ДВА ПОДХОДА К ДЕКОМПОЗИЦИИ СЛОЖНЫХ ИЕРАРХИЧЕСКИ СТОХАСТИЧЕСКИХ СИСТЕМ. НЕПРЕРЫВНО ВЗАИМОДЕЙСТВУЮЩИЕ ПОДСИСТЕМЫ

Владимир Рыков
(Москва, Россия)

Светлой памяти
Бориса Владимировича
Гнеденко посвящаю

АННОТАЦИЯ

Рассматриваются два подхода к решению проблемы декомпозиции процессов функционирования сложных иерархических систем, опирающиеся па условия взаимодействия между их подсистемами. Первый подход, которому посвящена первая часть работы (настоящая статья), связан с изучением систем с непрерывно взаимодействующими подсистемами. Типичным примером таких систем являются информационно-вычислительные сети. Традиционно такие системы исследуются с помощью моделей, сетей массового обслуживания, для которых проблема декомпозиции решается в рамках мультипликативною представления для стационарно го распределения вероятностей их состояний. Путем расширения метода дополнительной переменной доказана допустимость мультипликативного представления на случай иерархической сети с зависимыми длительностями обслуживания, характерный для систем передачи данных.

study

1. ВВЕДЕНИЕ

Многие современные производственные, экономические информационные социальные и др. системы имеют сложную иерархическую структуру. Под системой будем понимать объект, допускающий разложение на части (подсистемы), сохраняющие, по крайней мере, частично функции всей системы. Сложной называется система, подсистемы которой допускают дальнейшее разложение, неразложимые части системы называются ее элементами, а максимальное число уровней разложения - рангом сложной системы.

Заметим, что системы допускают, вообще говоря, различное разложение относительно различных функций и процессов в них происходящих.

Наибольший интерес с точки зрения анализа систем представляет исследование поведения процессов, протекающих в сложных иерархических структурах, которое в значительной степени зависит от режима взаимодействия ее подсистем. Сложность этих процессов побуждает искать пути их декомпозиции. В работе рассматриваются два подхода к декомпозиции процессов в сложных иерархических стохастических системах, связанные с режимом взаимодействия их подсистем, — один из них допускает декомпозицию при непрерывном взаимодействии, другой использует идею взаимодействия подсистем в дискретные, вообще говоря, случайные моменты времени. Первая часть посвящена системам с непрерывно взаимодействующими подсистемами на примере информационно-вычислительной сети (ИВС). Во второй рассматриваются системы с дискретно взаимодействующими подсистемами.

Исследуемая здесь модель предложена в работе [1], где она была использована для декомпозиции ИВС, в том числе в предположениях, обеспечивающих описание процессов функционирования сети дискретным марковским процессом. Ниже уточняются и развиваются результаты этой работы применительно к сетям с сообщениями произвольно распределенного объема, не изменяющегося в процессе их передачи по сети, но, возможно, передающихся с различными скоростями различными каналами.

Материал первой части представлен следующим образом. В разделе 2 приводится краткий обзор известных результатов но декомпозиции систем с непрерывно взаимодействующими подсистемами и на примере ИВС предложена модель непрерывно взаимодействующей стохастической иерархической сети. В разделах 3. 4 метод дополнительных переменных применен для исследования ИВС с независимыми на различных каналах длительностями передачи сообщений. Отправляясь от полученных результатов с помощью замены переменных мы перейдем к выражению соответствующих характеристик в терминах объемов (длин) сообщений и скоростей работы каналов с учетом их различия. Наконец, в разделах 5, б полученные результаты обобщаются на случай сетей, объемы сообщений в которых не изменяются при передаче по сети, а длительности их передачи определяются скоростью работы соответствующих каналов. Для наглядности используется терминология ИВС.

2. НЕПРЕРЫВНО ВЗАИМОДЕЙСТВУЮЩИЕ ПОДСИСТЕМЫ. МАТЕМАТИЧЕСКАЯ МОДЕЛЬ

Наиболее ярким примером непрерывно взаимодействующих стохастических систем являются современные ИВС. Технология открытости привела к их бурному развитию и объединению локальных сетей в глобальные, имеющие сложную иерархическую структуру, примером которой является, в частности, Internet [2]. Ввиду все возрастающей нагрузки в сетях проблема анализа их производительности приобретает все более важное значение.

Адекватным аппаратом исследования ИВС являются сети массового обслуживания (СеМО), а многочисленные особенности работы ИВС (различные протоколы и способы передачи и обработки информации, используемые в ИВС) приводят к разнообразию используемых моделей. В частности, т.к. наиболее узким местом работы сети являются каналы связи, обеспечивающие ресурсами всю сеть, в то время как серверы решают локальные задачи, то при моделировании ИВС в качестве узлов сети целесообразно рассматривать каналы, отводя серверам положение накопителей информации.

Одним из наиболее эффективных способов передачи сообщений является режим коммутации пакетов, при котором сообщение разбивается на части (пакеты), передающиеся независимо друг от друга наряду с пакетами от других сообщений и собирающиеся вместе при приеме. Естественной моделью организации пакетного режима передачи информации является дисциплина разделения времени. Эта дисциплина не только создает у пользователей иллюзию непрерывной работы в сети, но и обеспечивает определенные "хорошие" свойства работы системы.

Как проблемам моделирования ИВС с помощью СеМО, так и вопросам анализа последних посвящена многочисленная литература (см., например, [3] - - [7] и др.). Сложность анализа сетей побуждает искать пути их декомпозиции. Постановка задачи и первые результаты в этом направлении принадлежат Джексону [8, 9], который показал возможность декомпозиции экспоненциальных сетей при естественной дисциплине обслуживания в порядке посту пения требований (FIFO) с марковской маршрутизацией сообщений. В дальнейшем этот результат обобщался в различных направлениях, что привело к исследованию сетей типа Джексона и так называемых BCMP-сетей (см. [10], а также, например, обзоры [11, 12]). В этих сетях предполагается марковская маршрутизация и независимое обслуживание в узлах, что не всегда адекватно отражает ситуацию в реальных сетях передачи данных. Следующий шаг в направлении исследования СеМО был сделан Келли [13], который расширил принципы маршрутизации в сетях, допустив возможность априорного задания маршрута сообщения по сети, и развил общую концепцию декомпозиции, опирающуюся на обратимость соответствующего марковского процесса. Однако используемый им аппарат не позволил ему исследовать сети с зависимыми длительностями передачи сообщений.

В реальных ИВС длина сообщения в процессе его передачи от одного узла к другому не меняется. Это приводит к необходимости исследования сетей с зависимыми длительностями передачи сообщений на каналах.
Для решения этой задачи в настоящей работе обобщается классический метод введения дополнительных переменных, который используется при доказательстве мультипликативного представления стационарных вероятностей состояний сложных иерархических стохастических сетей, работающих в режиме разделения времени, как в случае независимых, так и зависимых длительностей передачи сообщений. При этом мы не затрагиваем тонких вопросов существования стационарных распределений для общих сетей, которые представляют собой предмет самостоятельных исследований (см., например, обзор [14]). Полученные результаты обеспечивают декомпозицию сложных иерархических стохастических систем, работающих в режиме разделения времени.

С точки зрения анализа процессов передачи сообщений сложная иерархическая система может быть представлена в виде сети, изображенной на рис. 2.

Потоки сообщений в такую сеть образованы суперпозицией независимых потоков от многих источников и согласно предельным теоремам для потоков независимых событий могут рассматриваться как пуассоновские. Мы будем предполагать их стационарными с интенсивностями λ_{ij} для потока от i-го источника в j-ый. Объемы (длины) сообщений для простоты будем измерять в единицах объема пакета, а скорости их передачи - в количестве пакетов в единицу времени. Тогда время T передачи сообщения будет пропорционально его объему V и обратно пропорционально скорости μ соответствующего канала, а распределение времени передачи сообщения *B(t)*) связано с распределением его объема *G(x)* соотношением

(1) $$B(x) = \mathbf{P}\{T < x\} = \mathbf{P}\{V/\mu < x\} = \mathbf{P}\{V < \mu x\} = G(\mu x).$$

Дополнительные предположения относительно объемов сообщений и скоростей их обслуживания будут сделаны в дальнейшем при исследовании конкретных моделей. Начнем с исследования поведения изолированного узла с независимыми длительностями передачи сообщений.

3. НЕЗАВИСИМЫЕ ДЛИТЕЛЬНОСТИ ПЕРЕДАЧИ СООБЩЕНИЙ. МОДЕЛЬ УЗЛА

Пусть в узел поступают r независимых пуассоновских потоков сообщений с интенсивностями λ_i, длительности передачи которых независимы и имеют общие распределения $B_i(\cdot), i=\overline{1,r}$. Предположим, что функции распределения (ФР) $B_i(\cdot)$ обладают плотностями и средними значениями, которые обозначим через $b_i(\cdot)$ и m_i^B, соответственно. Обозначим через $\rho_i = \lambda_i m_i^B$ загрузку канала сообщениями типа i, а через $\rho = \rho_1 + \ldots + \rho_r$ - полную его загрузку.

В качестве состояний узла рассмотрим два набора переменных (n, x), где набор n=($n_1,\ldots,n_r$) представляет собой количества сообщений различных типов с общим их числом $n = \sum 1 \le i \le r n_i$, а набор x = $(x_{ij}, j=\overline{1,n_i}, i=\overline{1,r})$ описывает остаточные длительности передачи ожидающих сообщений, расположение в произвольном (случайном) порядке. Рассмотрим случайный процесс

X(t) = (n, x), если в момент t система находится в состоянии (n,x)

и предположим, что плотности (по переменным x_{ij}) распределения его вероятностей состояний существуют (можно показать, что это имеет место при сделанных предположениях). Справедлива теорема

<u>Теорема 1.</u> При условии

(2) $$\rho < 1$$

процесс X(t) положительно возвратен по Харрису, а его предельные стационарные плотности p(n,x) распределения вероятностей состояний существует, и имеет вид

(3) $$p(\mathbf{n},\mathbf{x}) = (1-\rho)\frac{n!}{n_1!\ldots n_r!}\prod_{1\le i\le r}\prod_{1\le j\le n_i}\lambda_i(1-B_i(x_{ij})).$$

При этом распределение вероятностей макросостояний дается формулой

(4) $$p(\mathbf{n}) = (1-\rho)\frac{n!}{n_1!\ldots n_r!}\rho_1^{n_1}\ldots\rho_r^{n_r},$$

которой отвечает геометрическое распределение суммарного количества требований в системе

(5) $$p(n) = (1-\rho)\rho^n.$$

Доказательство. Если процесс X(t) положительно возвратен по Харрису, то согласно теории марковских процессов [15] его предельное распределение вероятностей состояний существует, а систему уравнений равновесия для него можно получить приравнивая к нулю частные производные по времени в соответствующей системе уравнений Колмогорова. Последняя выводится обычным методом путем перехода к пределу в системе разностных уравнений, связывающих вероятности состояний процесса в близкие моменты времени.

Для удобства записи этих уравнений введем операторы на пространстве состояний системы, описывающие их изменения при поступлении A_{ij} требования i-го типа на j-ую позицию и окончании обслуживания D_{ij} требования i-го типа на j-ой позиции. Формальное описание этих операторов более громоздко, чем содержательное, поэтому приведем последнее, причем применительно к обратным операторам, которые участвуют в записи соответствующих уравнений. Оператор A_{ij}^{-1} удаляет компоненту x_{ij} из набора x, оператор D_{ij}^{-1} добавляет на j-ое место i-го набора компоненту x_{ij} со значением ноль. С использованием введенных операторов и обозначения e_i для набора i-ая компонента которого равна 1, а все остальные 0, уравнения для стационарных вероятностей состояний процесса X(t) имеют вид

$$
\begin{aligned}
&-\frac{1}{n}\sum_{1\le i\le r}\sum_{1\le j\le n_i}\frac{\partial p(\mathbf{n},\mathbf{x})}{\partial x_{ij}}+\lambda p(\mathbf{n},\mathbf{x})= \\
&=\sum_{1\le i\le r}\sum_{1\le j\le n_i}\frac{\lambda_i}{n_i}p(\mathbf{n}-\mathbf{e}_i,A_{ij}^{-1}\mathbf{x})b_i(x_{ij})+ \\
&\quad+\sum_{1\le i\le r}\sum_{1\le j\le n_i+1}\frac{1}{n+1}p(\mathbf{n}+\mathbf{e}_i,D_{ij}^{-1}\mathbf{x})
\end{aligned}
\tag{6}
$$

с естественными изменениями для уравнений на границе области дискретных значений процесса. По непрерывным переменным эти уравнения выполняются всюду в области их изменения $0 \le x_{ij} < \infty$, $j=\overline{1,n_i}$, $j=\overline{1,r}$, включая границу.

Для объяснения приведенной системы уравнений заметим, что

а) при увеличении реального времени на величину h остаточное время каждого из n ожидающих требований уменьшается на величину h/n,
б) при поступлении сообщения i-го типа в систему время его обслуживания с равной вероятностью занимает любую из n_i позиций,
и, наконец,
в) чтобы одно из $(n + 1)$ сообщений покинуло систему, оно должно иметь остаточное время обслуживания в $(k/(n + 1))$ - окрестности точки ноль.

Непосредственной проверкой легко убедиться, что выражение (3) удовлетворяет этому уравнению. Интегрируя его по дополнительным переменным x_{ij} получаем распределение вероятностей (4) для макросостояний системы, а суммируя получившийся результат по $n_1 \ldots n_r$ таким, что $\sum 1 \le i \le r n_i = n$ и распределение (5) для суммарного количества требований в системе.

Чтобы показать, что условие (2) достаточно для положительной возвратности процесса X(t) по Харрису [15] заметим, что процесс X(t) неразложим, так как состояние "0", например, достижимо из любого другого с положительной вероятностью и, наоборот, любое состояние можно достичь из нулевого с положительной вероятностью. Так как нулевое состояние является положительным атомом для этого процесса, то для доказательства его положительной возвратности по Харрису достаточно показать конечность среднего времени возвращения в нулевое состояние. Однако, так как время возвращения складывается из конечного (со средним $1/\lambda$) свободного периода и периода занятости системы, то достаточно показать конечность в среднем последнего. Наконец, период занятости этой системы исследуется классическими методами (см., например, [16]), и условие (2) обеспечивает конечность его среднего значения.

Чтобы перейти к соответствующему выражению вероятностей состояний через распределение объемов V сообщений с учетом скорости работы канала воспользуемся заменой (1), из которой следует, что соответствующие средние значения связаны соотношением $m^B = m/\mu$. Поэтому, подставляя последнее выражение в формулу (3), получим

Следствие. В терминах распределений объемов сообщений стационарные плотности распределения вероятностей состояний имеют вид

$$p(\mathbf{n}, \mathbf{x}) = (1-\rho)\frac{n!}{n_1!\dots n_r!}\prod_{1\le i\le r}\prod_{1\le j\le n_i}\lambda(1-G_i(\mu x_{ij})) \tag{7}$$

с распределением макросостояний вида (4), где

$$\rho_i = \frac{\lambda_i m_i}{\mu}, \qquad \rho = \sum_{1\le i\le r}\rho_i.$$

4. НЕЗАВИСИМЫЕ ДЛИТЕЛЬНОСТИ ПЕРЕДАЧИ СООБЩЕНИЙ. МОДЕЛЬ СЕТИ

Переходя к исследованию модели сети, ограничимся исследованием ее фрагмента, состоящего из двухуровневой подсети (см. рис. 2) и рассмотрим сначала случай независимых длительностей передачи сообщений на различных уровнях сети с распределениями $B_{ij}^{(1)}(\cdot)$ и $B_{ij}^{(2)}(\cdot)$ для длительностей передачи сообщений на каналах первого и второго уровней, соответственно. Для описания пространства состояний сети используем набор переменных (n,x), где n = ($n_i^{(1)}$, $n^{(2)}$), а $n_i^{(1)}$=($n_{ij}^{(1)}$), $n^{(2)}$=($n_i^{(2)}$), $i, j = \overline{1, r}$ и $n_{ij}^{(1)}$ означает число сообщений, передающихся *i*-ым каналом в *j*-ом направлении на нижнем уровне, $n_i^{(2)}$ - число сообщений, передающихся на верхнем уровне в i -ом направлении, x = ($x_i^{(1)}$, $x^{(2)}$), а переменные $x_i^{(1)}$=$(x_{ijl}^{(1)}, l = \overline{1, n_{ij}^{(1)}})$; $x^{(2)}$=$(x_{il}^{(2)}, l = \overline{1, n_i^{(2)}})$ описывают остаточные длительности передачи сообщений на разных каналах в рассматриваемый момент времени. Эти переменные располагаются в произвольном (случайном) порядке в своих наборах. Введем процесс X(t) = (n, x), если в момент t система находится в состоянии (n,x) и предположим существование плотностей (по переменным $x_{ijl}^{(1)}, x_{il}^{(2)}$ распределения вероятностей его состояний. Справедлива теорема.

Теорема 2. Если процесс X(£) положительно возвратен по Харрису, то он обладает предельным распределением, вероятностей, плотности p(n, x) которого допускают мультипликативное представление

(8) $$p(\mathbf{n},\mathbf{x}) = \prod_{1\le i\le r} \hat{p}(\mathbf{n}_i^{(1)},\mathbf{x}_i^{(1)}) \times \hat{p}(\mathbf{n}^{(2)},\mathbf{x}^{(2)}),$$

где распределения вероятностей состояний узлов p(п, x) задаются формулой (3) с соответствующими распределениями длительностей передачи сообщений, а интенсивности потоков на каналы нижнего и верхнего уровней равны

(9) $$\lambda_{ij}^{(1)} = \lambda_{ij}, \quad (j\ne i), \quad \lambda_{ii}^{(1)} = \lambda_{ii} + \sum_{j\ne i}\lambda_{ji}, \qquad \lambda_i^{(2)} = \sum_{j\ne i}\lambda_{ji}.$$

Соответствующее распределение вероятностей макросостояний имеет вид

(10) $$p(\mathbf{n}) = \prod_{1\le i\le r} \hat{p}(\mathbf{n}_i^{(1)}) \times \hat{p}(\mathbf{n}^{(2)}),$$

где распределения p(п) определяются формулами (4) с соответствующими наборами п числа сообщений и нагрузками, определяемыми соотношениями

(11) $$\rho_{ij}^{(1)} = \lambda_{ij}^{(1)} m_{ij}^B, \quad \rho_i^{(1)} = \sum_{1\le j\le r}\rho_{ij}^{(1)}, \qquad \rho_i^{(2)} = \lambda_i^{(2)} m_{ii}^B.$$

Доказательство. Если процесс X(t) положительно возвратен но Харрису, то согласно теории марковских процессов его предельное распределение вероятностей состояний существует, а систему уравнений стационарного режима для их плотностей можно получить, приравнивая к нулю производные но времени в системе дифференциальных уравнений Колмогорова. Последняя получается обычным методом путем предельного перехода в соответствующей системе разностных уравнений, связывающих вероятности состояний процесса в близкие моменты времени.

Для удобства записи этих уравнений введем операторы на пространстве состояний системы, описывающие их изменения при поступлении сообщения типа *ij* на *l*-ую позицию - A_{ijl}, при передаче (*ij*)-сообщения с *l*-ой позиции нижнего уровня на *m*-ую позицию верхнего - $T_{ij}^{(12)}(lm)$, и *j*-сообщения с *m*-ой позиции верхнего на *l*-ую позицию нижнего - $T_j^{(21)}(ml)$, и при окончании обслуживания требования *i*-го типа на *l*-ой позиции - D_{iil}.

Так как в уравнениях Колмогорова участвуют обратные операторы, приведем сразу их содержательное описание:

- оператор A_{ijl}^{-1} удаляет компоненту $x_{ijl}^{(1)}$ из набора x,
- оператор $\left(T_{ij}^{(12)}(lm)\right)^{-1}$ добавляет компоненту $x_{ijl}^{(1)}$ со значением ноль и удаляет компоненту $x_{jm}^{(2)}$,
- оператор $\left(T_{ij}^{(21)}(ml)\right)^{-1}$ добавляет компоненту $x_{jm}^{(2)}$ со значением ноль и удаляет

компоненту $x_{jjl}^{(1)}$, и, наконец,

- оператор D_{iil}^{-1} добавляет компоненту $x_{iil}^{(1)}$ со значением ноль.

С использованием введенных операторов уравнения равновесия принимают вид

$$
\begin{aligned}
-\sum_{i,j,l}\frac{1}{n_i^{(1)}}\frac{\partial p(\mathbf{n},\mathbf{x})}{\partial x_{ijl}^{(1)}} - \frac{1}{n^{(2)}}\sum_{j,l}\frac{\partial p(\mathbf{n},\mathbf{x})}{\partial x_{jl}^{(2)}} + \lambda p(\mathbf{n},\mathbf{x}) &= \\
= \sum_{i,j,l}\frac{\lambda_{ij}}{n_{ij}^{(1)}} p(\mathbf{n}-\mathbf{e}_{ij}^{(1)}, A_{ijl}^{-1}\mathbf{x})\, b_{ij}^{(1)}(x_{ijl}^{(1)}) &+ \\
+\sum_{i,j\neq i,l,m}\frac{1}{(n_i^{(1)}+1)n_j^{(2)}} p\left(\mathbf{n}+\mathbf{e}_{ij}^{(1)}-\mathbf{e}_j^{(2)}, \left(T_{ij}^{(12)}(lm)\right)^{-1}\mathbf{x}\right) b_j^{(2)}(x_{jl}^{(2)}) &+ \\
+\sum_{j,l,m}\frac{1}{n_{jj}^{(1)}(n^{(2)}+1)} p\left(\mathbf{n}+\mathbf{e}_j^{(2)}-\mathbf{e}_{jj}^{(1)}, \left(T_{jl}^{(21)}(ml)\right)^{-1}\mathbf{x}\right) b_{jj}^{(1)}(x_{jjl}^{(1)}) &+ \\
+\sum_{i,l}\frac{1}{n_i^{(1)}+1} p(\mathbf{n}+\mathbf{e}_{ii}^{(1)}, D_{iil}^{-1}\mathbf{x}). & \qquad (12)
\end{aligned}
$$

Суммирование в этих уравнениях ведется по всем допустимым значениям индексов *i, j, l, m,* что обеспечивает их выполнение на границе области дискретных значений процесса. По непрерывным переменным эти уравнения выполняются всюду в области допустимых значений переменных, включая границу

$$
\begin{aligned}
&0 \le x_{ijl}^{(1)} < \infty, \quad l = \overline{1, n_{ij}^{(1)}}, \quad i,j = \overline{1,r}, \\
&0 \le x_{jm}^{(2)} < \infty, \quad m = \overline{1, n_j^{(2)}}, \quad j = \overline{1,r}.
\end{aligned}
$$

Для пояснения этих уравнений обозначим через 1 набор из единиц соответствующей размерности и заметим, что переход в состояние (n, x) за малое время *h* возможен следующими несовместимыми способами:

а) из состояния $(\mathrm{n}, \mathrm{x}_i^{(1)} - (h/n_i^{(1)})1, \mathrm{x}^{(2)} - (h/n^{(2)})1)$, если за это время не поступит ни одного сообщения;

б) из состояния $(\mathrm{n} - \mathrm{e}_{ij}^{(1)}, A_{ijl}^{-1}\mathrm{x})$ при поступлении (с плотностью распределения вероятностей $b_{ij}^{(1)}(x_{ijl}^{(1)})$ сообщения типа (*ij*) с длительностью передачи $x_{ijl}^{(1)}$, которое с равной вероятностью занимает любую из $n_{ij}^{(1)}$ позиций среди переменных $x_{ijl}^{(1)}$;

в) из состояния $(\mathrm{n} + \mathrm{e}_{ij}^{(1)} - \mathrm{e}_j^{(2)}, \left(T_{ij}^{(12)}(lm)\right)^{-1}\mathrm{x})$,при переходе сообщения с первого уровня на второй, при этом, длительность $x_{ijl}^{(1)}$ сообщения, передача которого заканчивается, должна находиться в $h/(n_i^{(1)}+1)$ - окрестности нуля, а время его передачи на канале второго уровня приобретает (с плотностью распределения вероятностей $b_j^{(2)}(x_{jm}^{(2)})$ значение $x_{jm}^{(2)}$, при этом на следующем канале оно займет с равной вероятностью любую из $n_j^{(2)}$ позиций;

г) из состояния $(\mathrm{n}+\mathrm{e}_j^{(2)}-\mathrm{e}_{jj}^{(1)}, \left(T_{jl}^{(21)}(ml)\right)^{-1}\mathrm{x})$, при переходе сообщения со второго уровня на первый, при этом длительность сообщения, передача которого заканчивается, должна находиться в $h/(n^{(2)}+1)$- окрестности нуля, а время его передачи на следующем канале приобретает значение $x_{jjl}^{(1)}$ (с плотностью распределения вероятностей $b_{jj}^{(1)}(x_{jjl}^{(1)})$ и займет с равной вероятностью любую из $n_{jj}^{(1)}$ позиций в своем наборе;

д) из состояния $(\mathrm{n}+\mathrm{e}_{ii}^{(1)}, D_{iil}^{-1}\mathrm{x})$, при окончании передачи сообщения, при этом длительность сообщения, передача которого заканчивается, должна находиться в $h/(n_i^{(1)}+1)$- окрестности нуля.

Непосредственной проверкой можно убедиться, что выражение (8), где распределения вероятностей состояний узлов $p(\mathrm{n},\mathrm{x})$ задаются формулой (3), а интенсивности потоков на каналы нижнего и верхнего уровней — выражениями (9), удовлетворяют этому уравнению.

Распределение вероятностей макросостояний (10) для этой системы получается интегрированием по дополнительным переменным $x_{ijl}^{(1)}, x_{il}^{(2)}$.

Полученное мультипликативное представление распределения вероятностей состояний сети обеспечивает ее декомпозицию. При этом, помимо прочего, эти соотношения демонстрируют нечувствительность распределения вероятностей макросостояний модели к виду функций распределения длительностей передачи сообщений.

Как и в случае изолированного узла, справедливо следствие

Следствие. При заданных распределениях объемов сообщений и различных скоростям их передачи в каналах выражение для плотностей распределения вероятностей состояний процесса по-прежнему имеет вид (8), где плотности распределения вероятностей состояний узлов задаются формулой (7), а распределение вероятностей макросостояний совпадает с (10) при нагрузках, задаваемых выражениями

(13) $$\rho_{ij}^{(1)}=\frac{\lambda_{ij}^{(1)}m_{ij}}{\mu_i^{(1)}},\quad \rho_i^{(1)}=\sum_{1\le j\le r}\rho_{ij}^{(1)},\qquad \rho_i^{(2)}=\frac{\lambda_i^{(2)}m_{ii}}{\mu^{(2)}}.$$

Доказательство получается простой заменой в соотношении (3) функций $B_{ij}^{(k)}(x)$ на $G_{ij}(\mu^{(k)}x)$.

Таким образом, при сделанных предположениях сеть допускает декомпозицию.

В реальных сетях связи предположение о независимости длительностей передачи сообщения на отдельных каналах неестественно, так как передается одно и то же сообщение, хотя, возможно, с различной скоростью разными каналами. Рассмотрим, поэтому вопрос о возможности декомпозиции сети в предположении, что объем сообщения не меняется при его передаче, хотя возможны различные скорости его передачи различными каналами. Чтобы ввести соответствующие обозначения, анализ сети начнем с изучения простой тандем-системы с неизменными объемами сообщений.

5. НЕИЗМЕННЫЕ ОБЪЕМЫ СООБЩЕНИЙ. ТАНДЕМ-СИСТЕМА

Рассмотрим систему, функционирующую в режиме разделения времени и состоящую из двух последовательно соединенных каналов, работающих со скоростями $\mu^{(1)}$ и $\mu^{(2)}$, соответственно, на которую поступает простейший поток сообщений интенсивности λ, объемы V которых распределены по закону $G(\cdot)$ с заданными плотностью распределения $g(\cdot)$ и средним значением m (рис. 3).

Предполагается, что объемы сообщений не меняются при их передаче, а длительности их передачи выражаются через их объемы формулой (1), так что распределения длительностей передачи и их плотности связаны соотношениями

(14)
$$\frac{1}{\mu^{(1)}}b^{(1)}(\frac{x}{\mu^{(1)}}) = \frac{1}{\mu^{(2)}})b^{(2)}(\frac{x}{\mu^{(2)}}) = g(x).$$
$$B^{(1)}(\frac{x}{\mu^{(1)}}) = B^{(2)}(\frac{x}{\mu^{(2)}}) = G(x),$$

Для построения марковского процесса при описании пространства состояний системы нам потребуется информация не только об остаточных, но также и об исходных длительностях передачи сообщений, находящихся на первой фазе. Поэтому для описания состояния системы используется набор (n,x,y), где переменные n = $(n^{(1)}, n^{(2)})$ указывают число сообщений на первой и второй фазах обслуживания, соответственно, переменные x $=(x_{1i}, i=\overline{1,n^{(1)}}), (x_{2i}, i=\overline{1,n^{(2)}})$ означают остаточные длительности передачи сообщений на первой и второй фазах, соответственно, и переменные y = $(y_i, i=\overline{1,n^{(1)}})$ обозначают исходные длительности передачи сообщений, находящихся на первой фазе. При этом переменные x_{1i} и x_{2i} располагаются в произвольном (случайном) порядке, однако порядок расположения переменных x_{1i} и y_i совпадает. Введем далее процесс X(t) соотношением X(t) = (n,x,y), если в момент t система находится в состоянии (n,x,y) и предположим, что плотности (по переменным x, y) распределения вероятностей его состояний существуют. Для процесса X(t) справедлива теорема.

Теорема 3. При условии, что процесс X(t) положительно возвратен по Харрису, его предельные плотности распределения вероятностей состояний p(n,x,y) существуют и допускают, мультипликативное представление

(15)
$$p(\mathbf{n},\mathbf{x},\mathbf{y}) = (1-\rho_1)(1-\rho_2)\prod_{1\le i\le n_1}\lambda\Theta(y_i-x_{1i})g(\mu^{(1)}y_i)\prod_{1\le j\le n_2}\lambda\left(1-G(\mu^{(2)}x_{2j})\right).$$

где

$$\rho_1 = \frac{\lambda m}{\mu^{(1)}}, \qquad \rho_2 = \frac{\lambda m}{\mu^{(2)}},$$

и $\Theta(x)$ есть Θ-функция: $\Theta(x)$=0 при $x < 0$ и $\Theta(x)$=0 при $x > 0$.

Доказательство. Если процесс X(t) положительно возвратен по Харрису, то согласно теории марковских процессов [15] его предельное распределение вероятностей состояний существует, а систему уравнений для его плотностей можно получить, приравнивая к нулю частные производные по времени в соответствующих уравнениях Колмогорова. Последние получаются обычным методом путем перехода к пределу в системе разностных уравнений, связывающих плотности распределения вероятностей состояний процесса в близкие моменты времени.

Для упрощения записи этой системы уравнений используем, как и ранее, операторы преобразования состояний A_i, T_{ij}, D_i, соответствующие поступлению сообщения, его переходу с первой фазы на вторую и окончанию обслуживания. Действие этих операторов в рассматриваемом случае отличается от описанного ранее. Оператор A_i добавляет переменные x_{1i} и y_i одновременно, оператор T_{ij} действует на границе при достижении переменной x_{1i} значения ноль, при этом он удаляет ее, помещая на позицию (*2j*) значение скорости передачи сообщения x_{2j}, соответствующее значению y_i с сохранением его объема, т.е. в соответствии с соотношением: $\mu^{(1)}y_i = \mu^{(2)}x_{2j}$, оператор D_i удаляет переменную x_{2j} по достижении ею значения ноль. Заметим, что в уравнениях равновесия, которые соответствуют обратным уравнениям Колмогорова, участвуют обратные операторы переходов.

Теперь область значений процесса X(t) и соответственно область определения этих уравнений имеет вид

$$0 \le x_{1i} \le y_i < \infty, \quad i = \overline{1, n_1}, \qquad 0 \le x_{2j} < \infty, \quad j = \overline{1, n_2}$$

и при поступлении требования процесс оказывается на границе $x_{1i} = y_i$, $i = \overline{1, n_1}$, что в уравнениях равновесия должно отражаться специальными граничными условиями. Однако с помощью δ-функции Дирака мы введем эти граничные условия в общую систему уравнений.

Заметим, кроме того, что так как дополнительными переменными являются длительности передачи сообщений, то для того чтобы выписать уравнения сразу в терминах распределения объемов сообщений мы воспользуемся формулой (14) для связи соответствующих плотностей.

С учетом введенных операторов и сделанных замечаний система уравнений для стационарных плотностей распределения вероятностей процесса X(t) принимает вид

(16)
$$\begin{aligned}
-\frac{1}{n^{(1)}} \sum_{1 \le i \le n_1} \frac{\partial p(\mathbf{n}, \mathbf{x}, \mathbf{y})}{\partial x_{1i}} - \frac{1}{n^{(2)}} \sum_{1 \le j \le n_2} \frac{\partial p(\mathbf{n}, \mathbf{x}, \mathbf{y})}{\partial x_{2j}} + \lambda p(\mathbf{n}, \mathbf{x}, \mathbf{y}) = \\
= \frac{\lambda}{n^{(1)}} \sum_{1 \le i \le n^{(1)}} p(\mathbf{n} - \mathbf{e}_1, A_i^{-1}(\mathbf{x}, \mathbf{y})) g(\mu^{(1)} y_i) \delta(y_i - x_{1i}) + \\
+ \frac{1}{(n^{(1)} + 1) n^{(2)}} \sum_{1 \le i \le n_1} \sum_{1 \le j \le n_2} p(\mathbf{n} + \mathbf{e}_1 - \mathbf{e}_2, T_{ij}^{-1}(\mathbf{x}, \mathbf{y})) + \\
+ \frac{1}{n^{(2)} + 1} \sum_{1 \le j \le n_2} p(\mathbf{n} + \mathbf{e}_2, D_j^{-1}(\mathbf{x}, \mathbf{y})).
\end{aligned}$$

Для пояснения этих уравнений заметим, как и ранее, что переход в состояние (n,x,y) за малое время h возможен следующими несовместимыми способами:

а) из состояния $(\mathrm{n}, \mathrm{x}_i^{(1)} - (h/n^{(1)})1, \mathrm{x}^{(2)} - (h/n^{(2)})1, \mathrm{y})$, если за это время не поступит ни одного сообщения;

б) из состояния $(\mathrm{n} - \mathrm{e}_i^{(1)}, A_i^{-1}(\mathrm{x}, \mathrm{y}))$ при поступлении (с вероятностью λh) сообщения длительности передачи y_i (с плотностью $b^{(1)}(y_i) = g(\mu^{(1)} y_i)$), совпадающее в этот момент с остаточной длительностью x_{1i}, которое с равной вероятностью занимает любую из $n^{(1)}$ позиций набора (x_{1i});

в) из состояния $(\mathrm{n} + \mathrm{e}_1 - \mathrm{e}_2, T_{ij}^{-1}\mathrm{x}, \mathrm{y})$, при переходе сообщения с первой фазы на вторую, при этом длительность сообщения x_{1i}, передача которого заканчивается, должна находиться в $h/(n_i^{(1)} + 1)$ -окрестности нуля, а остаточное время передачи сообщения на второй фазе x_{2j} принимает значение $x_{2j} = (\mu^{(1)}/\mu^{(2)}) y_i$, при этом на следующей фазе оно займет с равной вероятностью любую из $n^{(2)}$ позиций;

г) из состояния $(\mathrm{n} + \mathrm{e}_2, D_i^{-1}(\mathrm{x}, \mathrm{y}))$, при окончании передачи сообщения, при этом длительность сообщения, передача которого заканчивается, должна находиться в $h/(n^{(2)} + 1)$ - окрестности нуля.

Непосредственной проверкой легко убедиться, что плотность распределения вероятностей (15) удовлетворяет уравнению (16), при этом необходимо учитывать связь между распределениями длительностей передачи сообщений на различных фазах, даваемую соотношением (14).

Заметим, что выражение (15) имеет мультипликативную форму, в которой транзитным сообщениям соответствует плотность

$$\tilde{p}(\mathbf{n}, \mathbf{x}, \mathbf{y}) = (1-\rho) \prod_{1 \le i \le n} \lambda \Theta(y_i - x_i) g(\mu y_i),$$

а выходящим – плотность

$$\hat{p}(\mathbf{n}, \mathbf{x}) = (1-\rho) \prod_{1 \le j \le n} \lambda \left(1 - G(\mu x_j)\right).$$

Распределение вероятностей макросостояний для этой системы, которое получается интегрированием соотношений (15) по переменным x, y, имеет вид

$$p(n_1, n_2) = (1-\rho_1)(1-\rho_2)\rho_1^{n_1}\rho_2^{n_2},$$

который также показывает возможность декомпозиции соответствующей сети и нечувствительность стационарных вероятностей макросостояний к виду распределения объемов сообщений.

Не трудно проверить, что для тандем-системы, передающей r потоков сообщений с интенсивностями λ_i и распределениями объемов $G_i(\cdot), i=\overline{1,r}$ со скоростями $\mu^{(1)}, \mu^{(2)}$ на первой и второй фазах сохраняются аналогичные утверждения о мультипликативном представлении ее стационарных вероятностей состояний, причем транзитным сообщениям соответствует плотность

(17) $$\tilde{p}(\mathbf{n},\mathbf{x},\mathbf{y}) = (1-\rho)\frac{n!}{n_1!\cdots n_r!}\prod_{1\le i\le r}\prod_{1\le j\le n_i}\lambda_i\Theta(y_i-x_i)g(\mu^{(1)}y_i),$$

а выходящим, как и в случае независимых длительностей передачи, - плотность, задаваемая формулой (7).

6. НЕИЗМЕННЫЕ ОБЪЕМЫ СООБЩЕНИЙ. МОДЕЛЬ СЕТИ

На основании полученных результатов исследуем теперь поведение двухуровневой сети в случае, когда объемы сообщений произвольно распределены с ФР $G_{ij}(x)$ для сообщений типа (ij), но не изменяются при их передаче по различным каналам сети. Для исследования этой модели расширим пространство состояний системы, введенное а разделе 4, дополнительными переменными $\mathrm{y}=(y_{ij}^{(1)}, y_j^{(2)})$, указывающими исходные длительности передачи сообщений, находящихся на первом и втором уровнях сети. Обозначим через X(t) процесс X(t)=(n, x, y), если в момент t система находится в состоянии (n, x, y) и предположим, что плотности (по переменным $x_{ij}^{(1)}, x_j^{(2)}, y_{ij}^{(1)}, y_j^{(2)}$) распределения вероятностей его состояний существуют. Для процесса X(t) справедлива теорема

<u>**Теорема 4.**</u> При условии, что процесс X(t) положительно возвратен по Харрису, предельные стационарные плотности распределения вероятностей его состояний p(n,x,y) существуют и представимы в виде произведения

(18) $$p(\mathbf{n},\mathbf{x}) = \prod_{1\le i\le r}\tilde{p}(\mathbf{n}_i^{(1)},\mathbf{x}_i^{(1)},\mathbf{y}_i^{(1)}) \times \hat{p}(\mathbf{n}^{(2)},\mathbf{x}^{(2)},\mathbf{y}_i^{(2)}),$$

где плотности распределения вероятностей состояний узлов сети для транзитных $\tilde{p}(\mathrm{n},\mathrm{x},\mathrm{y})$ и выходящих $\hat{p}(\mathrm{n},\mathrm{x},\mathrm{y})$ задаются формулами (17), (7) со значениями интенсивностей сообщений, определяемыми формулами (9), и распределениями длительностей их передачи - соотношениями.

$$B_{ij}^{(1)}(x) = G_{ij}(\mu_i^{(1)}x), \quad j\neq i,$$

$$B_{ii}^{(1)}(x) = \frac{\lambda_{ii}}{\lambda_i^{(2)}}G_{ii}(\mu_i^{(1)}x) + \sum_{j\neq i}\frac{\lambda_{ji}}{\lambda_i^{(2)}}G_{ji}(\mu_i^{(1)}x),$$

$$B_i^{(2)}(x) = \sum_{j\neq i}\frac{\lambda_{ji}}{\lambda_i^{(2)}}G_{ji}(\mu(2)x).$$

Доказательство. Если процесс X(t) положительно возвратен но Харрису, то согласно теории марковских процессов его предельные вероятности состояний существуют, а система уравнений для их плотностей получается, как и ранее, из системы уравнений Колмогорова.

Для удобства записи этой системы уравнений используем, как и ранее, операторы преобразования состояний системы, соответствующие поступлению сообщений A_{ijl}, передаче их с одного уровня на другой $T_{ij}^{(12)}(lm)$, и окончанию обслуживания D_{iil}. Действие этих операторов в рассматриваемом случае отличается от предыдущих и объединяет их описание в разделах 4 и 5. Оператор A_{ijl} добавляет компоненты $x_{ijl}^{(1)}, y_{ijl}^{(1)}$ в наборы (x,y) одновременно, оператор $T_{ij}^{(12)}(lm)$ действует на границе области определения процесса, при обращении компоненты $x_{ijl}^{(1)}$ в ноль, удаляет ее и присваивает компоненте $x_{jm}^{(2)}$ значение времени передачи сообщения, соответствующее значению $y_{ijl}^{(1)}$ с сохранением его объема, т.е. в соответствии с соотношением $\mu_i^{(1)} y_{ijl}^{(1)} = \mu^{(2)}, x_{jm}^{(2)}$ одновременно удаляя последнюю, оператор $T_{ij}^{(21)}(ml)$ действует аналогично при переходе сообщения со второго на первый уровень, наконец оператор D_{iil} удаляет компоненту $x_{iil}^{(1)}$ со значением ноль.

Заметим, кроме того, что уравнения. Колмогорова имеют место в области допустимых значений процесса, которая в рассматриваемом случае имеет вид

$$0 \le x^{(1)}_{ijl} \le y^{(1)}_{ijl} < \infty,\quad l=\overline{1,n^{(1)}_{ij}},\ j\ne i,\quad i,j=\overline{1,r},$$

$$0 \le x^{(1)}_{iil} < \infty,\quad l=\overline{1,n^{(1)}_{ii}},\quad i=\overline{1,r}$$

$$0 \le x^{(2)}_{jl} \le y^{(2)}_{jl} < \infty,\quad l=\overline{1,n_j^{(2)}},\quad j=\overline{1,r}.$$

На границах этой области

$$x^{(1)}_{ijl}=y^{(1)}_{ijl},\quad l=\overline{1,n^{(1)}_{ij}}\quad j\ne i,\quad i,j=\overline{1,r},$$

и

$$x^{(2)}_{jm}=y^{(2)}_{jm},\quad m=\overline{1,n_j^{(2)}}\quad j=\overline{1,r},$$

соответствующих поступлению сообщений и их переход с одного уровня на другой, одновременно изменяются две координаты процесса, что необходимо учитывать с помощью специальных граничных условий. Однако, используя, как и ранее, δ - функции Дирака, мы включим эти граничные условия в общую систему уравнений.

Заметим, кроме того, что в приводимых ниже уравнениях учитывается связь (1) между распределениями объемов сообщений и длительностями их передачи. С использованием введенных операторов и с учетом сделанных замечаний уравнения равновесия для стационарных плотностей вероятностей процесса *X(t)* имеют вид

$$
\begin{aligned}
&-\sum_{i,j,l}\frac{1}{n_i^{(1)}}\frac{\partial p(\mathbf{n},\mathbf{x},\mathbf{y})}{\partial x_{ijl}^{(1)}}-\frac{1}{n^{(2)}}\sum_{j,m}\frac{\partial p(\mathbf{n},\mathbf{x},\mathbf{y})}{\partial x_{jm}^{(2)}}+\lambda p(\mathbf{n},\mathbf{x},\mathbf{y})=\\
&\quad=\sum_{i,j,l}\frac{\lambda_{ij}}{n_{ij}^{(1)}}p(\mathbf{n}-\mathbf{e}_{ij}^{(1)},A_{ijl}^{-1}(\mathbf{x},\mathbf{y}))g_{ij}(\mu_i^{(1)}y_{ijl}^{(1)})\delta(y_{ijl}^{(1)}-x_{ijl}^{(1)})+\\
&\quad+\sum_{i,j\neq i,l,m}\frac{1}{(n_i^{(1)}+1)n_j^{(2)}}p(\mathbf{n}+\mathbf{e}_{ij}^{(1)}-\mathbf{e}_j^{(2)},\left(T_{ij}^{(12)}(lm)\right)^{-1}(\mathbf{x},\mathbf{y}))\delta(y_{jm}^{(2)}-x_{jm}^{(2)})+\\
&\quad+\sum_{j,l,m}\frac{1}{n_{jj}^{(1)}(n^{(2)}+1)}p(\mathbf{n}+\mathbf{e}_j^{(2)}-\mathbf{e}_{jj}^{(1)}\left(T_j^{(21)}(ml)\right)^{-1}(\mathbf{x},\mathbf{y}))\delta(y_{jjl}^{(1)}-x_{jjl}^{(1)})+\\
&\quad+\sum_{i,l}\frac{1}{n_i^{(1)}+1}p(\mathbf{n}+\mathbf{e}_{ii}^{(1)},D_{iil}^{-1}\mathbf{x}).
\end{aligned}
\tag{19}
$$

Аргументация относительно вывода этих уравнений аналогична изложенной в разделах 4, 5, и мы ее опускаем. Непосредственной проверкой можно убедиться, что выражение (18) удовлетворяет этой системе уравнений.

Распределение вероятностей макросостояний для этой системы получается, как и ранее, интегрированием по переменным *(x,y)* и имеет вид (10), совпадающий с распределением вероятностей для системы с независимыми длительностями передачи сообщений.

Переход к выражениям для распределения вероятностей через распределения объемов сообщений осуществляется, как и ранее, простой заменой переменных.

Этот результат помимо возможности декомпозиции сети показывает также, что дисциплина разделения времени обеспечивает нечувствительность стационарного распределения вероятностей макросостояний системы не только к виду законов распределения длительностей обслуживания, но также, в некотором смысле, к структуре процессов обслуживания.

7. ЗАКЛЮЧЕНИЕ

В статье, которая является первой частью работы, посвященной проблемам декомпозиции сложных систем, приводится обобщение мультипликативного представления для стационарных вероятностей состояний сети массового обслуживания на случай зависимых длительностей передачи сообщений по каналам, который в действительности имеет место в реальных сетях передачи данных. Именно: предполагается, что сообщения имеют случайные объемы, которые, однако, остаются неизменными в процессе их передачи по сети, а длительности их передачи на различных каналах зависят от скорости последних и пропорциональны этим объемам. Полученные результаты позволяют рассчитывать различные характеристики производительности сетей в процессе их проектирования.

СПИСОК ЛИТЕРАТУРЫ

[1] Рыков В. В. О декомпозиции иерархических информационно-вычислительных сетей. Вестник РУДН, 1997 (в печати)

[2] Храмцов П. Лабиринт Internet. 000 "Электроинформ", 1996.

[3] Kleinrock L. Queueing systems. V. 1: Theory, John Wiley <k Sons. New York, 1975

[4] Kleinrock L. Queueing systems. V. 2: Computer Applications. John Wiley k. Sons. New York, 1976

[5] Шварц М. Сети связи: протоколы, моделирование и анализ. М.: "Наука", 1992.

[6] Бочаров П.П.. Башарин Г.П., Коган А.Я. Анализ очередей в вычислительных сетях. Теория и методы расчета. JVL: "Наука", 1989.

[7] Жожикашвили В.А., Вишневский В.М. Сети массового обслуживания. Теория и применения к сетям ЭВМ. М.: Радио и связь. 1988.

[8] Jackson J.R. Networks of wating lines. Operat. Res., V. 5 (1957), N 4, pp. 518-521.

[9] Jackson J.R. Job Shop-like Queueing Systems. Manag. Sci. V 10 (1963), N 1, pp. 131-142.

[10] BaskeU F., Chandy K.H, Muntz R.R., Falacios F.G. Open, closed and mixed networks of queues with different classes of customers. J. of ACM, V. 22 (1975), No 22, pp. 248-260.

[11] Башарин Г.П., Толмачев А.Л. Теория сетей массового обслуживания и ее приложения к анализу информационно-вычислительных систем. Итоги науки и техники. Теория вероятностей. Метем, статистика. Теоретич. киберн. М.: ВИНИТИ. Т. 21, 1983, 3-120.

[12] Bahama S. Properties and analysis of queueing network models with finite capacities. Performance'93. Rome, Springer-Verlag, 1993.

[13] Kelly F.F. Reversibility and Stochastic Networks. John Wiley & Sons. Chichester, U.K., 1979.

[14] Кельберт М.Я., Сухов Ю.М. Математические вопросы теории сетей с очередями. Итоги науки и техники. Теория вероятностей. Метем, статистика. Теоретич. киберн. Т. 26, с. 3-97. М.: ВИНИТИ, 1988.

[15] Нуммелин Э. Общие неприводимые цепи Маркова и неотрицательные операторы. М.: Мир, 1989.

[16] Климов Г.П. Стохастические системы обслуживания. М.: Наука, 1966.