Разработка модели и алгоритмов обслуживания трафика в системах облачных услуг

Волков Александр Олегович

Московский физико-технический институт, магистрант кафедры Инфокоммуникационных Систем и Сетей, Москва, Россия,
aleksandr.o.volkov@phystech.edu

Степанов Сергей Николаевич

Московский университет связи и информатики, зав. кафедрой, д.т.н., профессор, Москва, Россия,
stpnvsrg@gmail.com

Аннотация — В облачных системах существует серьезный разброс в требованиях к количеству предоставляемого ресурса и необходимость незамедлительной обработки поступающих заявок, что затрудняет эффективное использование облака. Предложенная аналитическая модель обслуживания системы облачных услуг с режимом обслуживания Processor Sharing (PS) решает эти проблемы. Поток запросов на обслуживание описывается моделью Пуассона. Длительность обслуживания запросов в модели существенно не зависит от требований клиентов к производительности. В рамках предложенной модели определения основных показателей эффективности передачи данных формулируются через значения вероятностей стационарных состояний модели. Предложенная модель и результаты ее анализа могут быть использованы для оценки основных характеристик производительности и качества обслуживания (QoS) облачных систем.

Ключевые слова: облачные вычисления, показатели производительности, Quality of Service (QoS), Processor Sharing (PS).

Введение

По мере развития сети Интернет появляется потребность в больших вычислительных мощностях и удобный доступ к ним. Облачные системы представляют собой новый подход, который может решить вышеупомянутую проблему, благодаря представлению эффективного и удобного доступа к вычислительным ресурсам через Интернет.

Для поставщиков облачных услуг одной из самых актуальных задач является поддержание требуемого качества обслуживания на приемлемом для клиентов уровне. Это дополнительно усложняет работу провайдеров, поскольку теперь им требуется не только управлять своими ресурсами, но и обеспечивать ожидаемое QoS для клиентов. Всё это требует точного и хорошо адаптированного механизма анализа производительности предоставляемого сервиса. По изложенным выше причинам разработка модели и алгоритмов оценки требуемого по нагрузке и качеству обслуживания ресурса в системах облачных услуг является актуальной задачей, которая играет значимую роль в контексте решения проблем производительности облачных систем. Процесс взаимодействия между клиентами и провайдерами облака показан на рис.1.

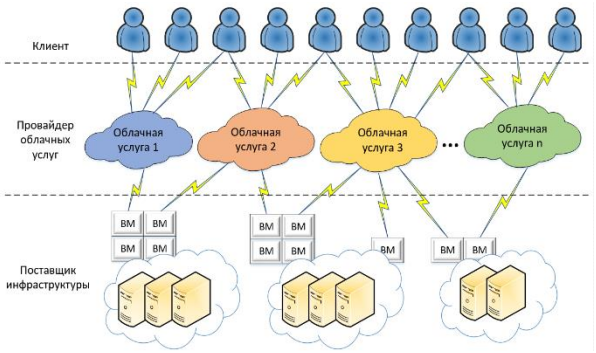

Рис. 1. Трехуровневая архитектура облачных вычислений

Важнейшей задачей при исследовании производительности облачных систем, описываемых с помощью теории массового обслуживания, является эффективное использование ограниченного ресурса передачи данных [1-3]. Для решения этой задачи необходимо проанализировать процесс совместного обслуживания поступающего трафика данных.

В контексте облачных вычислений наиболее подходящим сценарием обслуживания может быть Processor Sharing [4]. В данном режиме емкость вычислительных ресурсов в равной степени распределяется между всеми клиентами в системе. Использование Processor Sharing в качестве сценария обслуживания позволяет обрабатывать заявки сразу по прибытии, кроме того он эффективен даже при достаточно серьезной разнице в требованиях к объему предоставляемого ресурса.

Целью статьи является разработка и анализ модели обслуживания данных в системе облачных вычислений. Статья основана на результатах [2-7] и организована следующим образом: основная часть включает описание принципа работы модели облака, которая была использована для анализа, математический анализ модели на основе марковского процесса и основные определения показателей эффективности облачной системы через значения вероятностей стационарных состояний модели. Заключение же содержит в себе краткое изложение основных выводов и результатов работы.

Основная часть

В модели рассмотрен процесс одновременной обработки запросов облаком заказанных сервисов в режиме Processor Sharing. Каждый клиент может заказать один из n облачных сервисов, это означает, что в системе

имеется n потоков. Обозначим через C - производительность облака (общее количество ресурса), выраженную в операциях с плавающей запятой, flop/s. Для обслуживания k-го потока необходимо обеспечить ресурс в размере C_k с вероятностью более, чем $1-\varepsilon_k$, где ε_k - целевой показатель производительности системы.

Предположим, что в контексте k-го клиента объем требуемой производительности имеет экспоненциальное распределение со средним значением σ_k, выраженным в flop. Запросы для k-го потока поступают в облако согласно пуассоновскому процессу с интенсивностью λ_k. Тогда интенсивность поступления запросов от k-го потока и суммарная интенсивность всех запросов равны соответственно:

$$A_k=\lambda_k\sigma_k \text{ и } A=\sum_{k=1}^{n}A_k$$

Обозначим через $\alpha_k=A_k/C_k$ соответствующую интенсивность предложенного трафика, то есть среднее количество обслуживаемого ресурса с размером C_k. Одним из показателей качества обслуживания является коэффициент потенциальной загрузки облака, в данном случае он вычисляется как $\rho=A/C$. Процесс совместного использования общей производительности облака показан на рисунке 2.

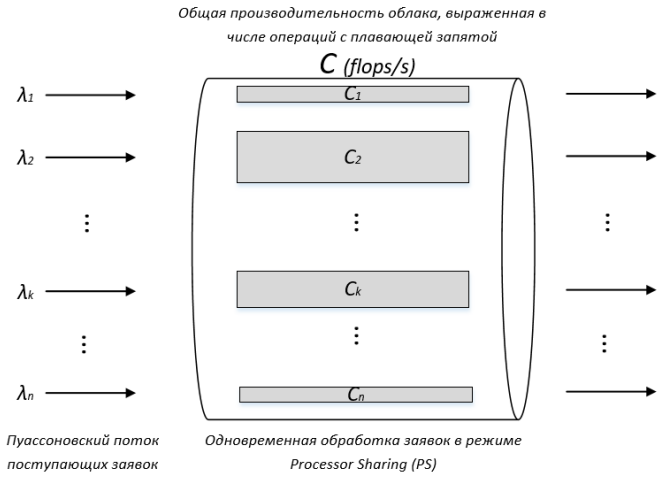

Рис. 2. Модель совместного использования общей производительности облака

Обозначим через $(i_1, i_2, \ldots, i_n)$ - состояние модели, где i_k - число заявок k-го потока, которые обслуживаются в данный момент, $k=1,2,\ldots,n$. Для заявки в k-го потоке выделяется производительность в размере C_k flop/s, если общее количество производительности выделенной всем потокам, включая текущую, не превосходит C. Если же данное условие является невыполненным, то все обслуживаемые заявки, делят между собой весь общий ресурс. Обозначим за $v_k(i_1, i_2, \ldots, i_n)$ - производительность облака, отданная для обслуживания k-му потоку в состоянии $(i_1, i_2, \ldots, i_n)$. При этом для каждого состояния системы выполнено:

$$\sum_{k=1}^{n}v_k(i_1, i_2, \ldots, i_n)\le C;$$

$$v_k(i_1, i_2, \ldots, i_n)\le i_kC_k,\quad k=1,2,\ldots,n$$

Пусть

$$i=i_1C_1+i_2C_2+\ldots+i_nC_n,\quad i=0,1,\ldots$$

Параметр i показывает значение возможной загрузки облака в состоянии $(i_1, i_2, \ldots, i_n)$ на обслуживание всех имеющихся заявок. В случае $i\le C$ каждый поток получает свою максимальную производительность при отсутствии задержек, то есть выполнено $v_k(i_1, i_2, \ldots, i_n)=i_kC_k$. Если же $i>C$, то суммарная производительность всех потоков равна общей производительности облака при наличии перегрузок. Запишем функцию баланса для этих двух случаев:

$$\Phi(i_1,\ldots,i_n)=\begin{cases}\prod_{k=1}^{n}\frac{1}{i_k!C_k^{\,i_k}} & i\le C\\ \frac{1}{C}\sum_{k=1}^{n}\Phi(i_1,\ldots,i_k-1,\ldots,i_n) & i>C\end{cases}$$

Приведенная выше модель может быть описана марковским процессом:

$$r(t)=(i_1(t),\ i_2(t),\ \ldots,\ i_n(t)),$$

где $i_k(t)$ - количество обслуживаемых заявок k-го потока в момент времени t.

Пусть $\pi(i_1, i_2, \ldots, i_n)$ - ненормированные стационарные распределения количества потоков каждого состояния. Для существования стационарности режима обслуживания необходимо:

$$\pi(i_1, i_2, \ldots, i_n)=\pi(0, 0, \ldots, 0)\,\Phi(i_1, i_2, \ldots, i_n)\prod_{k=1}^{n}A_k^{\,i_k}$$

В итоге получаем:

$$\pi(i_1,\ldots,i_n)=\begin{cases}\pi(0)\prod_{k=1}^{n}\frac{\alpha_k^{\,i_k}}{i_k!} & i\le C\\ \frac{1}{C}\sum_{k=1}^{n}A_k\pi(i_1,\ldots,i_k-1,\ldots,i_n) & i>C\end{cases}$$

Определим основные показатели обслуживания заявок. Одной из важных характеристик производительности системы является скорость образования задержек в системе (вероятность, что $i>C$):

$$G_k = \frac{\sum_{i>C} i_k \pi(i_1, \ldots, i_n)}{\sum i_k \pi(i_1, \ldots, i_n)}$$

В заключении определим еще несколько основных показателей QoS в для k-го потока: W_k - среднее время обслуживания для одной заявки, L_k - среднее число обслуживающихся заявок в системе, ϑ_k - среднее значение пропускной способности для обслуживания одной заявки.

$$L_k = \sum_{(i_1, \ldots, i_n)} p(i_1, \ldots, i_n) i_k;$$

$$W_k = \frac{L_k}{\lambda_k}; \quad \vartheta_k = \frac{\sigma_k}{W_k} = \frac{A_k}{L_k}, \quad k = 1, 2, \ldots, n.$$

Заключение

Построена и проанализирована модель совместного обслуживания системы облачных вычислений. Поток запросов на обслуживание описывается моделью Пуассона. При этом полученные результаты показывают, что в рассмотренной модели такие характеристики как среднее время обслуживания одной заявки и скорость образования задержек в системе не зависят от количества необходимого ресурса, что, в свою очередь, позволяет обслуживать клиентов с различными требованиями к количеству ресурса и увеличивает пропускную способность системы. В рамках предложенной модели даны определения основным показателям эффективности системы. Часть из них формулируются через значения вероятностей стационарных состояний модели. Предложен алгоритм оценки введенных показателей эффективности, основанный на рекурсивном решении системы уравнений состояния. Разработанный алгоритм может быть использован для пространства состояний модели, содержащего достаточно большое количество состояний. Оно является достаточным для большинства интересных с практической точки зрения случаев. Предложенная модель и результаты ее анализа могут быть использованы для оценки требуемой производительности облака и основных параметров эффективности обслуживания облачных вы. Полученные результаты могут быть также использованы для выявления «узких» мест в облаке.

Литература

[1] Степанов С.Н. Основы телетрафика мультисервисных сетей. М.: Эко-Трендз, 2010. 392 с

[2] Степанов С.Н. Теория телетрафика: концепции, модели, приложения / Серия «Теория и практика инфокоммуникаций». М.: Горячая линия–Телеком, 2015. 868 с.

[3] B. G. Batista, J. C. Estrella, C. H. Ferreira, D. M. Filho и др. Performance Evaluation of Resource Management in Cloud Computing Environments // PloS One. 2015. 10(11)

[4] Волков А.О., Степанов С.Н. Возможные режимы обслуживания ресурса и их оценка с точки зрения применимости к системам облачных услуг // Труды международной научно-технической конференции «Телекоммуникационные и вычислительные системы-2019». М.: МТУСИ. 2019. С. 9–11.

[5] J. Vilaplana, F. Solsona, I. Teixido, J. Mateo, F. Abella, J. Rius, A Queuing theory model for cloud computing. // Supercomput. 2014. № 69(1). С. 492–507

[6] Степанов С.Н., Степанов М.С. Планирование ресурса передачи при совместном обслуживании мультисервисного трафика реального времени и эластичного трафика данных. // Автоматика и телемеханика. 2017. № 11. С. 79–93.

[7] Степанов С.Н., Степанов М.С. Планирование ресурса передачи информации соединительных линий мультисервисных иерархических сетей доступа. // Автоматика и телемеханика. 2018. № 8. С. 66–80.