К АНАЛИЗУ ЗАДЕРЖЕК В ПРЕДОСТАВЛЕНИИ ОБЛАЧНЫХ УСЛУГ В МОДЕЛЯХ С СИСТЕМОЙ МОНИТОРИНГА

Масловская Н.Д., Протасова К.В, Шевякова К.А.

Российский университет дружбы народов,
maslov.natik@mail.ru, evsivia90@yandex.ru,
ksenia-sh@mail.ru

В статье представлен сравнительный анализ двух моделей облачной инфраструктуры, который позволяет оценить задержку в предоставлении облачных услуг, связанную с процессом мониторинга.

Ключевые слова: облачные вычисления, центр обработки данных, виртуальная машина, система мониторинга, рекуррентный алгоритм, вероятность блокировки, среднее время.

Введение

В последнее время широкое распространение и активное продвижение получила концепция «облачных вычислений» (cloud computing), которая реализует независящий от конечного абонентского оборудования доступ по требованию к разделяемому между многими пользователями набору разнородных вычислительных ресурсов, территориальное расположение и объем которых может меняться во времени.

Усилия по стандартизации облачных технологий консолидирует Международный союз электросвязи. Из требований, предъявляемых к облачной инфраструктуре, следует необходимость в мониторинге производительности центра обработки данных (ЦОД), в том числе, в отслеживании текущего числа включенных виртуальных машин (ВМ), которое производится с помощью системы мониторинга. При работе системы мониторинга происходит приостановка работы ВМ, соответственно, при мониторинге возникает задержка в обслуживании запросов [2].

В докладе описаны две аналитические модели, учитывающие время, затраченное на мониторинг. Их различие заключается в режиме работы системы мониторинга. В первой модели отслеживание текущего числа включенных ВМ происходит только по завершению обслуживания одного из запросов, находящегося на ВМ. Во второй модели система мониторинга включается также и при поступлении на обслуживание нового запроса. В результате чего погрешность измерений у системы мониторинга во второй модели уменьшается, но при этом увеличивается среднее время обслуживания запроса [1,3].

Построение моделей

В статье рассматривается приложение, размещенное на облачном ЦОД. ЦОД состоит из C ВМ, конечного буфера размером r и системы мониторинга (рис 1). Предполагается, что входящий поток запросов является пуассоновским с интенсивностью λ, а время предоставления облачной услуги распределено по экспоненциальному закону со средним $1/\mu$. Система мониторинга осуществляет свою работу по экспоненциальному закону за среднее время $1/\alpha$.

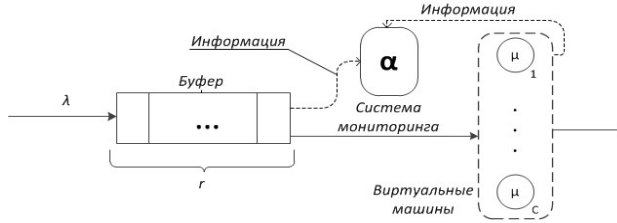

Рис. 1. Схема модели доступа к облачной инфраструктуре

Пусть n – число запросов в системе, а s обозначает состояние системы мониторинга («on»/«off»). Тогда вектор $(n,\ s)$ описывает состояние системы, а $p(n,\ s)$ – вероятность нахождения системы в состоянии $(n,\ s)$. $R = C + r$ – суммарное количество мест в системе. Пространство состояний для данной модели имеет вид:

$$X = \{(0,0), (n,s), n = 1, \dots, R, s = 0, 1\}.$$

Состояние $(0,0)$ описывает случай, когда вся система пуста. В состояниях $(n,\ 1)$ система мониторинга находится в режиме «on», в состояниях $(n,\ 0)$ в режиме «off». Переход системы мониторинга из состояния «off» в состояние «on» у первой модели происходит только в том случае, если один из запросов, находящихся на ВМ, завершил обслуживание. При этом запросы, поступившие, пока система мониторинга находилась в режиме «off», сразу поступают на обслуживание на ВМ. Это создает погрешность статистических данных, так как система мониторинга сможет отследить поступление новых запросов и занятие ими ВМ, не раньше, чем обслужится один из запросов. Во второй модели запросы, поступающие в систему, когда она находится в состояниях $(n,\ 0), n = \overline{1,\ C-1}$ будут сначала обслуживаться системой мониторинга, а затем будут распределяться по ВМ. В таком случае среднее время обслуживания запроса увеличивается, но при этом увеличивается и точность при отслеживании текущего числа занятых ВМ. В обеих моделях после того, как все ВМ будут заняты, новый поступивший запрос будет помещаться в буфер и находиться там до того момента, пока не обслужится один из запросов на ВМ.

Рекуррентный алгоритм расчета стационарных вероятностей

Диаграммы интенсивностей переходов двух моделей будут различаться переходами из состояния $(n,\ 0)$ при $n = 1, \dots, C-1$. Так как в первой модели при приходе нового запроса система мониторинга остается в режиме «off», это отражается на графе переходом из состояния $(n,\ 0)$ в состояние $(n+1,0)$ (рис. 2). В то время как во второй модели при поступлении новой заявки система будет переходить из состояния $(n,\ 0)$ в состояние $(n+1,1)$.

Алгоритм расчета стационарных вероятностей для первой и второй модели проводится аналогично, поэтому подробно рассматривается только вторая модель [3]. Вывод рекуррентного алгоритма расчета стационарного распределения вероятностей основан на линейных преобразованиях системы уравнений равновесия.

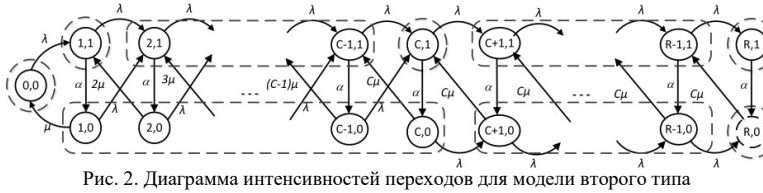

Рис. 2. Диаграмма интенсивностей переходов для модели второго типа

Для получения ненормированных вероятностей применяется алгоритм расчета:

1. $q(0,0) = 1$.
2. $q(1,0) = \frac{\lambda}{\alpha}$, $q(1,1) = \frac{\lambda}{\alpha}(\lambda + \mu)q(1,0)$.
3. $q(2,0) = \frac{\lambda}{2\alpha}[(\lambda + \alpha)q(1,1) - \lambda]$, $q(2,1) = \frac{\lambda}{2\alpha}(\lambda + 2\mu)q(2,0)$.
4. для $n = 3, \dots, C$

   $q(n,0) = \frac{\lambda}{n\alpha}\left[(\lambda + \alpha)q(n-1,1) - \lambda\, q(n-2,1) + (n-2)\mu\, q(n-2,0)\right]$

   $q(n,1) = 1/\alpha(\lambda + n\mu)q(n,0)$
5. $q(C+1,0) = \frac{\lambda}{C\alpha}\left[(\lambda + \alpha)q(C,1) - \lambda\, q(C-1,1) + (C-1)\mu\, q(C-1,0)\right]$

$q(C+1,1) = \frac{1}{\mu}[(\lambda + C\mu)q(C+1,0) - \lambda q(C,0)]$

6. для $n = C+2, \ldots, R-1$

$q(n,0) = \frac{1}{\alpha}[(\lambda + \alpha)q(n-1,1) - \lambda q(n-2,1)]$

$q(n,1) = \frac{1}{\mu}[(\lambda + C\mu)q(n,0) - \lambda q(n-1,0)]$

7. $q(R,0) = \frac{1}{\alpha}[(\lambda + \alpha)q(R-1,1) - \lambda q(R-2,1)]$

$q(R,1) = \frac{1}{C\mu}[\lambda q(R,0) - \lambda q(R-1,0)]$

Получив ненормированные вероятности, можно рассчитать нормирующую константу $G = \sum_{(i,j)\in \mathbf{X}} q(n,s)$, а затем найти стационарное распределение вероятностей:

$$p(n,s) = q(n,s)/G, (n,s) \in \mathbf{X}.$$

Основными вероятностно-временными характеристиками рассматриваемых моделей являются вероятность блокировки запроса на предоставление облачной услуги, $B = p(R,0) + p(R,1)$, среднее время задержки в предоставлении услуги по причине работы системы мониторинга $W = \frac{\sum_{i=C}^{R-1} (i-C)p(i,1) + \sum_{i=R}^{\infty} (R-C)p(i,1)}{\lambda(1-B)}$ и суммарное среднее время задержки: $W = w + \frac{\sum_{i=C}^{R-1} (i-C)p(i,1) + \sum_{i=R}^{\infty} (R-C)p(i,1)}{\lambda(1-B)}$.

Заключение

В докладе представлено сравнение двух вероятностных моделей облачной инфраструктуры с системой мониторинга, разработаны рекуррентные алгоритмы для расчета стационарного распределения вероятностей состояний моделей, а также получены формулы для вычисления вероятностно-временных характеристик. В результате проведенных исследований можно сделать вывод, что хоть использование модели первого типа дает бóльшую погрешность в определении текущего числа занятых ВМ, при небольших нагрузках ее использование целесообразнее с точки зрения снижения времени задержки в обслуживании.

Литература

1. Khaled S. and Boutaba R. Estimating service response time for elastic cloud applications // Proc. of the International Conference on Cloud Networking CLOUDNET. – IEEE. – 2012. – P. 12–16.
2. Мокров Е.В., Самуйлов К.Е. Модель системы облачных вычислений в виде системы массового обслуживания с несколькими очередями и с групповым поступлением заявок // T-Comm – Телекоммуникации и Транспорт. – 2013. – №11. – С. 139–141.
3. Гудкова И.А., Масловская Н.Д. Вероятностная модель для анализа задержки доступа к инфраструктуре облачных вычислений с системой мониторинга // T-Comm – Телекоммуникации и Транспорт. – 2014. (в печати)

FOR ANALYSING IMPACT OF DELAYS ON MEAN SERVICE TIME IN CLOUD COMPUTING IN MODELS WITH MONITORING SYSTEM

Maslovskaya N.D., Protasova K.V., Shevyakova K.A.

Peoples' Friendship University of Russia,
maslov.natik@mail.ru, evsivia90@yandex.ru,
ksenia-sh@mail.ru

The paper presents a comparative analysis of two models of cloud infrastructure, which allows to estimate the delay in providing cloud services related to the monitoring process.

Key words: cloud computing, data center, virtual machine, monitoring, recursive algorithm, blocking probability, mean time.