Информационные процессы, Том 5, № 3, 2005, стр. 194–200.

Динамическое управление приоритетами при дифференцированном обслуживании абонентов в фиксированной инфраструктуре подвижной сети

Н.Е. Богомолова, Я.В. Чернушевич

МТУСИ

Поступила в редколлегию 06, 06, 2005

Аннотация—Рассматривается метод повышения пропускной способности фиксированной инфраструктуры мобильной сети, основанный на дифференцированном обслуживании абонентов. На основе анализа действующих тарифных планов предложено выделять категории абонентов в мобильных сетях. Построена математическая модель дифференцированного обслуживания потоков заявок категорий на звене информационной сети, при этом через установление порога доступа в сеть абонентов низкой категории учтена загрузка сигнальной сети. Также приводится приближенная модель, обеспечивающая приемлемую точность расчета характеристик. На основании приближенной модели приведены расчеты, показывающие эффективность дифференцированного обслуживания.

1. ВВЕДЕНИЕ

При растущих объемах информационной и сигнальной нагрузки в цифровых мобильных сетях второго поколения и с учетом перспектив конвергенции стационарных и мобильных сетей весьма актуальной является задача разработки метода дифференцированного обслуживания абонентов.

В данной работе рассматриваются процессы, происходящие в фиксированной инфраструктуре мобильной сети. По определению ETSI [1] в фиксированную инфраструктуру мобильной сети входят все элементы коммутационной подсистемы SSS, связывающие их транзитные узлы мобильных сетей, а также линии связи стационарных сетей (ISDN, PSTN, PDN), участвующие в установлении соединений между абонентами.

Усложнение характера и рост объема информационной и сигнальной нагрузки приводит к тому, что требуемое качество обслуживания может быть обеспечено только при использовании эффективных методов повышения пропускной способности фиксированной инфраструктуры мобильных сетей. Одним из таких методов является метод дифференцированного обслуживания абонентов, т.е. выделение категорий абонентов, обслуживание которых происходит с повышенным качеством. В настоящее время этот способ может найти широкое применение в мобильных сетях с помощью предоставления абонентам разных уровней обслуживания, тем самым, снижая взаимное влияние пользователей разных категорий. При этом в многоуровневом соглашении о качестве обслуживании SLA (Service Level Agreement) за ежемесячную плату будет установлен определенный объем услуг, а за перерасход может взиматься дополнительная плата. При введении определенных ограничений, обязательно необходимо предложить пользователям дополнительные возможности.

В статье рассматривается динамическое управление приоритетами при дифференцированном обслуживании абонентов в фиксированной инфраструктуре подвижной сети.

На основании анализа действующих тарифных планов различных операторов и наблюдений за информационной и сигнальной нагрузками в сети крупного транзитного оператора мобильной связи было выделено две категории мобильных абонентов: “бизнес” и “экономные”.

Первая категория — “бизнес” абоненты. Она характеризуется меньшей длительностью разговора (в среднем 30 сек), но бо́льшим числом вызовов от абонента (3 вызова в ЧНН). Для данной категории характерно активное использование дополнительных услуг (таких как удержание вызова и/или переадресация). Вторая категория — это “экономные” абоненты. Появление данной категории обусловлено тем, что операторы сотовой связи ранее не тарифицировали первые несколько секунд соединения (обычно пять), чем пользовались некоторые абоненты, а в настоящее время операторы переходят на посекундную оплату с первой минуты. Для этой категории характерно очень низкое среднее время разговора (менее пяти секунд) и очень высокая интенсивность вызовов. Для категории “экономных” абонентов характерно активное использование услуги коротких сообщений. В основном эта категория абонентов создает нагрузку на сеть сигнализации. При исследованиях было выявлено, что основное влияние на поведение сети оказывают абоненты категорий “бизнес” и “экономные”. С учетом данного фактора для построения математической модели обслуживания заявок на звене сети будет достаточно рассмотреть две категории абонентов: приоритетная и неприоритетная.

2. МАТЕМАТИЧЕСКАЯ МОДЕЛЬ

Предметом дальнейших исследований будет модель обслуживания двух потоков сообщений звеном фиксированной инфраструктуры мобильной сети, имеющем v информационных каналов, при этом для требований второго потока доступны только $k, 0 \le k \le v$, каналов. Требования из первого потока образуют пуассоновский поток интенсивности λ_1 и занимают любой свободный канал на время обслуживания, имеющее экспоненциальное распределение с параметром μ_1. После завершения обслуживания требование может с вероятностью продолжить обслуживание, заказав некоторую дополнительную услугу. Продолжительность дополнительного обслуживания имеет экспоненциальное распределение с параметром μ_s. В том случае, когда требование не поступает на обслуживание, оно не может заказать и дополнительную услугу. Требования из второго потока образуют пуассоновский поток интенсивности λ_2 и занимают любой доступный им из k свободный канал на время обслуживания, имеющее экспоненциальное распределение с параметром μ_2. Для требований из второго потока дополнительное обслуживание не предусмотрено.

Такая модель описывается марковским процессом, состояния которого описываются трехмерным вектором (i_1, i_2, i_3), где i_1 — число линий, занятых обслуживанием требований первого потока, i_2 — число линий, занятых дообслуживанием требований первого потока, i_3 — число линий, занятых обслуживанием требований второго потока. Множество возможных состояний

$$S = \{(i_1, i_2, i_3) : i_1 + i_2 + i_3 \le v, 0 \le i_1 \le v, 0 \le i_2 \le v, 0 \le i_3 \le v\},$$

где v — число линий.

Расчет таких моделей с помощью решения уравнений статистического равновесия является достаточно сложной задачей при больших значениях v, поскольку размерность задачи растет как v^3, а трудоемкость решения системы уравнений равновесия еще быстрее, и, кроме того, существенными оказываются вычислительные проблемы, связанные с пропаданием порядка из-за того, что многие вероятности будут слишком малыми. Такие вычислительные проблемы изложены, например, в [2]. Поскольку в нашем случае основным расчетным средством будет приводимая ниже упрощенная модель, то решение рассматриваемой задачи проводится средствами статистического моделирования. Это тем более важно, что с использованием такого

средства можно исследовать и более сложные ситуации, когда времена обслуживания и интервалы между поступлением требований не являются показательными, а соответствующая модель не является марковской.

Стандартные средства численного моделирования состоят в том, что моделируются несколько независимых реализаций рассматриваемого процесса, обозначим их число через N, и результаты обрабатываются стандартными статистическими методами: вычисляется выборочное среднее и среднеквадратическое отклонение для вычисляемых характеристик. Число N выбирается настолько большим, чтобы обеспечить необходимую точность вычислений. Поскольку у нас цель моделирования состоит в том, чтобы оценить погрешность в определении характеристик, связанную с заменой исходной модели на приближенную, то точность выбиралась меньше погрешности в определении основных показателей качества, таких как вероятности потерь сообщений и получаемый доход.

Для построения траектории моделировались моменты поступления требований в систему и моменты освобождения линий, и, в зависимости от состояния, определялось дальнейшее поведение системы. Как указывалось выше, состояние системы задается вектором (i_1, i_2, i_3), где i_1 — число линий, занятых обслуживанием требований первого потока, i_2 — число линий, занятых дообслуживанием требований первого потока, i_3 — число линий, занятых обслуживанием требований второго потока, в момент времени t. Далее определялся момент наступления следующего события, когда возможно изменение состояния системы.

Изменение состояния системы возможно по следующим причинам:

поступление требования первого потока,

поступление требования второго потока,

освобождение линии, занятой обслуживанием требования первого потока,

освобождение линии, занятой дообслуживанием требования первого потока,

освобождение линии, занятой обслуживанием требования второго потока.

Используя свойства показательного распределения, получаем, что момент наступления следующего события произойдет через случайное время τ, имеющее показательное распределение с параметром

$$\lambda = \lambda_1 + \lambda_2 + \mu_1 i_1 + \mu_s i_2 + \mu_2 i_3,$$

при этом вероятность поступления требования первого потока $q_1 = \frac{\lambda_1}{\lambda}$, вероятность поступления требования второго потока $q_2 = \frac{\lambda_2}{\lambda}$, вероятность освобождения линии, занятой обслуживанием требования первого потока $q_3 = \frac{\mu_1 i_1}{\lambda}$, вероятность освобождения линии, занятой дообслуживанием требования первого потока $q_4 = \frac{\mu_s i_2}{\lambda}$, вероятность освобождения линии, занятой обслуживанием требования первого потока $q_5 = \frac{\mu_2 i_3}{\lambda}$, а стоимость обслуживания требований увеличивается и составит величину

$$R = R + \Delta R, \Delta R = (c_1 i_1 + c_s i_2 + c_2 i_3)\tau,$$

где c_i — стоимость обслуживания абонентов категории i;

Далее разыгрывается с помощью датчика случайных чисел наступление одного из этих пяти событий с указанными вероятностями.

Если наступило первое событие, то i_1 увеличивается на 1, если $i_1 + i_2 + i_3 < v$, и не изменяется в противном случае, а величина дохода P уменьшается на стоимость установления соединения α. Если наступило второе событие, то i_3 увеличивется на 1, если $i_1 + i_2 + i_3 < k$, и не изменяется в противном случае, а величина дохода P уменьшается на стоимость установления

соединения α. Если наступило третье событие, т.е. освобождение линии, занятой обслуживанием требования первого потока, то i_1 уменьшается на 1, а далее разыгрывается событие поступления требования на дообслуживание. Если выбранное равномерно распределенное на отрезке $[0,1]$ число меньше a, то требование поступает на дообслуживание и в этом случае i_2 увеличивается на 1, а величина дохода P уменьшается на стоимость установления соединения α, а если выбранное число оказалось больше a, то требование не поступает на дообслуживание и в этом случае не происходит изменение i_2 и P. Если наступило четвертое событие, т.е. освобождение линии, занятой дообслуживанием требования первого потока, то i_2 уменьшается на 1, а если наступило последнее из рассматриваемых пяти событий, т.е. освобождение линии, занятой обслуживанием требования второго потока то i_3 уменьшается на 1.

Таким образом, производится моделирование траектории при наступлении события в системе. После этого время t увеличивается на величину τ. Если получающее новое значение времени больше, чем время наблюдений за траекторией T, то моделирование траектории завершается, в противном случае снова определяется момент наступления следующего события и совершаются описанные выше действия.

Для реализации указанного способа моделирования траектории необходимо задать начальные значения используемых величин и параметров. Понятно, что t, R и P в начальный момент равны 0. Значения i_1, i_2, i_3 выбираются из значений стационарного распределения вероятностей состояний в упрощенной модели. Это позволяет уменьшить значение общей продолжительности моделирования траектории T, которое выбирается из тех соображений, что к этому моменту установился стационарный режим, и, следовательно, уменьшить трудоемкость расчетов.

3. ПРИБЛИЖЕННАЯ МОДЕЛЬ

Поскольку предметом исследований является расчет характеристик звеньев большой емкости, которые обслуживают потоки с большим количеством требований, то можно сделать некоторые упрощения, которые не должны существенно изменить результаты вычислений.

Первое упрощение состоит в том, что двухэтапное обслуживание требований первого потока заменяется одноэтапным; среднее время обслуживания считается равным $t^* = t_1 + at_s$ и стоимость его обслуживания равна $c_1 t_1 + ac_s t_s$. Здесь t_1 — средняя продолжительность обслуживания требований первого потока, $t_1 = \frac{1}{\mu_1}$, t_s — средняя продолжительность дообслуживания требований первого потока, $t_s = \frac{1}{\mu_s}$. Это упрощение основано на том факте, что характеристики качества обслуживания потоков большой емкости мало чувствительны к закону распределения времени обслуживания, а, в первую очередь, зависят от среднего значения времени обслуживания.

Второе упрощение состоит в том, что время обслуживания считается для всех потоков одинаковым, а интенсивности поступления требований пересчитываются таким образом, чтобы сохранить общую нагрузку от соответствующего потока в единицу времени: интенсивность первого потока принимается равной $\Lambda_1 = \lambda_1 t^*$, а интенсивность второго потока — равной $\Lambda_2 = \lambda_2 t_2$, где $t_2 = \frac{1}{\mu_2}$.

Процесс занятия линий в такой модели описывается уже одномерным марковским процессом, состояния которого задаются параметром i — числом занятых линий. Множество возможных состояний

$$S = \{(i) : 0 \le i \le v\}.$$

Получившийся марковский процесс является процессом рождения и гибели, для стационарных вероятностей которого получаем уравнения:

$$p_i = p_{i-1}\frac{\Lambda}{i}, 1 \le i \le k,$$

где

$$\Lambda = \Lambda_1 + \Lambda_2,$$

$$p_i = p_{i-1}\frac{\Lambda_1}{i}, k+1 \le i \le v.$$

Вычисление стоимостного функционала производится следующим образом. Вероятность отказа в обслуживании для требований из первого потока

$$\pi_1 = p_v,$$

а для второго —

$$\pi_2 = \sum_{i=k}^{v} p_i.$$

Полученные характеристики позволяют вычислить стоимостной функционал, показывающий эффективность работы звена сети за рассматриваемый промежуток времени, поскольку известны доли обслуженных требований первого и второго потоков соответственно:

$$R = \lambda_1\pi_1(c_1t_1 + ac_st_s) + \lambda_2\pi_2c_2t_2$$

— стоимость оказанных услуг,

$$P = R - \alpha(\lambda_1(1+a) + \lambda_2)$$

— доход, полученный от оказанных услуг с учетом стоимости установления соединения.

Исследуем эффективность введения динамического управления, направленного на ограничение доступа к сети требований второго потока в моменты, когда звено нагружено выше некоторого порога. Выбор порога осуществляется исходя из требования максимизации доходов сети, при этом возможны различные виды функционала качества работы сети. В первом случае учитывается только стоимость оказанных услуг R, во втором – доход сети, который обозначается P, определяется путем вычитания из стоимости расходов, связанных с установлением соединений, которые зависят от числа поступивших требований, но не зависят от продолжительности соединений. Понятно, что при малых значениях потерь в сети величины R и P отличаются незначительно и не меняют качественной картины.

Для обеспечения экономической эффективности работы сети необходимо добиваться максимально возможной загрузки сети, поэтому при выборе исходных данных для расчетов использовалось следующее соображение. Предполагается, что сеть спроектирована на обслуживание требований первого потока в ЧНН с необходимым качеством, однако из-за неравномерности нагрузки часть оборудования простаивает, поэтому допускается обслуживание требований второго потока с ограниченным доступом к сети и с не гарантированным качеством обслуживания, т.е. для требований второго потока потери не нормируются. Последнее обстоятельство позволяет рассматривать различные ситуации, когда потери сообщений второго потока могут быть весьма большими и интенсивность сообщений второго потока регулируется лишь соображениями практической целесообразности для абонентов — использовать сеть с большими потерями, с относительно низкими тарифами на обслуживание. При таких предположениях

разница между R и P может быть весьма значительной, что может изменять качественную картину на сети и, соответственно, влиять на выбор значения k, при котором обеспечивается максимальный доход от эксплуатации сети.

4. РЕЗУЛЬТАТЫ ЧИСЛЕННЫХ РАСЧЕТОВ

В таблице приведены результаты расчетов, связанные с выбором значения k, при котором обеспечивается максимизация функционала R — стоимость оказанных услуг. Величина λ_2 изменялась в процессе вычислений; ее значения приведены в первой строке таблицы с результатами расчетов. Первое значение λ_2 соответствует случаю, когда суммарный поток требований может быть обслужен с заданной степенью точности без использования ограничений на доступ к сети. При втором значении λ_2 ресурс сети использован практически полностью, поскольку суммарная средняя нагрузка при выбранных значения параметров составляет 170 Эрланг. Два следующих значения λ_2 соответствуют ситуациям, когда сеть перегружена и средняя нагрузка составляет уже 200 и 250 Эрланг соответственно.

Также в таблице приведены значения порога k, обеспечивающие максимальный доход сети при выбранных параметрах и далее приведены значения функционалов P и R соответственно для оптимального значения порога, а так же вероятность потерь требований первого потока. Отметим, что при отсутствии второго потока доход сети составит 349,99. Для более полной картины приведены значения дохода при увеличении стоимости соединения ($\alpha = 0,1$), когда при отсутствии второго потока доход сети составит 334,99, то есть, как видно из приведенных результатов, обслуживание требований второго потока становится экономически нецелесообразным.

Поскольку при оптимизации доходов потери требований первого потока могут быть весьма значительными, то рассмотрен случай, когда качество обслуживания требований первого потока гарантировано за счет выбора порога ограничений на доступ к сети требований из второго потока. Результаты расчетов для этого случая приведены в следующих строках таблицы.

Далее рассмотрен случай, когда регулирования доступа к сети нет. В этом случае, как видно из приведенных результатов, качество обслуживания и доходы могут оказаться существенно меньше даже чем в случае, когда требования из второго потока не обслуживаются вовсе, а приведенном примере (выделенном в таблице) он меньше, чем при более малых значениях λ_2.

Таблица

Результаты расчетов при $\lambda_1 = 100, \quad c_1 = 2, \quad c_2 = 1, \quad c_s = 3,$
$\mu_1 = 1, \quad \mu_2 = 10, \quad \mu_s = 1, \quad a = 0,5, \quad \alpha = 0,01, \quad v = 200.$

λ_2	100	200	500	1000
k	197	196	194	192
P	359,91	369,23	384,64	388,34
R	357,41	365,73	371,76	376,84
$R(\alpha = 0,1)$	334,91	334,23	319,64	273,34
p	0,0003	0,0014	0,012	0,016
$k(p \le 0,001)$	200	193	181	179
$P(p \le 0,001)$	357,40	365,66	378,56	380,53
$P(k = v)$	357,40	369,14	378,26	**353,83**
$p(k = v)$	0,0002	0,002	0,054	0,214

5. ВЫВОДЫ

Приближенная модель достигает требуемой точности при $v > 50$.

Без применения дифференцированного обслуживания качество обслуживания потока первой категории является низким.

Введение дифференцированного обслуживания позволяет гарантировать качество обслуживания, требуемое для потока первой категории, не значительно ухудшая качество второго потока.

При отстутствии порога для обслуживания абонентов второй категории, несмотря на большое количество обслуженных вызовов, общий доход сети падает.

При введении в модель повторных вызовов негативный эффект будет еще более значительным, что позволит выявить дополнительные преимущества дифференцированного обслуживания.

СПИСОК ЛИТЕРАТУРЫ

1. ETR-351. Digital cellular telecommunications system Technical performance objectives (GSM 03.05 version 5.0.0). 1996 - 25 с.

2. Степанов С.Н. Численные методы расчета систем с повторными вызовами. М.: Наука 1983, 230 с.