МОДЕЛИРОВАНИЕ СИСТЕМЫ МАССОВОГО ОБСЛУЖИВАНИЯ ВЗАИМОДЕЙСТВИЯ ТЕРМИНАЛЬНЫХ УСТРОЙСТВ И СЕРВИСОВ ПОСТАВЩИКОВ УСЛУГ В БАНКЕ

УДК 330.46

Иван Андраникович Мнацаканян,
аспирант, каф. Прикладной Математики Московского государственного университета экономики, статистики и информатики (МЭСИ)
Тел.: (919) 727-93-05
Эл. почта: mc_konyan@mail.ru

В статье основное внимание уделяется разработке математической модели и инструменту для оптимизации работы системы массового обслуживания в банке. В статье рассматриваются математические аспекты, позволяющие добиться перераспределения транзакционного потока, снижения времени пребывания запроса в очереди, что приводит к увеличению доходов банка и получению конкурентных преимуществ.

Ключевые слова: система массового обслуживания, оптимизация, банки, платежные транзакции

Ivan A. Mnatsakanyan,
Post-graduate student, the Department of Applied Mathematics, Moscow State University of Economics, Statistics and Informatics (MESI)
Tel.: (919) 727-93-05
E-mail: mc_konyan@mail.ru

MODELING QUEUING SYSTEM OF INTERACTION BETWEEN TERMINAL DEVICES AND SERVICES PROVIDERS IN THE BANK

The article focuses on the development of mathematical models and tools to optimize the system of queuing at the bank. The article discusses the mathematical aspects that will achieve redistribution of transaction flow, reduce the time of the request in the queue, increase the bank's profit and gain competitive advantage.

Keywords: queuing system, optimization, banks, payment transactions.

1. Введение

В настоящее время в крупном российском банке с каждым днем растет поток платежей от терминалов и банкоматов, автоматизированных систем, позволяющих производить автоплатежи (платежи по расписанию, или зависящие от параметров счета у поставщика услуг), систем, позволяющих удаленно оплачивать услуги банка через интернет. На сегодняшний день через исследуемый банк проходит около 5 млн. платежей. Этот показатель является максимальным среди российских банков. Рост потока платежей оказывает серьезное воздействие на автоматизированные системы банка. В связи с этим возникают проблемы, связанные с производительностью систем банка, взаимодействующих с ними систем поставщиков услуг, возникают сбои и простои в проведении платежей.

Исторически сложилось, что многие системы банка дублируют функционал друг друга и позволяют производить наборы платежей, схожих по составу. Причинами этого явилось то, что системы банка были децентрализованы по территориальному признаку. В каждом территориальном банке существовала собственная система, которая взаимодействовала с требуемым для данного региона набором платежей. Постепенно компании поставщики услуг начали укрупняться, перенося бизнес в другие регионы. В банке появилась задача централизовать все платежи. При централизации банк решил внедрить крупную автоматизированную систему и интегрировать ее с некоторыми системами, существующими в территориальных банках.

В настоящий момент в банке существует централизованная система, принимающая весь поток платежей и отправляющая его на серверы платежей, которые взаимодействуют с поставщиками услуг. Интенсивность потока платежей входящих в автоматизированную систему постоянно возрастает, но с недавнего времени она достигла предельного значения, при котором серверы платежей не могут обслуживать данные потоки. При этом простаивают серверы платежей, внедренных в банке до централизации, которые могли бы эти потоки обслужить. Следствием этого являются возрастающая нагрузка на сервер, отклоняемые платежи, увеличение времени нахождения запросов в системе.

Между банком и поставщиками услуг существуют договора, в которых банк устанавливает комиссию с проведенных платежей этих поставщиков услуг. В случае простоя сервера платежей, работающего с поставщиками услуг, банк теряет прибыль и клиентов, недовольных большим временем ожидания и отказами в проведении платежей.

Целью данной работы является разработка математической модели и инструмента для оптимизации работы автоматизированной системы.

Для достижения данной цели требуется решение следующих задач:

– Описание работы автоматизированной системы

– Оптимизация производительности работы сервера платежей.

– Оценка необходимости подключения дополнительного сервера для обработки потока платежей.

– Определение целесообразности отключения сервера.

2. Описание работы автоматизированной системы

Представим замкнутую систему массового обслуживания, содержащую множество групп запросов от банкоматов, автоматизированных систем, позволяющих производить автоплатежи, автоматизированных систем, позволяющих удаленно оплачивать услуги банка через интернет по разным поставщикам услуг I_j, $j = \overline{1,m}$, каждая из групп запросов характеризуется своей интенсивностью запросов к серверу очередей автоматизированной системы λ_j, $j = \overline{1,m}$. При этом сервер платежей автоматизированной системы P обрабатывает заявки, пришедшие через сервер очередей автоматизированной системы B с общей интенсивностью λ_{sum} с дисциплиной обслуживания FIFO. Закон распределения времени обработки

запросов сервера платежей P в общем случае неизвестен и может быть экспоненциальным (M), равномерным (R), детерминированным (D), нормальным (N), либо произвольным (G), если любая информация отсутствует. Сделаем допущение, что поток заявок является однородным.

3. Оптимизация производительности работы сервера и оценка необходимости подключения дополнительного сервера для обработки потока платежей

Поставим задачу направления на сервер только такого потока запросов, при котором среднее время ожидания в очереди не превысит предельно допустимого значения. Например, по технологическим требованиям время ожидания запроса не должно превышать 30 секунд. Будем решать задачу, описанную в п.2 аналитическими методами.

Рассмотрим вначале общий случай с произвольным законом распределения времени поступления входящих запросов в сервер очередей и произвольным временем обработки этих запросов сервером платежей. Для данной системы класса G/G/1 существуют приближения ее характеристик.

Пусть среднее время между входящими запросами в сервер очередей составляет $E(A_1)$, где A_1 случайная величина времени прихода следующего запроса, дисперсия соответственно $\sigma_{A_1}^2 = E\left(\left(A_1 - E\left(A_1\right)\right)^2\right.$, а коэффициент вариации $c_{A_1} = \sigma_{A_1}\left(E\left(A_1\right)\right)^{-1}$. Для сервера среднее время между обрабатываемыми сервером запросами платежей составляет $E(B_1)$, где B_1 случайная величина времени обработки запроса, дисперсия соответственно $\sigma_{B_1}^2 = E\left(\left(B_1 - E\left(B_1\right)\right)^2\right.$, а коэффициент вариации $c_{B_1} = \sigma_{B_1}\left(E\left(B_1\right)\right)^{-1}$. Пусть λ – интенсивность потока входящих запросов, μ – интенсивность потока обработки запросов.

Тогда существует оценка для среднего времени пребывания в очереди W_q:

$$W_q \approx \frac{\rho}{1-\rho}\cdot\frac{C_{A_1}^2 + C_{B_1}^2}{2}\cdot E\left(B_1\right), \quad (1)$$

где $\rho = \frac{\lambda}{\mu}$.

Для системы M/M/1 по формуле (1) получим:

$$W_q = \rho\cdot(1-\rho)^{-1}\cdot E(B_1) \quad (2)$$

Для M/G/l:

$$W_q \approx \rho\cdot(1+c_{B_1}^2)\cdot(2\cdot(1-\rho))^{-1}\cdot E(B_1) \quad (3)$$

Для G/M/1:

$$W_q \approx \rho\cdot(1+c_{A_1}^2)\cdot(2\cdot(1-\rho))^{-1}\cdot E(B_1) \quad (4)$$

Используя (1), получим среднее время обработки запроса в системе:

$$W \approx \frac{\rho}{1-\rho}\cdot\frac{c_{A_1}^2 + c_{B_1}^2}{2}\cdot E\left(B_1\right) + E\left(B_1\right) \quad (5)$$

Среднее количество заявок в очереди составит:

$$L_q = W_q\cdot(E(A_1))^{-1} \quad (6)$$

Среднее количество заявок в системе:

$$L = W\cdot(E(A_1))^{-1} \quad (7)$$

Если среднее время не превышает максимально допустимого, то привлечение нового, дополнительного сервера платежей для обработки потока не требуется. Если среднее время превышает максимально допустимое, возникает вопрос о выделении части потока запросов и перенаправления его на другой дополнительный сервер платежей.

На основе достижения текущим сервером заданного среднего времени обработки входящих запросов $T_{\max}$ определим необходимость подключения дополнительного сервера платежей для обработки потока запросов. Для этого вначале найдем максимальную интенсивность потока входящих запросов λ_{base}, которую текущий сервер платежей сможет обработать, удовлетворяя требованиям по среднему времени обработки запросов $T_{\max}$.

Исходя из формулы для среднего времени обработки запроса (4), устанавливаем ограничения по времени обслуживания входящих запросов, и решаем его относительно интенсивности:

$$\frac{\rho}{1-\rho}\cdot\frac{c_{A_1}^2 + c_{B_1}^2}{2}\cdot E\left(B_1\right) + E\left(B_1\right) \le T_{\max} \quad (8)$$

Определим интенсивность потока входящих запросов с учетом $\rho = \frac{\lambda}{\mu}$:

$$\lambda_{base} \le \frac{2\cdot T_{\max}\cdot\mu_{base} - 2\cdot E\left(B_1\right)\cdot\mu_{base}}{E\left(B_1\right)\left(c_{A_1}^2 + c_{B_1}^2\right) + 2\cdot T_{\max} - 2\cdot E\left(B_1\right)}. \quad (9)$$

Найдем максимальную интенсивность потока входящих запросов:

$$\lambda_{base} = \frac{2\cdot T_{\max}\cdot\mu_{base} - 2\cdot E\left(B_1\right)\cdot\mu_{base}}{E\left(B_1\right)\left(c_{A_1}^2 + c_{B_1}^2\right) + 2\cdot T_{\max} - 2\cdot E\left(B_1\right)}. \quad (10)$$

Коэффициенты вариации c_{A_1} и c_{B_1} должны устанавливаться для каждого конкретного закона, с соответствующей трансформацией уравнения (10). На дополнительный сервер перенаправляется оставшаяся интенсивность потока запросов λ_{new}:

$$\lambda_{new} = F(\lambda - \lambda_{base}) \quad (11)$$

где F – в общем случае функция, зависящая от разности исходной интенсивности входящего потока λ и интенсивности потока входящих запросов λ_{base}.

Для практических задач можно считать, что всякий поток, образующийся из любых нескольких независимых ординарных потоков, является простейшим, причем интенсивности суммируются. Соответственно разницу потоков (на основе свойства устойчивости) можно найти:

$$\lambda_{new} = \lambda - \lambda_{base} \quad (12)$$

Преобразуя уравнения для ограничения по быстродействию для дополнительного сервера платежей (4), найдем требуемую интенсивность обработки запросов на нем:

$$\frac{\rho}{1-\rho}\cdot\frac{c_{A_2}^2 + c_{B_2}^2}{2}\cdot E\left(B_2\right) + E\left(B_2\right) \le T_{\max r} \quad (13)$$

Определим требуемую интенсивность потока обработки запросов с учетом $\rho = \frac{\lambda}{\mu}$.:

$$\mu_{1,2} \ge \frac{1 + T_{\max r}\times\lambda_{new} \pm \sqrt{\left(1 + T_{\max r}\cdot\lambda_{new}\right)^2 - T_{\max r}\times\left(2\lambda_{new} - \lambda_{new}\left(c_{A_2}^2 + c_{B_2}^2\right)\right)}}{T_{\max r}}, \quad (14)$$

где $\mu_{1,2}$ – решение уравнения (13).

Установим требуемую интенсивность обработки запросов μ для дополнительного сервера платежей на минимальном уровне:

$$\mu_{\min} = \frac{1 + T_{\max r}\times\lambda_{new} + \sqrt{\left(1 + T_{\max r}\times\lambda_{new}\right)^2 - T_{\max r}\left(2\lambda_{new} - \lambda_{new}\left(c_{A_2}^2 + c_{B_2}^2\right)\right)}}{T_{\max r}}. \quad (15)$$

Можно отметить, что для конкретных законов генерации и обработки запросов, максимально допустимая

интенсивность поступающих запросов для исходного сервера платежей и требуемая интенсивность обработки запросов для дополнительного сервера платежей трансформируются в более индивидуальные значения. В частности для детерминированного и экспоненциального закона данные формулы (10) и (15) преобразуются более простые, так как их коэффициенты вариации равны соответственно 0 и 1.

Рассмотрим систему R/M/1, содержащую коэффициент вариации, зависящий от интенсивности запросов, что приводит к другому ходу решения. Для данной системы, принадлежащей общему классу G/M/1, уравнение трансформируется:

$$W_q \approx \frac{\rho}{1-\rho}\cdot\frac{c_{A_1}^2+1}{2}\cdot E(B_1)+E(B_1) \quad (16)$$

Для равномерно распределенной на отрезке $[a, b]$ случайной величины дисперсия равна $\sigma_{A_1}^2 = (b-a)^2\cdot(12)^{-1}$, математическое ожидание равно $E(A_1) = 0{,}5\cdot(a+b) = \lambda^{-1}$. Таким образом, квадрат коэффициента вариации:

$$c_{A_1}^2 = \frac{\sigma_{A_1}^2}{E^2(A_1)} = \frac{(b-a)^2}{3(a+b)^2}. \quad (17)$$

Учитывая экспоненциальное распределение времени обработки у сервера, получаем: $E(B_1) = \mu^{-1}$.

Таким образом, получим общую формулу среднего времени ожидания обработки запроса для системы R/M/1:

$$W_q \approx \frac{\rho}{1-\rho}\cdot\frac{\frac{(b-a)^2}{3(a+b)^2}+1}{2}\cdot\mu^{-1}+\mu^{-1} \quad (18)$$

Найдем максимальную интенсивность потока входящих запросов λ_{base}, которую текущий сервер платежей сможет обработать, удовлетворяя требованиям по среднему времени обработки запросов $T_{\max}$, решив неравенство (19):

$$\frac{\rho}{1-\rho}\cdot\frac{\frac{(b-a)^2}{3(a+b)^2}+1}{2}\cdot\mu^{-1}+\mu^{-1}\le T_{\max} \quad (19)$$

Обозначив интервал существования равномерной функции $\Delta = b-a$, и преобразуя неравенство, получаем:

$$\lambda^2(24\mu T_{\max}-12)+\lambda(24\mu-24\mu^2 T_{\max})+\Delta^2\le 0 \quad (20)$$

Отсюда, находим:

$$\lambda_{1,2base} = \frac{24\mu_{base}^2 T_{\max}-24\mu_{base}\pm\sqrt{\left(24\mu_{base}-24\mu_{base}^2 T_{\max}\right)^2-4\cdot(24\mu_{base}T_{\max}-12)\Delta^2}}{2\cdot(24\mu_{base}T_{\max}-12)}. \quad (21)$$

где $\lambda_{1,2base}$ – решение уравнения (21).

При этом, одним из вариантов решения может быть установка меры отклонения в зависимости от интенсивности: $\Delta = k\cdot\lambda$, где k – коэффициент отклонения:

$$\lambda_{base}^2(24\mu_{base}T_{\max}-12+k^2)+k_s(24\mu_{base}-24\mu_{base}^2T_{\max})\le 0 \quad (22)$$

Откуда, учитывая $\lambda_{base} > 0$, получаем:

$$\lambda_{base} = \frac{24\mu_{base}-24\mu_{base}^2T_{\max}}{12-24\mu_{base}T_{\max}-k^2}. \quad (23)$$

Определим требуемую интенсивность обработки запросов для дополнительного сервера платежей:

$$\mu_{new}^2(-24\lambda_{new}T_{\max})+\mu_{new}\times(24\lambda_{new}+24\lambda_{new}^2T_{max})-12\lambda_{new}^2+\Delta^2\le 0 \quad (24)$$

Решив неравенство (24), получим требуемую интенсивность обработки потока запросов обработки для дополнительного сервера платежей:

$$\mu_{1.2} = \frac{-\left(24\lambda_{new}+24\lambda_{new}^2T_{\max}\right)\pm\sqrt{\left(24\lambda_{new}+24\lambda_{new}^2T_{\max}\right)^2-4\left(-24\lambda_{new}T_{\max}\right)\times\left(-12\lambda_{new}^2+\Delta^2\right)}}{2\cdot\left(-24\lambda_{new}T_{\max}\right)}. \quad (25)$$

С учетом $\Delta = k\cdot\lambda$,:

$$\mu_{new}^2(-24\lambda_{new}T_{\max})+\mu_{new}\times\times(24\lambda_{new}+24\lambda_{new}^2T_{\max})-12\lambda_{new}^2+(k\cdot\lambda)^2\le 0 \quad (26)$$

$$\mu_{1.2} = \frac{-\left(24\lambda_{new}+24\lambda_{new}^2T_{\max}\right)\pm\sqrt{\left(24\lambda_{new}+24\lambda_{new}^2T_{\max}\right)^2-4\left(-24\lambda_{new}T_{\max}\right)\times\left(-12\lambda_{new}^2+(k\cdot\lambda)^2\right)}}{2\cdot\left(-24\lambda_{new}T_{\max}\right)}. \quad (27)$$

4. Оценка целесообразности отключения сервера

Решение, представленное в п.3, можно использовать и для решения обратной задачи – оценки определения целесообразности отключения сервера платежей и направления его потока запросов на один из оставшихся серверов.

Предположим, есть два отдельных сервера платежей, обрабатывающих идентичные потоки запросов. По некоторым причинам нужно отключить один из серверов платежей от автоматизированной системы, при этом до принятия решения целесообразно проверить, справится ли оставшийся сервер платежей с увеличившимся потоком запросов.

Предположим, что отключаемый сервер платежей обрабатывает поток заявок с интенсивностью μ_{del}, а интенсивность потока входящих заявок, направленная на этот сервер – λ_{del}. Аналогично, оставшийся сервер платежей обрабатывает поток заявок с интенсивностью μ_{cur}, а интенсивность потока входящих заявок, направленная на этот сервер – λ_{cur}.

В общем случае суммировать интенсивности потоков λ_{cur} и λ_{del} недопустимо, но для практических задач можно считать, что всякий поток, образующийся из любых нескольких независимых ординарных потоков, является простейшим, причем интенсивности суммируются.

Отсюда найдем сумму интенсивностей потоков λ_{cur} и λ_{del}:

$$\lambda_{res} = \lambda_{cur}+\lambda_{del} \quad (28)$$

Используя формулу (15), получим:

$$\mu_{res} = \frac{2+2\cdot T_{\max}\cdot\lambda_{res}\pm\sqrt{4\left(1+T_{nax}\lambda_{res}\right)^2-8T_{\max}\left(2\lambda_{res}-\lambda_{res}\left(c_{A_2}^2+c_{B_2}^2\right)\right)}}{4T_{\max}}. \quad (29)$$

Сравнивая требуемую интенсивность работы μ_{res} с μ_{cur}, можно принять решение о допустимости направления на данный сервер дополнительного потока запросов.

5. Заключение

Используя данную математическую модель можно добиться следующих результатов:

Получить инструмент для распределения потока запросов на серверы платежей;

Достичь заданное минимальное временя ожидания запроса в очереди;

Обеспечить экономию при отключении сервера платежей, в случае, когда остальные успевают обрабатывать поток запросов;

Добиться перераспределения за-

просов в случае проведения плановых работ или технологического сбоя;

Представленные модели могут быть доработаны для учета индивидуальных особенностей серверов, в частности, можно построить модели класса G/G/M с индивидуальными приоритетами внутри очередей.

После внедрения данной системы в банке можно будет подсчитать экономические эффекты:

– Прямой экономический эффект от внедрения системы будет влиять на увеличение доходов банка с комиссий проведенных платежей за счет снижения количества сбойных транзакций и снижения среднего времени пребывания транзакции в очереди.

– Косвенный экономический эффект от внедрения системы будет проявляться в увеличении числа клиентов за счет повышения конкурентоспособности банка, в снижении трудозатрат расчетного центра и технической поддержки на обслуживание клиентских запросов, поиск сбойных транзакций, а также в снижении расходов при отключении сервера.

– Снижения риска ухода клиента в другие банки.

Литература

1. Leonard Kleinrock Queueing Systems: Volume II – Computer Applications. // New York: Wiley Interscience, 1976.

2. Клейнрок Л. Вычислительные системы с очередями// изд. «Машиностроение» 1979 г.

3. Кухарев, В.Н. Анализ существующих методов и алгоритмов проектирования информационных систем и их потоков данных // Юж.-Рос. гос. техн. ун-т., 2005. – 31с., №1480 2005г.

4. Горемыкина Г.И., Мастяева И.Н. Моделирование системы управления качеством корпоративной информационно-вычислительной сети // Журнал «Национальные интересы: приоритеты и безопасность» 2013 г.

5. Снегова Е.Г., Мастяева И.Н.Сети социальных связей как инструмент моделирования рисков мошенничества в экспресс-кредитовании // Журнал «Финансы и кредит» 2013.

6. Стрелков С.В. Мастяева И.Н. Распределение рискового капитала на основе кооперативных игр. // «Экономика и математические методы», ЦЭМИ РАН 2013.

References

1. Leonard Kleinrock Queueing Systems: Volume II – Computer Applications. // New York: Wiley Interscience, 1976.

2. Kleinrock L. Queueing Systems// izd. «Mashinostroenie» 1979 g.

3. Kukhrev V.N. Analysis of existing methods and algorithms for designing information systems and their data streams// Yuzh.-Ros. gos. tehn. un-t., 2005. – 31s., №1480 2005g.

4. Goremykina G.I., Mastyaeva I.N. Modeling system of quality management of corporate information network / / Zhurnal «Nacionalnye interesy: prioritety i bezopasnost» 2013g.

5. Snegova E.G., Mastyaeva I.N. Social networks as a tool for modeling the risks of fraud in the express loans / / Zhurnal «Finansy i kredit» 2013.

6. Strelkov S.V., Mastyaeva I.N. The distribution of risk capital on the basis of cooperative games. // «Ekonomika i matematicheskie metody», CEMI RAN 2013.