УДК 519.872

ПРИМЕНЕНИЕ АСИМПТОТИЧЕСКОГО МЕТОДА ДЛЯ АНАЛИЗА ЗАМКНУТОЙ ДВУХФАЗНОЙ МОДЕЛИ ВЫЧИСЛИТЕЛЬНОГО УЗЛА ОБРАБОТКИ ДАННЫХ

И.Л. Лапатин
Томский государственный университет
Россия, 634050, Томск, Ленина пр., 36
E-mail: ilapatin@mail.ru

О.Д. Лизюра
Томский государственный университет
Россия, 634050, Томск, Ленина пр., 36
E-mail: oliztsu@mail.ru

А.А. Назаров
Томский государственный университет
Россия, 634050, Томск, Ленина пр., 36
E-mail: nazarov.tsu@gmail.com

С.В. Пауль
Томский государственный университет
Россия, 634050, Томск, Ленина пр., 36
E-mail: paulsv82@mail.ru

Ключевые слова: сети массового обслуживания, замкнутая модель, метод асимптотического анализа, узел обработки данных.

Аннотация: В работе рассмотрена базовая модель вычислительного узла в виде замкнутой двухфазной сети массового обслуживания. Заявки отождествляются с виртуальными машинами (ВМ), запущенными на узле, их количество фиксировано. Фазы отождествляются с режимами работы ВМ: активный режим и режим ожидания (пассивный). При этом в модели учитывается зависимость времени пребывания на фазе от количества заявок на ней, что позволяет учитывать конкуренцию ВМ за ресурсы вычислительного узла. Для предложенной модели проиллюстрирована возможность применения метода асимптотического анализа и найдена Гауссовская аппроксимация числа заявок на фазах. Данный подход предполагается развивать для более сложных по структуре замкнутых моделей, учитывающих разнообразные фазы работы ВМ и маршрутизацию переходов между ними.

1. Введение

Современную жизнь уже невозможно представить без наличия облачных сервисов, которые позволяют пользователям без покупки дорогостоящего оборудования использовать удаленные вычислительные ресурсы для обработки данных. Провайдеры таких удаленных сервисов сталкиваются с проблемой балансирования между желанием получить как можно больше клиентов и опасностью ухудшения качества предоставляемых для них услуг. Ухудшение качества происходит из-за того, что фактически разным клиентам облачного узла приходится конкурировать за его физические ресурсы [1–3], прежде всего вычислительные. А стохастическая природа возникновения запросов от каждого клиента не позволяет точно спрогнозировать суммарную нагрузку в каждый момент времени. Таким образом, у провайдеров возникает проблема управления допустимым числом клиентов (нагрузкой), для получения максимальной прибыли.

Для моделирования работы вычислительных узлов [4–6] хорошо подходит теория массового обслуживания, математический аппарат которой позволяет учитывать стохастическую природу моментов возникновения задач и их размера. Некоторым авторам [4, 7, 8] удается учитывать эффект снижения производительности узла при увеличении нагрузки на него, но в результате исследования часто удается найти лишь усредненные числовые характеристики модели или, в ином случае, деградация моделируется в виде ступенчатой функции, что ограничивает применимость таких моделей. А для анализа производительности и управления нагрузкой необходимо знание не только о средних значениях, но и виде распределения вероятностей возникающей конкуренции. Это необходимо чтобы оценивать с какой вероятностью снижение производительности не превысит некоторого недопустимого значения. Предлагаемый асимптотический подход позволяет оценивать именно распределение вероятностей, активно работающих ВМ, которое и определяет уровень конкуренции на узле.

Будем полагать, что количество ВМ, которое запущено на вычислительном узле фиксировано и не изменятся. Но каждая ВМ имеет два разных режима работы, которые условно можно разделить на активную и пассивную фазу работы, которые могут меняться в случайном порядке. В активной фазе работы происходит непосредственно обработка данных с использованием вычислительных ресурсов узла, а в пассивной ВМ находится в режиме ожидания и почти не потребляет ресурсов. Заявки в модели будем отождествлять с ВМ на вычислительном узле, а фазы обслуживания с фазами работы ВМ. При этом интенсивность обслуживания на каждой фазе зависит от числа ВМ на активной и пассивной фазе, что позволяет учитывать в модели эффект конкуренции и снижения производительности вычислительного узла.

2. Описание математической модели

Рассмотрим замкнутую двухфазную сеть массового обслуживания с конечным числом приборов N (рис. 1). Каждый прибор имеет два состояния по числу фаз обслуживания. Так как система не имеет входов и выходов, модель описывает процесс циркулирования конечного числа N виртуальных машин в облачном узле.

В данной модели также учитывается деградация скорости обслуживания, которая зависит от числа заявок (виртуальных машин) на каждой фазе.

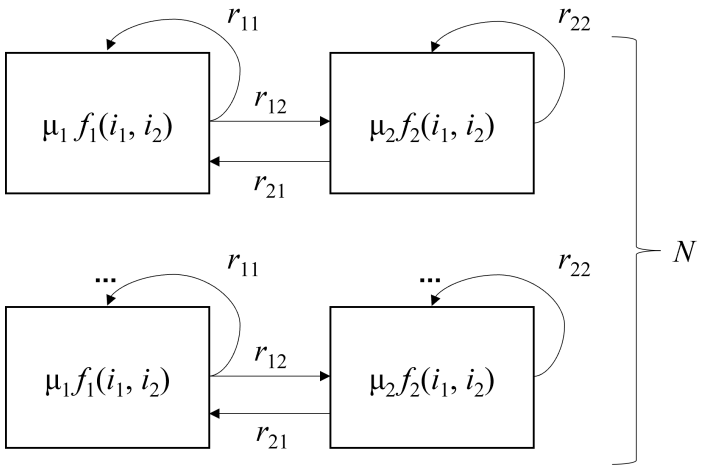

Рис. 1. Замкнутая N-линейная СМО с двумя фазами и деградацией скорости обслуживания

Обозначим процессы $i_1(t)$ и $i_2(t)$ – число заявок, которые характеризуют число виртуальных машин на первой и второй фазе соответственно. Отметим, что для данной модели в любой момент времени $i_1(t) + i_2(t) = N$.

Интенсивность обслуживания на k-той фазе равна $\mu_k f_k(i_1, i_2)$, где $k = 1, 2$ – номер фазы. Здесь μ_k – интенсивность обслуживания заявки без конкуренции на узле, $f_k(i_1, i_2) > 0$ – функция деградации, зависящая от числа заявок на фазах.

Заявка переходит с фазы k_1 на фазу k_2 с вероятностью $r_{k_1 k_2}$, где $k1, k2 = 1, 2$. Условие нормировки для распределений переходов между фазами имеет вид

$$\sum_{k_2=1}^{2} r_{k_1 k_2} = 1, \quad \forall k_1 = 1,\ 2.$$

Целью исследования является нахождение распределения вероятностей числа заявок на первой и второй фазах при наличии деградации скорости обслуживания на обеих фазах.

Для процессов $i_1(t)$ и $i_2(t)$ обозначим двумерное стационарное распределение вероятностей

$$P(i_1, i_2) = P\{i_1(t) = i_1,\ i_2(t) = i_2\} \tag{1}$$

числа заявок на первой и второй фазе.

Обозначим $\mathbf{i}(t) = \{i_1(t),\ i_2(t)\}$, тогда распределение вероятностей (1) перепишем в векторной форме

$$P(\mathbf{i}) = P\{\mathbf{i}(t) = \mathbf{i}\}.$$

Также обозначим векторы $\mathbf{e}_1 = \{1, 0\}$, $\mathbf{e}_2 = \{0, 1\}$, $\mathbf{e} = \{1, 1\}$.

В Таблице 1 показаны обозначения параметров модели.

Таблица 1. Параметры модели

N	Число заявок, циркулирующих между фазами
$\mathbf{R}$	Матрица вероятностей $r_{k_1k_2}$ того, что заявка перейдет в k_2-ю фазу после окончания k_1-й фазы $k_1, k_2 = 1, 2$
$\mu_k f_k(i_1, i_2) = \mu_k f_k(\mathbf{i})$	Интенсивность обслуживания на k-й фазе, $k = 1, 2$
$f_k(i_1, i_2) = f_k(\mathbf{i})$	Функция деградации скорости обслуживания на k-й фазе, $k = 1, 2$

Составим систему уравнений Колмогорова для стационарного распределения вероятностей $P(\mathbf{i})$ двумерного процесса $\mathbf{i}(t)$ в матричной форме

$$-P(\mathbf{i})\left(\sum_{k=1}^{2} i_k\mu_k f_k(\mathbf{i})\right)+\sum_{k_1=1}^{2}\sum_{k_2=1}^{2} P(\mathbf{i}+\mathbf{e}_{k_1}-\mathbf{e}_{k_2})(i_{k_1}+1)\mu_{k_1} f_{k_1}(\mathbf{i}+\mathbf{e}_{k_1}-\mathbf{e}_{k_2})r_{k_1k_2} = 0. \tag{2}$$

Реализуя метод асимптотического анализа, функцию деградации скорости обслуживания $f_k(\mathbf{i})$ на k-ой фазе обозначим $f_k(\mathbf{i}) = \dfrac{1}{N}\widetilde{f}_k\left(\dfrac{\mathbf{i}}{N}\right)$, где функция $\widetilde{f}_k(\mathbf{x})$ – некоторая заданная дифференцируемая функция непрерывного аргумента $\mathbf{x}$, тогда уравнение (2) примет вид

$$-P(\mathbf{i})\left(\sum_{k=1}^{2}\frac{i_k}{N}\mu_k\widetilde{f}_k\left(\frac{\mathbf{i}}{N}\right)\right)+$$

$$+\sum_{k_1=1}^{2}\sum_{k_2=1}^{2} P(\mathbf{i}+\mathbf{e}_{k_1}-\mathbf{e}_{k_2})\frac{i_{k_1}+1}{N}\mu_{k_1}\widetilde{f}_{k_1}\left(\frac{\mathbf{i}+\mathbf{e}_{k_1}-\mathbf{e}_{k_2}}{N}\right)r_{k_1k_2} = 0. \tag{3}$$

Уравнение (3) будем решать методом асимптотического анализа в предельном условии растущего числа заявок в системе ($N \to \infty$).

3. Метод асимптотического анализа

Предельное условие растущего числа заявок в системе определяется бесконечно большим параметром $N \to \infty$. Обозначим и сделаем следующие замены в уравнении (3)

$$\frac{i_k}{N} = \varepsilon i_k \triangleq x_k,\ \frac{\mathbf{i}}{N} = \varepsilon\mathbf{i} \triangleq \mathbf{x},\ P(\mathbf{i}) = P_1(\mathbf{x}, \varepsilon),\ k = 1, 2.$$

Тогда уравнение (3) примет вид

$$-P_1(\mathbf{x}, \varepsilon)\left(\sum_{k=1}^{2}\mu_k x_k\widetilde{f}_k(\mathbf{x})\right)+$$

$$+\sum_{k_1=1}^{2}\sum_{k_2=1}^{2} P_1(\mathbf{x}+\varepsilon\mathbf{e}_{k_1}-\varepsilon\mathbf{e}_{k_2}, \varepsilon)\mu_{k_1}(x_{k_1}+\varepsilon)\widetilde{f}_{k_1}(\mathbf{x}+\varepsilon\mathbf{e}_{k_1}-\varepsilon\mathbf{e}_{k_2})r_{k_1k_2} = 0. \tag{4}$$

Реализуя метод асимптотического анализа в предельном условии растущего числа заявок в системе, решая уравнение (4), на первом этапе получим

$$\kappa_{k_1}\mu_{k_1}\widetilde{f}_{k_1}(\kappa) - \sum_{k_2=1}^{2}\kappa_{k_2}\mu_{k_2}\widetilde{f}_{k_2}(\kappa)r_{k_1k_2} = 0,\ k_1,k_2 = 1,2, \tag{5}$$

определяющее асимптотическое среднее значение $\kappa = \{\kappa_1, \kappa_2\}$ числа заявок в системе на первой и второй фазах соответственно.

На втором этапе реализации метода в уравнении (3) введем обозначение $\varepsilon^2 = \frac{1}{N}$ и сделаем замены

$$\frac{i_k}{N} = \varepsilon^2 i_k \triangleq x_k + \varepsilon y_k,\ \frac{\mathbf{i}}{N} = \varepsilon^2\mathbf{i} \triangleq \mathbf{x} + \varepsilon\mathbf{y},\ P(\mathbf{i}) = P_2(\mathbf{y},\varepsilon),\ k = 1,2,$$

и запишем уравнение второй асимптотики

$$-P_2(\mathbf{y},\varepsilon)\left(\sum_{k=1}^{2}\mu_k(x_k+\varepsilon y_k)\widetilde{f}_k(\mathbf{x}+\varepsilon\mathbf{y})\right) +$$

$$+\sum_{k_1=1}^{2}\sum_{k_2=1}^{2}P_2(\mathbf{y}+\varepsilon\mathbf{e}_{k_1}-\varepsilon\mathbf{e}_{k_2},\varepsilon)\mu_{k_1}(x_{k_1}+\varepsilon(y_{k_1}+\varepsilon))\widetilde{f}_{k_1}(\mathbf{x}+\varepsilon(\mathbf{y}+\varepsilon\mathbf{e}_{k_1}-\varepsilon\mathbf{e}_{k_2}))r_{k_1k_2} = 0. \tag{6}$$

Решая систему (6) в условии растущего числа заявок в системе, получим предельную гауссовскую плотность распределения вероятностей $\mathbf{i}(t)$ числа заявок на первой и второй фазах при наличии деградации скорости обслуживания на обеих фазах

$$p(\mathbf{i}) = \frac{1}{(2\pi)^{k_1}\det\mathbf{K}}\exp\left\{-\frac{1}{2}\sum_{k_1,k_2=1}^{2}K^{-1}_{k_1k_2}(i_{k_1}-\kappa_{k_1})(i_{k_2}-\kappa_{k_2})\right\}. \tag{7}$$

Корреляционные моменты $K_{k_1k_2}$, которые образуют матрицу ковариации $\mathbf{K}$, являются решением системы уравнений

$$\sum_{l=1}^{2}a_{k_1l}K_{k_1l} = b_{k_1},$$

$$\sum_{l=1}^{2}a_{k_1l}K_{k_2l} + b_{k_1k_2} + \sum_{l=1}^{2}a_{k_2l}K_{k_1l} + b_{k_2k_1} = 0, \text{при } k_1 \neq k_2. \tag{8}$$

Здесь величины $b_{k_1k_2}$, b_{k_1} определяются системой

$$b_{k_1} = x_{k_1}\mu_{k_1}\widetilde{f}_{k_1}(\mathbf{x}),$$

$$b_{k_1k_2} = x_{k_1}\mu_{k_1}\widetilde{f}_{k_1}(\mathbf{x})r_{k_1k_2}, k_1 \neq k_2, \tag{9}$$

при $\mathbf{x} = \kappa$. Величины $a_{k_1k_2}$ – системой

$$a_{k_1k_1} = \mu_{k_1}\widetilde{f}_{k_1}(\mathbf{x}) + x_{k_1}\mu_{k_1}\frac{\partial\widetilde{f}_{k_1}(\mathbf{x})}{\partial x_{k_1}} - \mu_{k_1}\widetilde{f}_{k_1}(\mathbf{x})r_{k_1k_1} -$$

$$-\sum_{k_2=1}^{2} x_{k_2}\mu_{k_2}\frac{\partial\widetilde{f}_{k_2}(\mathbf{x})}{\partial x_{k_1}}r_{k_2k_1}, \quad k_1 = k_2,$$

$$a_{k_1l} = x_{k_1}\mu_{k_1}\frac{\partial\widetilde{f}_{k_1}(\mathbf{x})}{\partial x_l} - \mu_l\widetilde{f}_l(\mathbf{x})r_{lk_1} - \sum_{k_2=1}^{2} x_{k_2}\mu_{k_2}\frac{\partial\widetilde{f}_{k_2}(\mathbf{x})}{\partial x_l}r_{k_2k_1}, \quad k_1 \neq k_2, \tag{10}$$

при $\mathbf{x} = \kappa$.

4. Заключение

Предложено исследование замкнутой двухфазной сети массового обслуживания, в которой циркулирует N заявок при наличии деградации скорости обслуживания зависящей от числа заявок в системе. Получена гауссовская плотность распределения вероятностей числа заявок в системе на каждой фазе.

Список литературы

1. Huber N., von Quast M., Brosig F., Hauck M., Kounev S. A method for experimental analysis and modeling of virtualization performance overhead // Cloud Computing and Services Science. 2012. P. 353–370.
2. Bermejo B., Juiz C. A general method for evaluating the overhead when consolidating servers: performance degradation in virtual machines and containers // The Journal of Supercomputing. 2022. Vol. 78, No. 9. P. 11345–11372.
3. Xu F., Liu F., Jin H., Vasilakos A. V. Managing performance overhead of virtual machines in cloud computing: A survey, state of the art, and future directions // Proceedings of the IEEE. 2013. Vol. 102, No. 1. P. 11–31.
4. Liu X., Li S., Tong W. A queuing model considering resources sharing for cloud service performance // The Journal of Supercomputing. 2015. Vol. 71. P. 4042–4055.
5. Fdida S., Mailles D., Pujolle G. Queueing systems with resource sharing // Journal of Systems and Software. 1986. Vol. 6, No. 1–2. P. 23–29.
6. Bai W. H., Xi J. Q., Zhu J. X., Huang S. W. Performance analysis of heterogeneous data centers in cloud computing using a complex queuing model // Mathematical Problems in Engineering. 2015. Vol. 2015. P. 980945.
7. Choudhary A., Chakravarthy S. R., Sharma D. C. Analysis of MAP/PH/1 Queueing System with Degrading Service Rate and Phase Type Vacation // Mathematics. 2021. Vol. 9, No. 19. P. 2387.
8. Goswami V., Patra S. S., Mund G. B. Performance analysis of cloud with queue-dependent virtual machines // Proceedings of the 2012 1st international conference on recent advances in information technology (RAIT). Dhanbad, 15–17 March 2012. Dhanbad: IEEE, 2012. P. 357–362.