ИССЛЕДОВАНИЕ МЕТОДА «НАПРАВЛЯЮЩИХ ГИПЕРБОЛ» ДЛЯ ОПТИМИЗАЦИИ ОДНОРОДНЫХ НЕМАРКОВСКИХ СЕТЕЙ С ОГРАНИЧЕННЫМИ РАЗМЕРАМИ БУФЕРОВ

Ю.Г. Галич (Омск)

Формализованным представлением сложных организационно-технических систем, таких как, производственные, транспортные или компьютерные сети, которые предназначены для обслуживания дискретных потоков однотипных заявок, выступают сети массового обслуживания (СеМО). Узлами сети являются системы массового обслуживания (СМО), которые обслуживают заявки, и передают их, в соответствии с заданными переходными вероятностями, на входы других СМО или на выход из сети. Часто производительность СеМО оценивают по времени прохождения через нее заявки. Стационарное среднее время E прохождения заявки через СеМО зависит от распределения ресурса сети между ее узлами, поэтому актуальна задача оптимизации этого распределения.

Методы оптимизации марковских сетей успешно разрабатываются в [1–4]. Существуют различные попытки оптимизации немаркоских сетей [5–7]. В работе [8] предложен эффективный аналитико-имитационный метод «направляющих гипербол» (НГ) для оптимизации немарковских СеМО, содержащих десятки и сотни СМО с бесконечными размерами буферов, который характеризуется хорошей точностью и приемлемой вычислительной трудоемкостью.

Решение задачи оптимального распределения ресурса по узлам открытой сети, в каждом из которых существует ненулевая вероятность потерь заявок, обусловленная ограниченными размерами буферов, а также ограничена суммарная вероятность потерь, определяет актуальность исследования возможности применения для таких сетей метода НГ.

В работе [1] изложена формулировка задачи оптимизации однородной замкнутой экспоненциальной сети, ее обобщенная версия для немарковской сети – в работе [8]. В условие обобщенной задачи для открытой немарковской сети добавим ограничения: размеры буферов каждого узла не превосходят N, а суммарная вероятность потерь заявок – Q. Сформулируем эту задачу, изложим основные этапы метода НГ и проблемы, которые возникли в процессе нашего исследования.

В открытую однородную немарковскую сеть поступает рекуррентный поток заявок с интенсивностью Λ. Интервалы времени поступления заявок – независимые случайные величины с функцией распределения $A(t)$. Заявка из входного потока сети с вероятностью p_{0i} попадает в i-й узел или с вероятностью Q_i покидает сеть, если буфер i-го узла переполнен, $i=\overline{1,n}$. В любом из K_i каналов i-го узла время обслуживания заявки также независимая случайная величина с функцией распределения $B_i(t)$. После обслуживания в i-м узле заявка случайно и независимо, в соответствии с заданными переходными вероятностями p_{ij}, выбирает один из узлов j для продолжения своего маршрута. Если буфер j-го узла переполнен, то заявка покидает сеть. Вероятности p_{ij}, $i,j=\overline{0,n}$ задаются неразложимой стохастической матрицей $\mathrm{P}=\left\|p_{ij}\right\|$.

Ресурс M сети задается как функция вектор интенсивностей $\vec{\mu}=(\mu_1,\ldots,\mu_n)$ обслуживания в узлах $i=\overline{1,n}$:

$$M(\vec{\mu})=\sum_{i=1}^{n} c_i\mu_i, \tag{1}$$

где c_i – стоимостные коэффициенты.

Требуется найти вектор $\vec{\mu} = \vec{\mu}_{\text{opt}}$, принадлежащий области допустимых решений (ОДР):

$$M(\vec{\mu}) = \sum_{i=1}^{n} c_i \mu_i, = M^*, \quad \mu_i \geq \mu_{i\min}, \quad i = \overline{1,n}, \tag{2}$$

где $\mu_{i\ \min} = \lambda_i / K_i$ (граница области стационарности для СМО с бесконечными размерами буферов), доставляющий минимум функции $E = E(\vec{\mu})$:

$$E(\vec{\mu}) = \sum_{i=1}^{n} \alpha_i \left(w_i(\vec{\mu}) + \frac{1}{\mu_i} \right) \to \min_{\vec{\mu}}, \tag{3}$$

где α_i – среднее число посещений i-го узла заявкой за время ее прохождения через сеть;

w_i – среднее время ожидания заявки в очереди i-го узла.

Для СМО с бесконечными размерами буферов коэффициенты α_i однозначно определяются из системы уравнений баланса [8]. С учетом добавленных в условие задачи ограничений эта система уравнений принимает следующий вид:

$$\alpha_i = \sum_{j=0}^{n} \alpha_j p_{ij} - Q_i, \quad i = \overline{0,n}, \quad \alpha_0 = 1, \tag{4}$$

$$\sum_{i=1}^{n} Q_i = Q$$

где .

Через коэффициенты α_i последовательно определяются интенсивности $\lambda_i = \Lambda\alpha_i$ входных потоков узлов, их коэффициенты загрузки $\rho_i = \lambda_i/(\mu_i K_i)$ и проверяются условия стационарности $\rho_i \leq 1$, $i = \overline{1,n}$. Значения w_i в формуле (3) определяются посредством имитационного моделирования (ИМ).

Метод НГ решения оптимизационной задачи представляет собой ускоренный градиентный поиск точки $\vec{\mu}_{\text{opt}}$, использующий ИМ сети и сепарабельную аппроксимацию целевой функции [8].

Начальный этап. Задаем число итераций $K > 2$, две точки $\vec{\mu}^1 = \vec{\mu}_c$ и $\vec{\mu}^2 \neq \vec{\mu}_c$, принадлежащие ОДР. Координаты центра $\vec{\mu}_c$ ОДР для СМО с бесконечными размерами буферов, определяемого условием равной загрузки узлов, вычисляем по формулам:

$$\mu_i = M^* (\alpha_i / K_i) \left(\sum_{j=1}^{n} c_j \alpha_j / K_j \right)^{-1}, \quad i = \overline{1,n}. \tag{5}$$

С помощью ИМ в этих точках вычисляем оценки среднего времени $\hat{E}^1$ и $\hat{E}^2$, и, соответственно оценки среднего времени ожидания $\left(\hat{w}_1^1, \ldots, \hat{w}_n^1\right)$, $\left(\hat{w}_1^2, \ldots, \hat{w}_n^2\right)$ в узлах $i = \overline{1,n}$. Полагаем $k = 2$.

Основной цикл.

1. Используя оценки $\left(\hat{w}_1^1,\ldots,\hat{w}_n^1\right)$, $\left(\hat{w}_1^2,\ldots,\hat{w}_n^2\right)$ по формулам (6) находим коэффициенты R_i и S_i, $i=\overline{1,n}$, аппроксимации $E^{ap}(\vec{\mu})$:

$$S_i=\frac{\hat{w}_i^k\mu_i^k-\hat{w}_i^{k-1}\mu_i^{k-1}}{\hat{w}_i^k-\hat{w}_i^{k-1}},\quad R_i=\hat{w}_i^{k-1}\left(\mu_i^{k-1}-S_i\right),\quad i=\overline{1,n}, \tag{6}$$

Аппроксимация $E^{ap}(\vec{\mu})$ целевой функции $E(\vec{\mu})$ представляет собой сепарабельную функцию варьируемых переменных μ_i:

$$E^{ap}(\vec{\mu})=\sum_{i=1}^{n}\alpha_i\left(W_i(\vec{\mu})+\frac{1}{\mu_i}\right),$$

где
$$W_i(\vec{\mu}_i)=\begin{cases}\dfrac{R_i}{\mu_i^k-S_i},\ \text{если } \hat{w}_i^k\neq\hat{w}_i^{k-1},\\ \hat{w}_i^k,\ \text{если } \hat{w}_i^k=\hat{w}_i^{k-1}.\end{cases} \tag{7}$$

Вычисляем градиент функции E^{ap} по формуле (8):

$$\nabla E^{ap}(\vec{\mu}^k)=\left(\alpha_1\frac{\partial W_1}{\partial\mu_1}-\frac{\alpha_1}{(\mu_1)^2},\ldots,\alpha_n\frac{\partial W_n}{\partial\mu_n}-\frac{\alpha_n}{(\mu_n)^2}\right), \tag{8}$$

где
$$\frac{\partial W_i}{\partial\mu_i}=\begin{cases}\dfrac{-R_i}{\left(\mu_i^k-S_i\right)^2},\ \text{если } \hat{w}_i^k\neq\hat{w}_i^{k-1},\\ 0,\ \text{если } \hat{w}_i^k=\hat{w}_i^{k-1},\end{cases}\qquad i=\overline{1,n}.$$

Направление $-\nabla E^{ap}(\vec{\mu}^k)$ наискорейшего убывания функции $E^{ap}(\vec{\mu})$ проецируем на гиперплоскость ограничений (2), и проекция L на нее направления антиградиента есть направление вектора $\vec{e}=-\nabla E^{ap}(\vec{\mu}^k)+\vec{n}(\vec{n}\nabla E^{ap}(\vec{\mu}^k))$, где $\vec{n}=\vec{c}/|\vec{c}|$ нормаль к гиперплоскости ограничений, $\vec{c}=(c_1,\ldots,c_n)$ – вектор стоимостных коэффициентов. Валидная часть $[L]$ проекции L ограничена точками $\vec{\mu}^k$ и $\vec{\mu}=\vec{\mu}^k+h\vec{e}$, где $h=\min\{h_1,h_2\}$,

$$h_1=\min\left\{h_{1i}:h_{1i}>0;i=\overline{1,n}\right\},$$
$$h_2=\min\left\{h_{2i}:h_{2i}>0;i=\overline{1,n}\right\},$$
$$h_{1i}=-(\mu_i^k-\mu_{i\min})/e_i,\quad h_{2i}=-(\mu_i^k-S_i)/e_i,\quad i=\overline{1,n}.$$

Проекцию L направления антиградиента строим пошагово, как исходящую из точки $\vec{\mu}^k$ ломаную, точки $\vec{\mu}$ которой есть проекции на гиперплоскость (2) равноотстоящих с малым шагом $D\cdot10^{-3}$, где диаметр D ОДР – длина максимального из диапазонов варьирования переменных μ_i: $D=\max\{l_i\}$, где $l_i=\mu_{i\max}-\mu_{i\min}$ и

$$\mu_{i\max}=\left(M^*-\sum_{j\neq i}^{n}c_j\mu_{i\min}\right)c_i^{-1},\quad i=\overline{1,n}.$$

Для каждой очередной точки $\vec{\mu}$ проверяем условия валидности $\mu_i > \mu_{i\min}$ и $(\mu_i - S_i)(\mu_i^k - S_i) > 0, i = \overline{1,n}.$ Построение [L] завершается получением и отбрасыванием точки $\vec{\mu},$ которая не удовлетворяет условиям валидности или не имеет проекции на гиперплоскость ограничений (2)

2) В качестве следующей точки $\vec{\mu}^{k+1}$ выбираем решение задачи одномерной оптимизации $E^{ap}(\vec{\mu}) \to \min, \vec{\mu} \in [L].$

3) Полагаем $k = k + 1$. С помощью ИМ вычисляем оценки $\left(\hat{w}_1^k, \ldots, \hat{w}_n^k\right)$ и $\hat{E}^k$. Если $k < K$, то переходим к шагу 1, иначе – к шагу 4.

4) Точку $\vec{\mu}^* \in \left\{\vec{\mu}^1, \ldots, \vec{\mu}^K\right\}$ с оценкой $\hat{E}\left(\vec{\mu}^*\right) = \min\left\{\hat{E}^1, \ldots, \hat{E}^K\right\}$ принимаем в качестве приближенного решения задачи. Конец алгоритма.

В [8] испытание метода НГ показано, в том числе, на примере тестовой СеМО-1. Суммарный ресурс $M = 30$ распределяется при $\vec{c} = (1, 2, 1, 1, 1, 1, 1, 3, 1)$, т.е. $c_i = K_i$. Типы распределений $B_i(t)$ для узлов $i = \overline{1,9}$ определены как R, R, R, M, M, E^2, E^2, E^2, R соответственно, где M – экспоненциальное распределение, R – равномерное, E^2 – эрланговское распределение второго порядка. Входящий поток сети пуассоновский и имеет интенсивность $\Lambda = 1$. Переходные вероятности равны соответственно $p_{0,1} = 0{,}2$, $p_{0,2} = 0{,}3$, $p_{0,3} = 0{,}5$, $p_{2,4} = 0{,}7$, $p_{2,5} = 0{,}3$, $p_{4,6} = 0{,}3$, $p_{4,7} = 0{,}4$, $p_{4,9} = 0{,}3$, $p_{5,8} = 0{,}9$, $p_{5,9} = 0{,}1$. Решение, определяемое за 7–11 итераций, обеспечивает среднее время прохождения заявки через сеть $E \approx 7{,}49...7{,}51$. Однако в работе не приводятся точки $\vec{\mu} = \vec{\mu}_{\text{opt}}$, доставляющие минимум функции $E = E(\vec{\mu})$, поэтому перед началом испытания метода НГ на сетях с ограниченными размерами буферов каждого узла метод НГ был испытан на тестовой СеМО-1 из [8]. Получен был близкий результат $E = 7{,}514$, который достигается в точке $\vec{\mu}$ = (0,787; 0,616; 1,407; 3,117; 9,683; 1,292; 1,509; 2,875; 2,348).

На этапе формулировки задачи перед нами возникают следующие проблемы:

– необходимо ли условие (2) для СМО с ограниченными размерами буферов, если условия стационарности всегда выполняется?

– как решить систему (4), если вероятности потерь заявок в i-ом узле неизвестны?

– является ли целевая функция (3) в этом случае выпуклой?

При наличии потерь формула (5) неверна, и это порождает еще одну проблему – какую точку выбрать в качестве начальной точки оптимизации?

Обозначенные проблемы требуют дальнейшего исследования.

В настоящем исследовании не решаются все возникшие проблемы, а тестируется сам алгоритм метода НГ для СеМО с ограниченными размерами буферов. Необходимость условия (2) в нашем случае обусловлена тем, что если размеры буферов превышают средние длины очередей СМО, то условие стационарности может нарушаться. Таким же соображением мы руководствовались при выборе начальной точки оптимизации, определяемой условием равной загрузки узлов, как в первоисточнике. Мы сохранили конфигурацию СеМО-1 (рис. 1), типы распределений, переходные вероятности, добавили ограничения N и Q и получили следующие результаты. Для значений N в диапазоне от 7 до 10 суммарная вероятность потерь не

превышала 1%, колеблясь от 0,5% до 0,9% при $N = 7$ и от 0,07% до 0,2% при $N = 10$, на каждом шаге метода НГ.

Решение задачи оптимального распределения ресурса сети методом НГ также за 6-11 итераций дает среднее время прохождения заявки через сеть близкое к 7,5, что соответствует результату испытания сети с неограниченными длинами очередей. Так, при $N = 7$ получили $E = 7{,}414$, которое достигается в точке $\vec{\mu} = (0{,}760;\ 0{,}537;\ 1{,}248;\ 3{,}095;\ 10{,}690;\ 1{,}302;\ 1{,}436;\ 2{,}757;\ 2{,}124)$; при $N = 8$ получили $E = 7{,}453$ в точке $\vec{\mu} = (0{,}978;\ 0{,}485;\ 1{,}429;\ 2{,}929;\ 10{,}229;\ 1{,}249;\ 1{,}699;\ 2{,}757;\ 2{,}246)$; при $N = 10$ получили $E = 7{,}503$, в точке $\vec{\mu} = (0{,}717;\ 0{,}580;\ 1{,}405;\ 2{,}902;\ 10{,}142;\ 1{,}278;\ 1{,}500;\ 2{,}893;\ 2{,}218)$.

С уменьшением значения N увеличивается значение Q. В случае $N = 6$ суммарная вероятность потерь бала чуть более 1%, а при $N = 3$ достигала 15%, и за 30 шагов так и не произошло «отбрасывания» очередного приближения $\vec{\mu}^{k+1}$ от искомой точки оптимума.

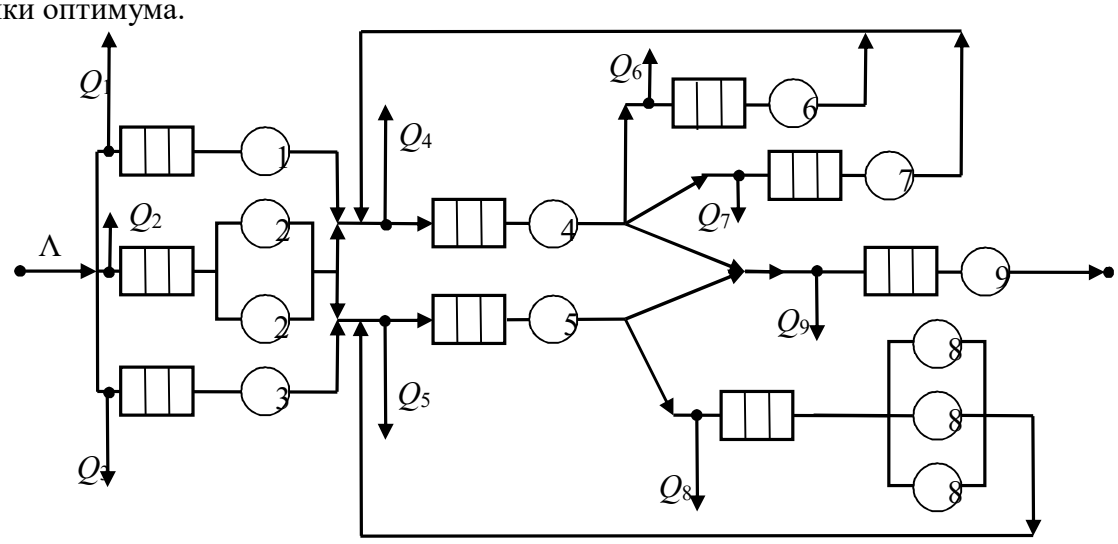

Рис. 1. Тестовая СеМО-1 с учетом введенных ограничений

Можно сказать, что при $N = 6$ и $N = 5$, когда суммарная вероятность потерь не превышала 4%, метод работает. Так, при $N = 5$ за 11 итераций получили среднее время прохождения заявки через сеть $E = 6{,}788$ в точке $\vec{\mu} = (0{,}892;\ 0{,}566;\ 1{,}439;\ 3{,}016;\ 9{,}742;\ 1{,}351;\ 1{,}644;\ 2{,}815;\ 2{,}284)$. На рис. 2 приведена часть траектории изменения целевой функции в процессе оптимизации методом НГ при различных значениях N.

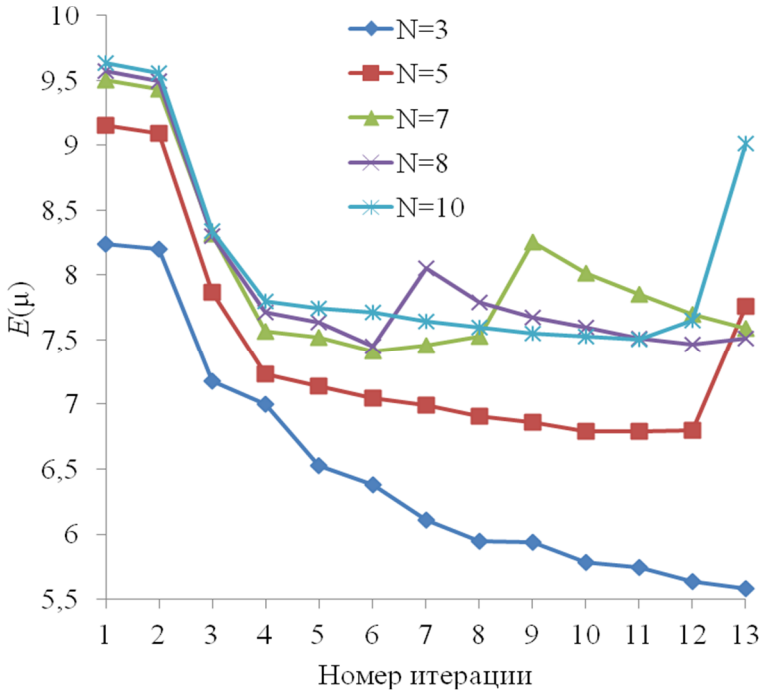

Рис. 2. Изменение E в процессе оптимизации при различных значениях N

При решении задачи оптимального распределения ресурса по узлам открытой однородной сети, в каждом из которых существует ненулевая вероятность потерь заявок, а суммарная вероятность потерь ограничена, мы использовали градиентный аналитико-имитационный метод НГ. Результаты экспериментов показывают, что, если суммарная вероятность потерь не превышает 1%, его применение для таких сетей оправдано и дает хорошую точность при малых затратах машинного времени. Если суммарная вероятность потерь не превышает 4% и это допустимо в каких-то технических системах, то метод НГ тоже работает. Это позволяет рекомендовать его для практического применения при проектировании и модернизации СеМО с ограниченными размерами буферов. Вместе с тем, с увеличением значений Q уменьшение среднего времени прохождения заявки через сеть происходит в основном за счет потерь заявок, и использование метода НГ для решения задачи оптимизации теряет смысл.

Литература

1. **Вишневский В.М.** Теоретические основы проектирования компьютерных сетей. М.: Техносфера, 2003. 512 с.
2. **Клейнрок Л.** Вычислительные системы с очередями / пер. с англ. под ред. Б. С. Цыбакова. М.: Мир, 1979. 600 с.
3. **Клейнрок Л.** Теория массового обслуживания – М.: Машиностроение, 1979. 432 с.
4. **Рыжиков Ю.И.** Имитационное моделирование. Теория и технологии. – СПб.: КОРОНА принт; М.: Альтекс-А, 2004. 384 с.
5. **Задорожный В.Н.** Распределение каналов в однородных немарковских сетях с очередями // Омский научный вестник, 2010. № 1(87). С. 5-10.
6. **Задорожный В.Н.** Полуаналитические методы оптимизации транспортных сетей // Вестник кибернетики. 2018. № 4 (32). С. 16-28.
7. **Gabriel R. Bitran, Reinaldo Morabito.** Open Queueing Networks: Optimization and Performance Evaluation Models for Discrete Manufacturing Systems / Сайт

Массачусетского технологического института. – URL: http:// dspace.mit.edu/bitstream/handle/1721.1/2537/SWP-3743- 31904719.pdf?sequence=1.

8. **Задорожный В.Н.** Оптимизация однородных немарковских сетей массового обслуживания // Проблемы управления, 2009. № 6. С. 68-75.