y \\ D_j^{cloud} \le \bar{D}_j \\ y_j \ge 0, \forall j \in M \\ f_j^{\min} \le f_j \le f_j^{\max}, \forall j \in m \\ n_j \in \{0,1,2,\ldots,n_j^{\max}\}, \forall j \in M \\ \sigma_j \in [0,1], \forall j \in M \end{cases}.$$

При условии, что рабочая нагрузка Y распределена по вычислительной облачной подсистеме, вторая подзадача является задачей частично целочисленного нелинейного программирования, решение которой представляет некоторую сложность. Например, может быть применен обобщённый метод декомпозиции Бендера, для которого в [1] разработан вычислительный алгоритм. После нахождения оптимальной рабочей нагрузки y_j^* для j-го облачного серве-

ра и оптимального решения f_j^*, n_j^* и σ_j^*, можно вычислить энергопотребление и вычислительную задержку в облачной вычислительной подсистеме, соответственно:

$$\begin{cases} P^{cloud}(Y) = \sum_{j \in M} \sigma_j^* n_j^* \left[A_j (\sigma_j^*)^p + B_j \right] \\ D^{cloud}(Y) = \sum_{j \in M} D_j^{cloud*} = \sum_{j \in M} \sigma_j^* \bar{D}_j \end{cases},$$

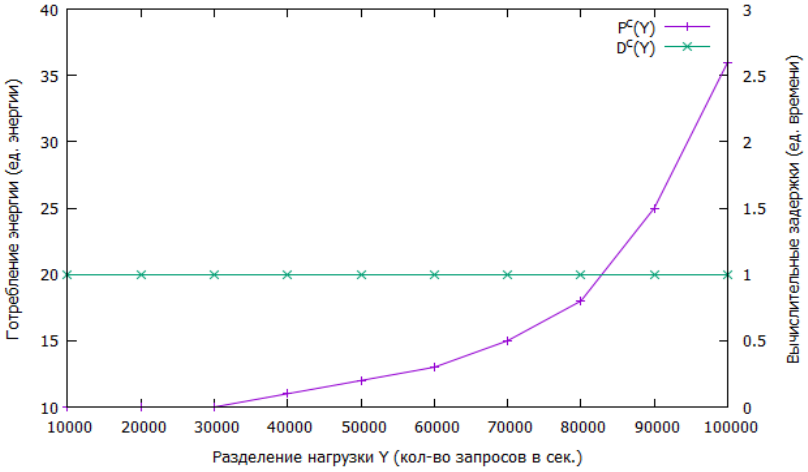

Рис. 2. Компромисс в подсистеме облачного вычисления

3. Минимизация коммуникационной задержки для отправок. Рассматриваем поток с интенсивностью λ_{ij} для минимизации коммуникационной задержки в сетевой подсистеме. В этом случае формализация третьей подзадачи состоит в следующем:

$$\min_{\lambda_{ij}} \sum_{i \in N} \sum_{j \in M} d_{ij} \lambda_{ij},$$

$$\begin{cases} l_i - x_i = \sum_{j \in M} \lambda_{ij}, \forall i \in N \\ \sum_{j \in N} \lambda_{ij} = y_j, \forall j \in M \\ 0 \le \lambda_{ij} \le \lambda_{ij}^{\max}, \forall i \in N; \forall j \in M \end{cases}.$$

Теперь, при условии, что рабочая нагрузка X распределена для туманного устройства и рабочая нагрузка Y распределена для облачной подсистемы, можно получить оптимальную рабочую нагрузку x_i^* на i-м туманном устройстве и y_j^* на j-м облачном сервере. Для нахождения x_i^* и y_j^*, третья подзадача рассматривается как задача назначений. Так как подобная задача может быть эффективно решена с помощью венгерского метода, в [1] разработан алгоритм на основе этого метода. После получения потока трафика λ_{ij}, отправленного из i-го

туманного устройства на j-й облачный сервер, можно вычислить коммуникационную задержку в сетевой подсистеме следующим образом:

$$D^{comm}(X,Y) = \sum_{i \in N} \sum_{j \in M} d_{ij} \lambda^*_{ij},$$

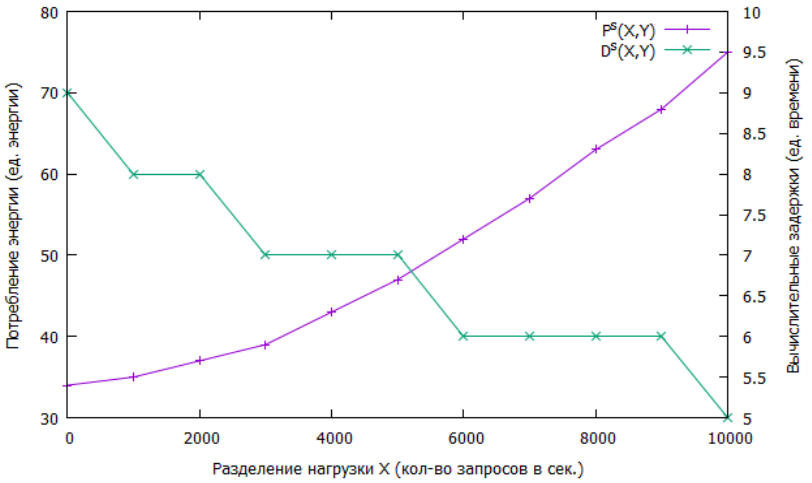

Рис. 3. Компромисс в системе туманно-облачного вычисления

Основываясь на декомпозиции, и решении трех подзадач, идея которых изложена выше, энергопотребление туманно-облачной вычислительной системы (рис. 3) может быть представлено в следующем виде:

$$P^{sys}(X,Y) \triangleq P^{fog}(X) + P^{cloud}(Y).$$

Это означает, что энергопотребление исходит как из туманных устройств, так и из облачных серверов. С другой стороны, функция задержки может быть переписана как:

$$D^{sys}(X,Y) \triangleq D^{fog}(X) + D^{cloud}(Y) + D^{comm}(X,Y),$$

что в свою очередь означает, что задержка системы состоит из вычислительной задержки туманных устройств и облачных серверов, а также из коммуникационной составляющей задержки.

После решения трех выше сформулированных подзадач, мы можем сформулировать исходную задачу:

$$\min_{X,Y} P^{sys}(X,Y), \tag{1}$$

$$\begin{cases} D^{sys}(X,Y) \le \bar{D} \\ X + Y = L \end{cases},$$

которая может быть решена итерационно. Уровень аппроксимации зависит от двух изменяемых параметров: η_j и D_j. Если подобрать эти параметры верно, решение задачи (1) является оптимальным решением исходной задачи. Задача оценки точности аппроксимации является предметом предмет дальнейших исследований.

Выводы

В работе представлена идея формализации оптимизационных задач для систем туманно-облачных вычислений. Построена модель для исследования задачи нахождения компромисса энергопотребления и задержек в туманно-облачной вычислительной системе. Также сформулирована задача разделения рабочей нагрузки, она была декомпозирована на три подзадачи, которые могут быть решены для каждой из подсистем в отдельности. Численные результаты представлены для иллюстрации того, как туманные вычисления расширяют и дополняют облачные вычисления.

Автор выражает благодарность студенту РУДН Мацкевичу Ивану за помощь в расчетах.

Литература / References

1. Deng R., Lu R., Lai C., Luan T.H., Liang H. Optimal Workload Allocation in Fog-Cloud Computing Toward Balanced Delay and Power Consumption // IEEE Internet of Things Journal. 2016. Vol. 3. Iss. 6. pp. 1171–1181. DOI: 10.1109/JIOT.2016.2565516.

2. Gorbunova A. V., Zaryadov I. S., Matushenko S. I., Sopin E. S. The Estimation of Probability Characteristics of Cloud Computing Systems with Splitting of Requests // Nineteenth International Scientific Conference "Distributed Computer and Communication Networks: Control, Computation, Communications" (DCCN). 2016. Vol. 3. pp. 467–472.

Самуйлов Константин Евгеньевич – доктор технических наук, профессор, РУДН, Москва, 117198; ФИЦ ИУ РАН, Москва, 119333 Российская Федерация, samuylov_ke@rudn.university

Samuylov Konstantin – Doctor of Technical Sciences, Professor, RUDN University, Moscow, 117198; IPI RAN, Moscow, 119333, Russian Federation, samuylov_ke@rudn.university