Вероятностные модели компьютерных архитектур

А. В. ФИЛИН
Институт проблем информатики, Москва

В. А. МАЛЫШЕВ
I.N.R.I.A.

А. Д. МАНИТА*
Московский государственный университет им. М. В. Ломоносова

УДК 519.248:62+519.872

Ключевые слова: компьютерные архитектуры, приоритетные системы массового обслуживания, динамические системы, жидкостная аппроксимация.

Аннотация

Связь между приоритетными системами массового обслуживания и компьютерными архитектурами хорошо известна. Но, насколько нам известно, до сих пор отсутствовали формулировки точных моделей для достаточно общих компьютерных архитектур. Эта работа преследует две цели: первая состоит в том, чтобы предложить такие формулировки на точном математическом языке; вторая, и более важная, состоит в том, чтобы предложить новый подход к приоритетным сетям в целом. Этот подход основан на недавних успехах, связанных с применением динамических систем в сетях массового обслуживания, что в частных случаях представляет собой хорошо известную жидкостную аппроксимацию. Это приводит к новому подходу к оценке производительности заданной компьютерной архитектуры. Здесь мы применяем этот метод к простейшей одношинной архитектуре. Эту работу следует рассматривать как первый шаг в развитии этого подхода.

Abstract

A. V. Filin, V. A. Malyshev, A. D. Manita, Probabilistic models for computer architectures, Fundamentalnaya i prikladnaya matematika vol. 3(1997), № 1, p. 263–301.

Connections between priority queueing models and computer architectures are widely known. But, as far as we know, there was no formulation of exact models for sufficiently general computer architecture models. This paper has two goals: the first and the smaller one is just to give this formulation in exact mathematical terms. The second and the most important one is to present a new approach to

*При частичной поддержке Франко-русского центра им. А. М. Ляпунова и грантов Российского фонда фундаментальных исследований № 95-01-00018 и INTAS № 0893.

Фундаментальная и прикладная математика 1997, том 3, № 1, с. 263–301.

priority networks themselves. This approach is based on recent advances in the dynamical system approach to queueing networks, which in some very particular cases becomes a well-known fluid approximation. This gives a new approach to performance evaluation of a given computer architecture. We apply this method here to the simplest architecture with the unique bus. This paper can be considered as the first step in the development of this approach.

1 Введение

Цель этой и последующих работ — предложить некоторые глобальные вероятностные модели компьютерных архитектур. Математическое моделирование компьютерных архитектур предполагает знание статистической структуры всех потоков данных между различными частями компьютера. Понятно что точное знание этой статистики невозможно, так как она сильно зависит от задач, решаемых компьютером в данный момент времени, и, более того, она также зависит от «психологии» программиста. Но, тем не менее, некоторые модели могут быть полезными, если

i) они приводят к интересной математике;

ii) они дают качественное понимание некоторых практических явлений, которые могут быть интересными для проектировщиков компьютеров;

iii) принимая во внимание, что книги по компьютерным архитектурам непонятны для математиков, формулирование на строгом уровне точных моделей могло бы сделать эту область более притягательной для математиков.

Компьютер может представлять из себя очень большую и сложную систему, на которую можно смотреть с разных точек зрения и на разных шкалах.

Точка зрения, которую мы выбираем здесь, состоит в том, чтобы представить компьютер как *большую приоритетную систему массового обслуживания.*

Тем самым, наши модели отличаются от моделей, рассмотренных в работе [1], которая имеет дело с неприоритетными дисциплинами: первым пришел — первым обслужился, time sharing, последним пришел — первым обслужился, дисциплина без ожидания.

В своем изложении мы смешиваем нестрогие рассуждения со строгими моделями. Поскольку наша работа ориентирована на математиков, мы объясняем некоторые вещи, которые тривиальны для специалистов по компьютерным архитектурам.

Мы представляем здесь модели в порядке возрастания сложности. Параграфы 2.1 и 2.2 охватывают уровень «железа» (hardware) и микропрограммный уровень. Параграф 2.3 содержит также программный уровень. В математической части работы мы анализируем простейшую одношинную архитектуру (см. рис. 1).

2 Модели прямого доступа

2.1 Простейшая модель

На рис. 1 показана структура с одной шиной, используемой всеми устройствами компьютера.

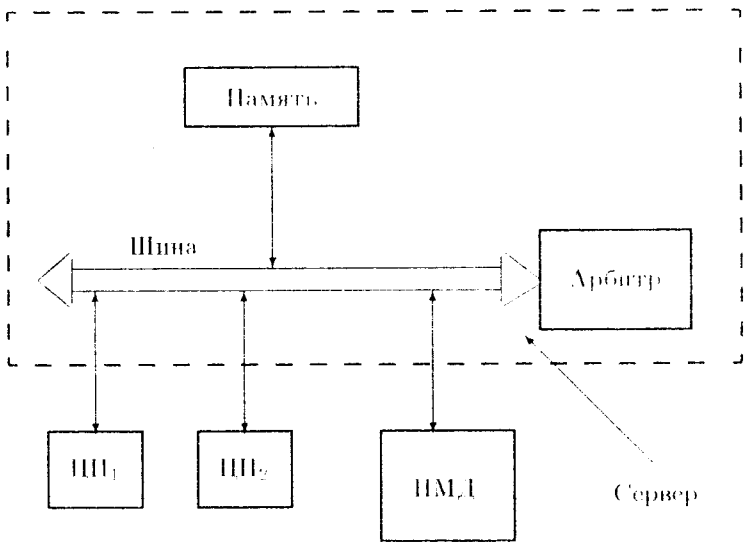

Рис. 1. Структура компьютера с одной шиной

Эта структура (модель 1) обладает следующими свойствами:

• Под памятью здесь мы понимаем основную память. Между памятью и любой другой единицей структуры возможна связь в обоих направлениях.

• Только один акт связи (память $\longleftrightarrow$ другое устройство) может иметь место в заданный момент времени, так как существует только один ресурс (шина) для этой связи.

• В качестве единиц A_i могут выступать центральные (ЦП) или периферийные процессоры, накопители на магнитных дисках (НМД), стриммеры, устройства ввода-вывода. Устройство называется устройством прямого доступа, если оно может контролировать без участия процессора обмен информацией между собой и памятью.

• Система «память—шина—арбитр» рассматривается как сервер (обслуживающий прибор) в смысле теории массового обслуживания, а устройства A_i — как источники требований, которые должны быть обслужены сервером. Наше основное предположение состоит в следующем.

Предположение А. *Все источники и их требования независимы друг от друга.*

Строго говоря, это предположение всегда ложно. Например, в зависимости от исполняемой программы память должна обращаться за данными к НМД. Это противоречит предположению о том, что НМД шлет свои требования независимо. Но вместе с тем ясно, что эта аппроксимация может быть разумной в ряде случаев.

• Задана некоторая приоритетная дисциплина P_a. К примеру, это может быть относительный приоритет P_{rel}, абсолютный приоритет P_{abs} (с дообслуживанием или без) или другие. Каждому устройству A_i приписан номер приоритета $P(i)$. Проблема состоит в том, чтобы найти функцию $P(i)$, которая лучше подходит для некоторых важных показателей компьютерной системы.

• Некоторые из устройств A_i рассматриваются как «бесконечные источники», то есть они шлют бесконечную последовательность единичных требований. Другие рассматриваются как «конечные источники», если каждое требование из узла i пытается попасть на обслуживающий прибор (в узел 0) независимо от остальных требований того же узла и всего имеется конечное число этих требований.

• Основные параметры следующие:

(а) для системы в целом: распределения занятых и незанятых периодов и длин очередей;

(б) для устройств также время ожидания обслуживания.

Теперь мы опишем *математическую модель массового обслуживания* (см. рис. 2) в более точных терминах.

Мы рассмотрим сеть типа BCMP со смешанной (открыто-замкнутой) структурой (см. [1]). Есть «центральный» узел 0 и k других узлов $1, \ldots, k$ (которые соответствуют конечным источникам). Каждому $i = 1, \ldots, k$ соответствуют N_i требований, m_i из которых располагаются в очереди к узлу 0, а прочие $N - m_i$ находятся в узле i. Каждое требование из узла i независимо от других переходит в узел 0 через экспоненциальное время с интенсивностью λ_i и возвращается в узел i после обслуживания в узле 0. В узле 0 выбирается некоторая дисциплина P_a.

Существуют также внешние входящие потоки $k+1, \ldots, N$. Предполагается, что они пуассоновские с интенсивностями λ_i. Возможны также ситуации (таймеры, ...), когда требования поступают в виде регулярной последовательности. Требования становятся в очередь к узлу 0 и обслуживаются им в соответствии с дисциплиной P_a. После завершения обслуживания они покидают систему. Времена обслуживания требований типа i независимы и имеют (произвольное) распределение S_i.

Приоритетные дисциплины.

Дисциплина обслуживания — это правило, определяющее какое из требований, ожидающих обслуживания, должно занять обслуживающий узел, когда

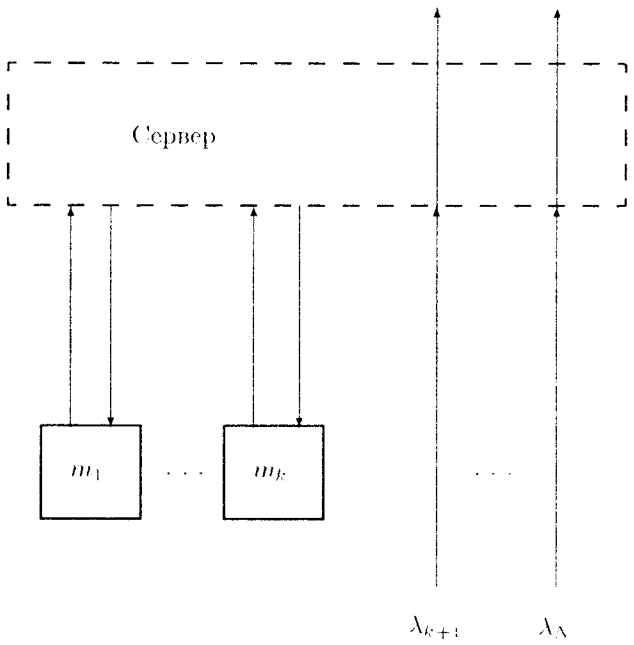

Рис. 2. Математическая модель

тот будет готов обслужить очередное требование. Приоритетные дисциплины — это специальный класс дисциплин, которые предполагают, что требования бывают различных типов и эти типы имеют разную важность. Обычно типы требований нумеруются натуральными числами в порядке убывания важности: тип i имеет приоритет над типом j, если $i < j$. Различают *абсолютный* и *относительный* приоритеты. Различие между ними состоит в том, что происходит в ситуации, когда сервер занят некоторым требованием, но в это же время в систему поступает другое требование с более высоким приоритетом.

В случае дисциплины *абсолютный* приоритет требование с высшим приоритетом прерывает обслуживание требования с низшим приоритетом и немедленно занимает обслуживающий прибор.

1. Дисциплина обслуживания называется абсолютным приоритетом *с дообслуживанием*, если требование, обслуживание которого было прервано, должно вернуться на узел обслуживания (когда в системе не окажется требований с более высоким приоритетом) и обслуживаться в течении времени, оставшегося после предыдущего обслуживания этого требования.

2. Дисциплина называется абсолютным приоритетом *без дообслуживания*,

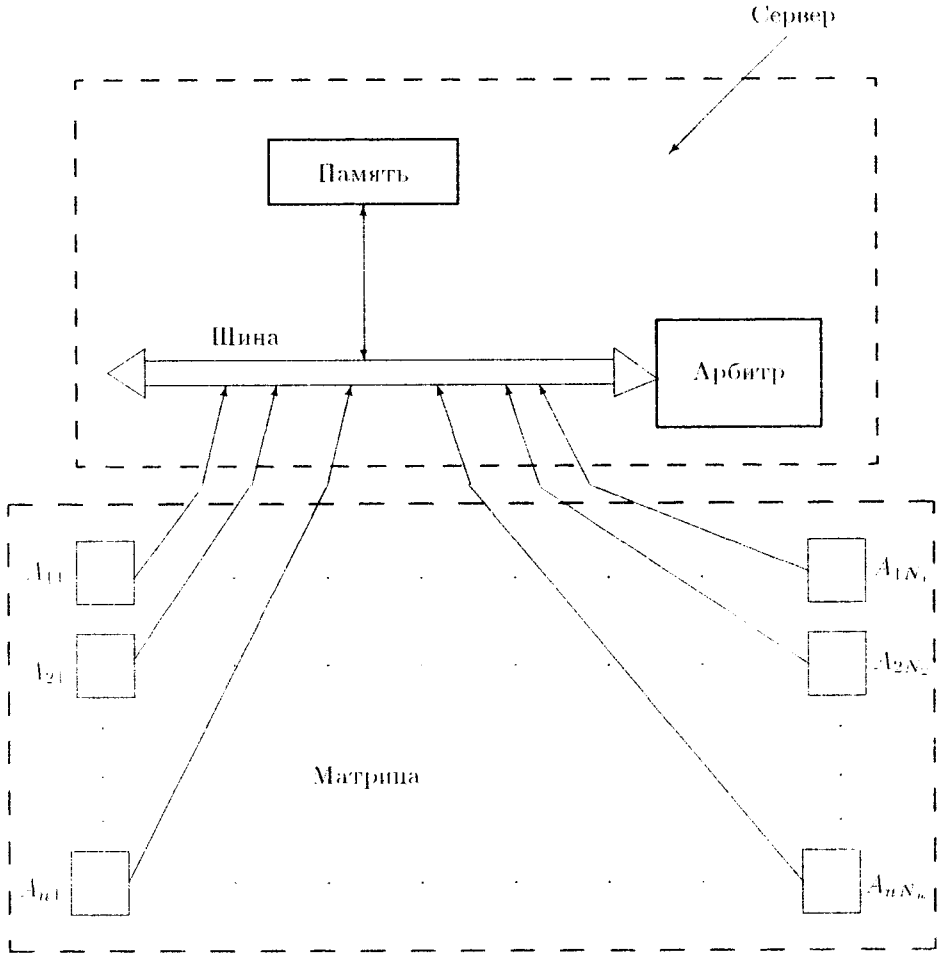

Рис. 3. Матричная модель

если требование, обслуживание которого было прервано, покидает систему.

В случае *относительного* приоритета требование с высшим приоритетом ожидает, пока сервер не закончит обслуживание требования с низшим приоритетом.

2.2 Матричная модель

Здесь ситуация такая же, как и раньше, но модули (устройства) занумерованы двумя индексами (см. рис. 3):

$$A_{ij},\ i = 1, \ldots, n;\ j = 1, \ldots, N_i; \quad \sum_{i=1}^{n} N_i = N.$$

Приоритеты выстраиваются в лексикографическом порядке. Индекс i определяет уровень, приоритеты между уровнями — абсолютные с дообслуживанием (наиболее частый случай) или иногда относительные. Внутри уровней — дисциплина относительный приоритет.

2.3 Система прямого доступа с программируемой подсистемой

Рассмотрим следующую модель *системы прямого доступа с простейшей подсистемой программируемых устройств* (см. рис. 4).

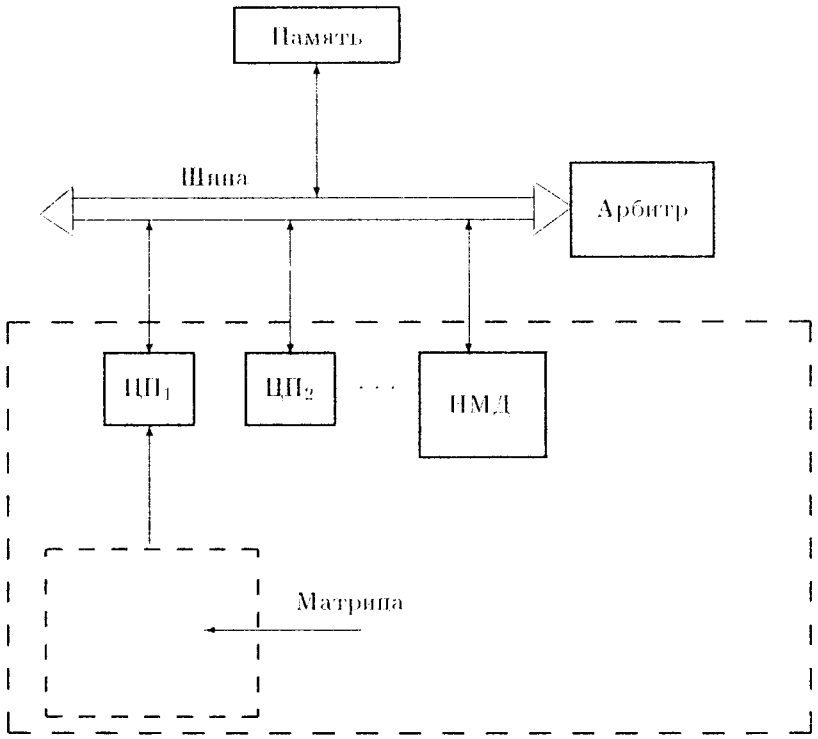

Рис. 4. Система с подсистемой программируемых устройств

Ситуация такая же, как на рис. 1, только у одного из модулей (процессор A_1) есть матричная подсистема. Этот процессор A_1 является сервером для этой матричной подсистемы, которая подчинена дисциплине, описанной

в параграфе 2.2. В то же время процессор A_1 выставляет требования к шине (основному серверу).

Предположение Б. *Процессор A_1 выставляет требования к шине только тогда, когда он работает как сервер (т. е. когда существует очередь к A_1). В остальном его требования не зависят от других деталей (какое устройство a_{ij} обслуживается процессором, какого рода требования идут к этому процессору от a_{ij} и т. п.). Более точно, процессор становится конечным или бесконечным источником в течении времени, когда он является сервером.*

Это предположения также ложно во всех случаях, но оно может служить грубой аппроксимацией и дать некоторую качественную информацию. Если A_1 — конечный источник, это предположение нуждается в уточнении, что может быть сделано различными способами.

2.4 Несколько подсистем программируемых устройств

На рис. 5 показана более сложная система, которая получена путем суперпозиции более простых, рассмотренных в предыдущем параграфе.

Пирамидальный компьютер для распознавания образов

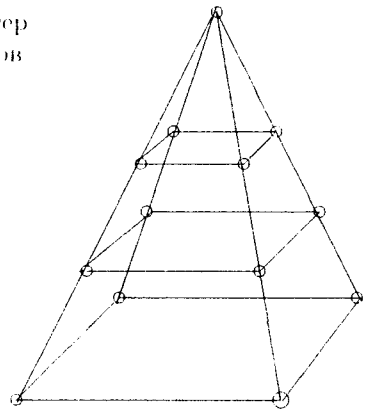

Рис. 5. Пирамида

Так, некоторые из A_i могут быть серверами для матричных подсистем $M(i) = (a^i_{jk})$. Некоторые из a^i_{jk} (если они являются процессорами) могут быть серверами для других матричных подсистем $M(i; j, k)$ и так далее по индукции. Однако предполагается, что любой процессор может быть сервером для самое большее одной матричной подсистемы. Таким образом, мы имеем ситуацию рис. 6.

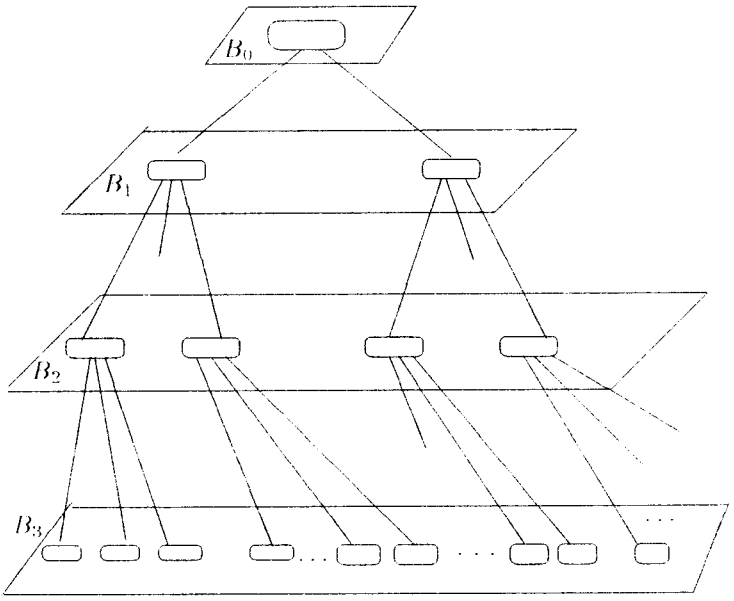

Рис. 6. Иерархия матричных подсистем

Итак, мы имеем конечную цепь процессоров B_0, $A_i = B_1$, B_2, ..., B_n, где B_0 — основной сервер (память—шина—арбитр), а B_i — сервер для B_{i+1}, $i = 0, \ldots, n-1$. Основное предположение аналогично предположению Б.

Предположение В. *Процессор B_i может выставлять требования к процессору B_{i-1} только в течении времени, когда он работает как сервер (т. е. когда есть непустая очередь требований от B_{i+1}).*

3 Вероятностный анализ

Мы будем рассматривать здесь модель параграфа 2.1.

Настоящий раздел организован следующим образом. В параграфе 3.1 мы даем грубый вероятностный анализ системы бесконечных источников с абсолютным приоритетом с дообслуживанием. Мы не делаем никаких ограничительных предположений относительно входных потоков и времен обслуживания.

В следующем параграфе мы делаем некоторые предположения, позволяющие использовать марковское описание приоритетных систем. Мы рассматриваем следующие дисциплины: абсолютный приоритет с дообслуживанием, абсолютный приоритет без дообслуживания, относительный приоритет. В этом случае приоритетная система попадает в специальный класс счетных

цепей Маркова. Качественный анализ таких приоритетных сетей массового обслуживания можно вести к изучению специального класса случайных блужданий в $\mathbf{Z}_+^\nu$. Этот анализ использует идеологию *индуцированных цепей Маркова* и *второго векторного поля* для ассоциированных случайных блужданий (см. [3,5]). Этот подход оказался успешным в решении ряда задач ([2,3,5,9]). В добавлении мы приводим краткую сводку обозначений, определений и некоторых результатов, которыми мы будем активно пользоваться.

Наша цель состоит в нахождении условий эргодичности приоритетных систем и вычислении некоторых параметров эргодичного режима (среднее время ожидания, среднее время обслуживания, и т. п.). Математическое понятие эргодичности или устойчивости систем массового обслуживания соответствует ситуации, когда средняя длина очереди остается ограниченной с ростом времени. Свойство эргодичности важно для разработчиков компьютерных архитектур, так как оно означает неперегруженность системы: ситуация, когда длины очередей, соответвующие некоторым типам, стремятся к бесконечности, практически означает, что вновь приходящие требования этих типов не будут обслужены системой за реальное время.

Для того, чтобы сделать наш вероятностный анализ более прозрачным, мы начнем с простого случая двух типов требований, потом мы рассмотрим систему, состоящую из нескольких бесконечных или конечных источников, и, в заключение, составную приоритетную систему, состоящую из бесконечных и конечных источников.

Аналогичный анализ может быть сделан также для иерархической модели параграфа 2.3.

3.1 Абсолютный приоритет в общей ситуации

Мы рассматриваем дисциплину абсолютный приоритет с дообслуживанием. Предположим сначала, что все источники бесконечны. Мы не предполагаем, что входные потоки пуассоновские (но предположение A важно). Пусть L_i — среднее время между последовательными поступлениями требований i-го типа, а M_i — среднее время обслуживания. Определим нагрузку $\rho_i = \dfrac{M_i}{L_i}$. По соглашению тип i имеет абсолютный приритет над типом j, если $i < j$. Рассмотрим достаточно большой интервал времени $[0, T]$. В течение этого времени приблизительно поступят $\dfrac{T}{L_1}$ требований типа 1. Среднее время, необходимое для обслуживания всех этих требований, равно $\rho_1 T$. Шина свободна от требований типа 1 в течение времени $(1 - \rho_1)T$. То же самое верно для требований типа i: среднее число поступивших требований равно $\dfrac{T}{L_i}$, их среднее время обслуживания равно $\rho_i T$. Этот факт легко установить по индукции, при этом существенно, что требования обслуживаются в соответствии с дисциплиной

абсолютный приоритет с дообслуживанием, тем самым, они как бы «не замечают» интервалов прерывания. Таким образом, мы приходим к следующему утверждению.

Предложение 1. *Предположим, что все источники бесконечные. Тогда система эргодична*[1], *если*

$$\sum_{i=1}^{N} \rho_i < 1.$$

Среднее время, когда система свободна от требований всех типов, равно

$$\left(1 - \sum_{i=1}^{N} \rho_i\right) T.$$

Полезным является следующее понятие *эффективного времени* для i-го типа требований. Так как среднее время, когда система готова обслуживать требования i-го типа, равно $\left(1 - \sum_{j=1}^{i-1} \rho_j\right) T$, это может быть проинтерпретировано, как будто бы «внутреннее» время для типа i становится в $\left(1 - \sum_{j=1}^{i-1} \rho_j\right)^{-1}$ раз медленнее. Тогда верно следующее.

Предложение 2. *Время ожидания для i-го типа требований равно*

$$W_i = (1 - \rho_1 - \ldots - \rho_{i-1})^{-1} W_1^0,$$

где W_1^0 есть среднее время ожидания требования типа i в случае, когда тип i имеет наивысший приоритет (или другие типы отсутствуют вообще). Среднее время обслуживания требований типа i (с учетом возможных прерываний) равно

$$(1 - \rho_1 - \ldots - \rho_{i-1})^{-1} M_i.$$

Случай, когда есть также «конечные» источники, более сложен, но может быть проанализирован аналогичным образом.

3.2 Случай двух типов

Мы покажем в этом и последующих параграфах, что системы массового обслуживания с дисциплиной абсолютный приоритет при некоторых предположениях относительно входных потоков и распределений времен обслуживания можно расматривать как марковские процессы с пространством состояний $\mathbf{Z}_+^\nu$. Эти марковские процессы представляют собой максимально однородные случайные блуждания в непрерывном времени. Состояниями $x(t)$

[1] В этом предложении *эргодичность* означает, что существует стационарный режим, в котором средние длины всех очередей конечны.

таких процессов являются векторы, компоненты $x_{ij}(t)$ которых равны числу требований i-го типа в узле j. В случае относительного приоритета систему массового обслуживания можно свести к марковскому процессу на конечном объединении «октантов» $\mathbf{Z}_+^\nu$. Более того, стохастическая динамика внутри «октантов» $\mathbf{Z}_+^\nu$ также представляет собой максимально однородные случайные блуждания в непрерывном времени. Благодаря этому для изучения этих систем массового обслуживания мы можем применить новые конструктивные методы анализа счетных цепей Маркова (см. [3, 5]).

Оказывается, что для специального класса случайных блужданий, ассоциированных с приоритетными системами, можно в явном виде построить *второе векторное поле*. Этот подход приводит к конструктивным и практически содержательным методам строгого качественного анализа приоритетных систем массового обслуживания.

Ниже мы используем обозначения из [3, 5]. Определение индуцированной цепи и основные результаты, связанные с классификацией случайных блужданий в терминах второго векторного поля (ВВП), можно найти в главе 4 в [3].

3.2.1 Бесконечные источники: абсолютный приоритет

Рассмотрим простейший пример системы с одним узлом и двумя типами требований. Требования типа 1 имеют приоритет над требованиями типа 2. Предположим также, что входные потоки требований пуассоновские и что времена обслуживания распределены экспоненциально. Пусть λ_i — интенсивность входного потока требований типа i, μ_i — интенсивность обслуживания требований типа i. Пусть n_1 и n_2 обозначают количества требований типов 1 и 2 в системе. Таким образом, любому состоянию системы мы можем поставить в соответствие точку положительного квадранта $\mathbf{Z}_+^2$. Качественное поведение рассматриваемой приоритетной системы связано с качественным поведением *ассоциированного* случайного блуждания в $\mathbf{Z}_+^2$, которое мы определим ниже.

Выберем интенсивности скачков $\lambda_{\alpha\beta}$ следующим образом.

Если $\alpha \in B^{\{1,2\}} = \{(n_1, n_2)\colon\ n_1 > 0,\ n_2 > 0\}$ или $\alpha \in B^{\{1\}} = \{(n_1, 0)\colon\ n_1 > 0\}$, мы положим

$$\lambda_{\alpha\beta} = \begin{cases} \mu_1, & \text{если } \beta - \alpha = (-1, 0), \\ \lambda_1, & \text{если } \beta - \alpha = (1, 0), \\ \lambda_2, & \text{если } \beta - \alpha = (0, 1), \\ 0 & \text{для всех прочих } \beta \neq \alpha \end{cases}$$

для всех приоритетных дисциплин, упомянутых выше.

Если $\alpha \in B^{\{2\}} = \{(0, n_2)\colon\ n_2 > 0\}$ и дисциплина обслуживания есть *абсолютный приоритет с дообслуживанием*, мы полагаем

$$\lambda_{\alpha\beta} = \begin{cases} \lambda_1, & \text{если } \beta - \alpha = (1, 0), \\ \lambda_2, & \text{если } \beta - \alpha = (0, 1), \\ \mu_2, & \text{если } \beta - \alpha = (0, -1), \\ 0 & \text{для всех прочих } \beta \neq \alpha. \end{cases}$$

Если $\alpha \in B^{\{2\}}$ и дисциплина обслуживания есть *абсолютный приоритет без дообслуживания*, мы полагаем

$$\lambda_{\alpha\beta} = \begin{cases} \lambda_1, & \text{если } \beta - \alpha = (1,-1), \\ \lambda_2, & \text{если } \beta - \alpha = (0,1), \\ \mu_2, & \text{если } \beta - \alpha = (0,-1), \\ 0 & \text{для всех прочих } \beta \neq \alpha. \end{cases}$$

Для $\alpha = (0,0)$ мы положим

$$\lambda_{0\beta} = \begin{cases} \lambda_1, & \text{если } \beta = (1,0), \\ \lambda_2, & \text{если } \beta = (0,1), \\ 0 & \text{для всех прочих } \beta \neq 0 \end{cases}$$

в обоих случаях.

Теперь мы явно вычислим *второе векторное поле* в обоих случаях.

Прежде всего рассмотрим ВВП $v^{\{1,2\}}$ на грани $B^{\{1,2\}}$:

$$v^{\{1,2\}} = (\lambda_1 - \mu_1, \lambda_2).$$

Заметим, что $v_2^{\{1,2\}} > 0$, так что грань $B^{\{1\}} = \{(n_1, 0)\colon\ n_1 > 0\}$ всегда неэргодична. Грань $B^{\{2\}}$ эргодична, если $\lambda_1 < \mu_1$, и неэргодична, если $\lambda_1 \geqslant \mu_1$. Предположим, что $\lambda_1 < \mu_1$. Рассмотрим индуцированную цепь $\mathcal{L}^{\{2\}}$. Это обратимая цепь Маркова, стационарное распределение которой можно выписать явно:

$$\pi_i = \frac{1}{1 - \lambda_1/\mu_1} \left(\frac{\lambda_1}{\mu_1}\right)^i, \quad i \geqslant 0.$$

Таким образом, в случае *абсолютного приоритета с дообслуживанием* ВВП для грани $B^{\{2\}}$ равно

$$v^{\{2\}} = (\lambda_2 - \mu_2)\pi_0 + \lambda_2(1 - \pi_0) = \lambda_2 - \mu_2 \left(1 - \frac{\lambda_1}{\mu_1}\right).$$

В случае *абсолютного приоритета без дообслуживания* ВВП на грани $B^{\{2\}}$ равно

$$v^{\{2\}} = (\lambda_2 - \mu_2 - \lambda_1)\pi_0 + \lambda_2(1 - \pi_0) = \lambda_2 - (\mu_2 + \lambda_1) \left(1 - \frac{\lambda_1}{\mu_1}\right).$$

В терминах второго вектоного поля рассматриваемая система эргодична в том и только том случае, когда $v_1^{\{1,2\}} < 0$ и $v^{\{2\}} < 0$. Итак, система эргодична тогда и только тогда, когда

в случае *абсолютного приоритета с дообслуживанием*

$$\frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2} < 1,$$

в случае *абсолютного приоритета без дообслуживания*

$$\frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2 + \lambda_1} < 1.$$

3.2.2 Приоритетная система с конечным и бесконечным источниками

Рассмотрим приоритетную систему с одним обслуживающим узлом (узлом 0 в терминах параграфа 2.1) и двумя типами требований. Предположим, что система замкнута по отношению к требованиям типа 1 и открыта по отношению к требованиям типа 2. Всего существует N требований типа 1, и каждое из них может находится в буфере (в узле 1 в терминах параграфа 2.1) или ожидать обслуживания в обслуживающем узле. Каждое требование независимо от других ждет в буфере экспоненциальное время со средним θ^{-1}, затем покидает буфер и переходит в очередь к обслуживающему узлу. Времена обслуживания требований типа 1 распределены экспоненциально со средним μ_1^{-1}. После завершения обслуживания требования типа 1 возвращаются в буфер. Требования типа 2 поступают в очередь к обслуживающему узлу согласно пуассоновскому входному потоку с интенсивностью λ_2. После обслуживания требования типа 2 покидают систему. Времена обслуживания требований типа 2 распределены экспоненциально со средним μ_2^{-1}.

Мы будем рассматривать два случая.

Случай 1: требования типа 1 имеют абсолютный приоритет с дообслуживанием над требованиями типа 2.

Случай 2: требования типа 2 имеют абсолютный приоритет с дообслуживанием над требованиями типа 1.

Пространство состояний для этой приоритетной системы:

$$\{(n_1, n_2):\ 0 \leqslant n_1 \leqslant N,\ n_2 \geqslant 0\} = \{0, \ldots, N\} \times \mathbf{Z}_+.$$

Здесь n_1 — число требований типа 1 в очереди к узлу обслуживания, n_2 — число требований типа 2, ожидающих обслуживания в очереди к обслуживающему узлу.

Как и выше, мы рассмотрим *ассоциированные* случайные блуждания в полуполосе

$$\{0, \ldots, N\} \times \mathbf{Z}_+.$$

Интенсивности переходов определяются следующим образом.

В *случае 1* мы полагаем

$$\lambda_{\alpha\beta} = \begin{cases} (N - n_1)\theta, & \text{если } \beta - \alpha = (1, 0), \\ \mu_1, & \text{если } \beta - \alpha = (-1, 0), \\ \lambda_2, & \text{если } \beta - \alpha = (0, 1) \end{cases}$$

для $\alpha \in B^{\{1,2\}} \cup B^{\{1\}} = \{(n_1, n_2):\ 0 < n_1 \leqslant N,\ n_2 \geqslant 0\}$ и

$$\lambda_{\alpha\beta} = \begin{cases} N\theta, & \text{если } \beta - \alpha = (1, 0), \\ \mu_2, & \text{если } \beta - \alpha = (0, -1), \\ \lambda_2, & \text{если } \beta - \alpha = (0, 1) \end{cases}$$

для $\alpha \in B^{\{2\}} = \{(0, n_2) : n_2 > 0\}$.

В *случае 2* мы полагаем

$$\lambda_{\alpha\beta} = \begin{cases} (N - n_1)\theta, & \text{если } \beta - \alpha = (1, 0), \\ \lambda_2, & \text{если } \beta - \alpha = (0, 1), \\ \mu_2, & \text{если } \beta - \alpha = (0, -1) \end{cases}$$

для $\alpha \in B^{\{1,2\}} \cup B^{\{2\}} = \{(n_1, n_2) : 0 \leqslant n_1 \leqslant N,\ n_2 > 0\}$ и

$$\lambda_{\alpha\beta} = \begin{cases} (N - n_1)\theta, & \text{если } \beta - \alpha = (1, 0), \\ \mu_1, & \text{если } \beta - \alpha = (-1, 0), \\ \lambda_2, & \text{если } \beta - \alpha = (0, 1) \end{cases}$$

для $\alpha \in B^{\{1\}} = \{(n_1, 0) : n_1 > 0\}$.

Рассмотрим *индуцированную* цепь Маркова $\mathcal{L}^{\{2\}}$ с пространством состояний $\mathbf{Z}_N = \{0, \ldots, N\}$. В *случае 1* все множество $\mathbf{Z}_N$ представляет собой один единственный класс существенных состояний. Таким образом, согласно параграфу 3.1 в [3] цепь Маркова $\mathcal{L}^{\{2\}}$ эргодична тогда и только тогда, когда второе векторное поле на грани $B^{\{2\}}$ отрицательно:

$$v^{\{2\}} = \pi_0(\lambda_2 - \mu_2) + (1 - \pi_0)\lambda_2 = \lambda_2 - \pi_0\mu_2 < 0.$$

Легко проверить, что в этом случае

$$\pi_0 = \frac{(\mu_1/\theta)^N}{N!} \left(1 + \frac{\mu_1}{\theta} + \frac{1}{2!}\left(\frac{\mu_1}{\theta}\right)^2 + \cdots + \frac{1}{N!}\left(\frac{\mu_1}{\theta}\right)^N\right)^{-1}.$$

Итак, в *случае 1* условие эргодичности имеет вид: $\dfrac{\lambda_2}{\mu_2} < \pi_0$.

В *случае 2* цепь Маркова $\mathcal{L}^{\{2\}}$ имеет только одно существенное состояние $\{N\}$. Это означает, что приоритетная система эргодична в том и только том случае, когда вторая компонента $v_2^{\{1,2\}}$ *второго векторного поля* внутри $B^{\{1,2\}}$ отрицательна:

$$v_2^{\{1,2\}} = \lambda_2 - \mu_2 < 0.$$

Таким образом, в этом случае условие эргодичности очень просто

$$\lambda_2 < \mu_2$$

и не зависит от θ, μ_1.

3.2.3 Анализ дисциплины относительный приоритет

Мы рассмотрим здесь систему с двумя типами требований и одним обслуживающим узлом. Наши предположения следующие.

Оба типа 1 и 2 бесконечны, входные потоки пуассоновские, λ_1 и λ_2 — интенсивности этих потоков.

Тип 1 имеет относительный приоритет над типом 2. Это означает, что если в момент, когда поступает требование типа 1, очередь из требований типа 1 пуста, а сервер занят некоторым требованием типа 2, то требование типа 1 ждет, пока завершится обслуживание требования типа 2, а затем занимает обслуживающий узел.

Времена обслуживания обоих типов распределены экспоненциально, μ_1 и μ_2 — соответствующие интенсивности.

Обратим внимание на то, что знание количества требований типов 1 и 2 в системе недостаточно для марковского описания системы. Для того, чтобы иметь марковость, надо иметь дополнительную информацию о том, каков тип текущего обслуживаемого требования. Таким образом, в качестве пространства состояний нашей системы следует выбрать следующее множество:

$$\mathcal{S} = \{\varnothing\} \cup \{(\sigma, n_1, n_2):\ \sigma = 1, 2,\ n_i \in \mathbf{Z}_+\},$$

где σ — тип требования, которое обслуживается в настоящий момент времени, n_1 — число требований типа 1, ожидающих обслуживания, n_2 — число требований типа 2, ожидающих обслуживания; так что общее число требований в системе равно $n_1 + n_2 + 1$. Состояние $\{\varnothing\}$ означает, что в системе нет ни одного требования.

Наша цель — построить второе векторное поле. Рассмотрим индуцированную цепь Маркова $\mathcal{L}^{\{2\}}$ с пространством состояний $\mathcal{S}' = \{(\sigma, n_1):\ \sigma = 1, 2,\ n_1 \in \mathbf{Z}_+\}$ и интенсивностями переходов

$$\lambda^{\{2\}}_{\sigma_1 \sigma_2} = \sum_y \lambda_{(\sigma_1, x), (\sigma_2, y)},$$

где $x > 0$.

Удобно считать множеством состояний цепи Маркова $\mathcal{L}^{\{2\}}$ множество $\mathbf{Z}^1$: состояние $k > 0$ означает, что в системе имеется k требований типа 1 и текущее требование на обслуживающем узле имеет тип 1; состояние $k \leqslant 0$ означает, что в системе находятся $-k$ требований типа 1, а текущее требование на обслуживающем узле имеет тип 2. Легко видеть, что интенсивности переходов имеют следующий вид:

$$\lambda_{k,k+1} = \lambda_1, \quad \lambda_{k,k-1} = \mu_1 \qquad \text{для } k > 0,$$

$$\lambda_{0,-1} = \lambda_1,$$

$$\lambda_{k,k-1} = \lambda_1, \quad \lambda_{k,-k} = \mu_2 \qquad \text{для } k < 0$$

(см. рис. 7).

Легко доказать, что цепь Маркова $\mathcal{L}^{\{2\}}$ эргодична в том и только том случае, когда $\lambda_1 < \mu_1$. Более того, ее стационарное распределение $\{\pi_k,\ k \in \mathbf{Z}^1\}$

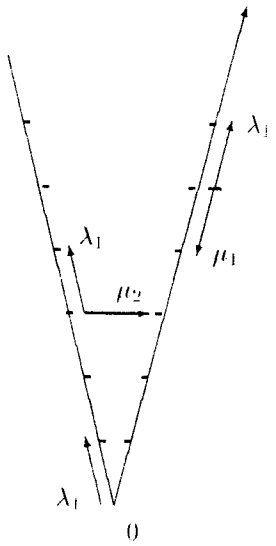

Рис. 7. Индуцированная цепь $\mathcal{L}^{\{2\}}$ в случае относительного приоритета

может быть найдено явно путем решения соответствующих разностных уравнений. Легко проверить, что в случае $\mu_1 \neq \lambda_1 + \mu_2$ мы имеем соотношения

$$\pi_{-k} = \pi_0 \left(\frac{\lambda_1}{\lambda_1 + \mu_2}\right)^k,$$

$$\pi_k = \pi_0 \left[\left(\frac{\lambda_1}{\mu_1}\right)^k + \frac{\lambda_1}{\mu_1 - \lambda_1 - \mu_2}\left(\left(\frac{\lambda_1}{\lambda_1 + \mu_2}\right)^{k-1} - \left(\frac{\lambda_1}{\mu_1}\right)^{k-1}\right)\right], \quad k \geqslant 1.$$

Ситуация, когда $\mu_1 = \lambda_1 + \mu_2$, есть *резонансный* случай для соответствующего разностного уравнения. В этом случае стационарные вероятности удовлетворяют таким соотношениям:

$$\pi_{-k} = \pi_0 \left(\frac{\lambda_1}{\mu_1}\right)^k, \quad \pi_k = \pi_0 k \left(\frac{\lambda_1}{\mu_1}\right)^k, \quad k \geqslant 1.$$

Предположим, что $\lambda_1 < \mu_1$. Тогда второе векторное поле на грани $B^{\{2\}}$ равно

$$v^{\{2\}} = \left(\sum_{k \leqslant 0} \pi_k\right)(\lambda_2 - \mu_2) + \left(\sum_{k > 0} \pi_k\right)\lambda_2.$$

Легко вычислить, что в случае $\mu_1 \neq \lambda_1 + \mu_2$ мы имеем

$$\sum_{k \leqslant 0} \pi_k = \pi_0 \frac{\lambda_1 + \mu_2}{\mu_2}.$$

$$\sum_{k>0}\pi_k=\pi_0\left[\frac{\lambda_1}{\mu_1-\lambda_1}+\frac{\lambda_1}{\mu_1-\lambda_1-\mu_2}\left(\frac{\lambda_1+\mu_2}{\mu_2}-\frac{\lambda_1}{\mu_1-\lambda_1}\right)\right]. \qquad (1)$$

Рассматриваемая система массового обслуживания эргодична тогда и только тогда, когда $v^{(2)}<0$. Итак, условия эргодичности в случае $\mu_1\neq\lambda_1+\mu_2$ принимают следующий вид:

$$\begin{cases}\lambda_1<\mu_1,\\ \dfrac{\lambda_1+\mu_2}{\mu_2}(\lambda_2-\mu_2)+\left[\dfrac{\lambda_1}{\mu_1-\lambda_1}+\dfrac{\lambda_1}{\mu_1-\lambda_1-\mu_2}\left(\dfrac{\lambda_1+\mu_2}{\mu_2}-\dfrac{\lambda_1}{\mu_1-\lambda_1}\right)\right]\lambda_2<0.\end{cases}$$

Если $\mu_1=\lambda_1+\mu_2$, то мы имеем дело с резонансным случаем и равенство (1) теряет смысл. В этом случае

$$\sum_{k>0}\pi_k=\pi_0\sum_{i=1}^{\infty}i\left(\frac{\lambda_1}{\mu_1}\right)^i=\pi_0\frac{\lambda_1}{\mu_1}\cdot\frac{1}{\left(1-\frac{\lambda_1}{\mu_1}\right)^2}.$$

Таким образом, условия эргодичности в случае $\mu_1=\lambda_1+\mu_2$ имеют вид

$$\begin{cases}\lambda_1<\mu_1,\\ \dfrac{\lambda_1+\mu_2}{\mu_2}(\lambda_2-\mu_2)+\dfrac{\lambda_1\mu_1}{(\lambda_1-\mu_1)^2}\lambda_2<0.\end{cases}$$

3.3 Требования от нескольких бесконечных источников на одном облуживающем узле

3.3.1 Абсолютный приоритет с дообслуживанием

Предположим, что существует K различных типов требований, требования типа 1 имеют абсолютный приоритет (с дообслуживанием) над требованиями типов $2,3,\ldots$, требования типа 2 имеют абсолютный приоритет (с дообслуживанием) над требованиями типов $3,4,\ldots$ и так далее. Мы будем предполагать также, что имеет место следующее предположение.

Предположение Г. *Параметры модели λ_i, μ_i, $i=1,\ldots,K$, таковы, что для всех k мы имеем* $\sum_{i=1}^{k}\dfrac{\lambda_i}{\mu_i}\neq 1$.

Это предположение не является слишком ограничительным, так как мы оставляем вне рассмотрения множество параметров нулевой лебеговой меры.

Как и в случае двух типов, мы свяжем с нашей системой случайное блуждание в $\mathbf{Z}_+^K$. Мы докажем, что условие

$$\sum_{k=1}^{K}\frac{\lambda_k}{\mu_k}<1 \qquad (2)$$

необходимо и достаточно для эргодичности, и дадим явное описание соответствующего второго векторного поля.

Предложение 3. (Явная конструкция второго векторного поля.) *Второе векторное поле обладает следующими свойствами.*

1. $\Lambda_1 = \{1, 2, \ldots, K\}$ — *выходящая грань для всех* $(K-1)$*-мерных граней* $\Lambda_1 - \{i\}$, $i \neq 1$, *и входящая для грани* $\Lambda_2 = \{2, 3, \ldots, K\}$, *если* $\dfrac{\lambda_1}{\mu_1} < 1$, *и выходящая для нее, если* $\dfrac{\lambda_1}{\mu_1} > 1$.

2. *Второе векторное поле на* $(K-k+1)$*-мерной грани* $\Lambda_k = \{k, k+1, \ldots, K\}$ *равно*

$$r^{\Lambda_k} = \left(0, \ldots, \lambda_k - \left(1 - \sum_{i=1}^{k-1} \frac{\lambda_i}{\mu_i}\right)\mu_k, \lambda_{k+1}, \ldots, \lambda_K\right).$$

Следовательно, Λ_k — *выходящая грань для всех* $(K-k)$*-мерных граней* $\Lambda_k - \{i\}$, $i = k+1, \ldots, K$. *Грань* Λ_k *является входящей для грани* $\Lambda_{k+1} = \{k+1, \ldots, K\}$, *если* $\displaystyle\sum_{i=1}^{k} \frac{\lambda_i}{\mu_i} < 1$, *и выходящей для нее, если* $\displaystyle\sum_{i=1}^{k} \frac{\lambda_i}{\mu_i} > 1$.

3. *Если грань* Λ_{k+1} *эргодична, обозначим через* $\sigma_{1,\ldots,k}$ *стационарную вероятность начала координат для индуцированной цепи* $\mathcal{L}^{\Lambda_{k+1}}$. *Тогда*

$$\sigma_{1,\ldots,k} = 1 - \sum_{i=1}^{k} \frac{\lambda_i}{\mu_i}.$$

(*Другими словами,* $\sigma_{1,\ldots,k}$ *есть стационарная вероятность события «в системе нет ни одного требования типов* $1, \ldots, k$*» для эргодической цепи Маркова, полученной из полной системы удалением требований типов* $k+1, \ldots, K$.)

Предложение 4. *Условие* (2) *необходимо и достаточно для эргодичности.*

Замечание. Из результатов пункта 3.3.2 следует, что условие (2) *не является* необходимым для эргодичности системы с дисциплиной абсолютный приоритет *без дообслуживания.*

Доказательство. Легко показать, что цепь $\mathcal{L}^{\Lambda_2}$ эргодична при $\dfrac{\lambda_1}{\mu_1} < 1$ и $\sigma_1 = 1 - \dfrac{\lambda_1}{\mu_1}$. Теперь мы можем явно найти *второе векторное поле* на грани

$\Lambda_2 = \{2, 3, \ldots, K\}$. Мы получим, что при $\dfrac{\lambda_1}{\mu_1} + \dfrac{\lambda_2}{\mu_2} < 1$ грань Λ_2 является входящей для грани Λ_3 и «урезанная» система $\mathcal{L}^{\Lambda_3}$, полученная удалением всех требований типов $3, 4, \ldots, K$, эргодична. Рассмотрим *уравнение баланса* для требований типа 2:

$$\lambda_2 - (\sigma_1 - \sigma_{1,2})\,\mu_2 = 0.$$

Отсюда немедленно получим, что

$$\sigma_{1,2} = 1 - \frac{\lambda_1}{\mu_1} - \frac{\lambda_2}{\mu_2}.$$

Теперь легко по индукции завершить доказательство. Уравнение баланса для типа k имеет следующий вид:

$$\lambda_k - (\sigma_{1,\ldots,k-1} - \sigma_{1,\ldots,k})\,\mu_k = 0.$$

Мы опускаем детали.

Определение. Длина интервала времени между моментом, когда узел начинает обслуживать требование типа i, и моментом, когда это требование покидает систему, назовем *фактическим временем обслуживания* для типа i.

В случае абсолютного приоритета с дообслуживанием фактическое время обслуживания для типа i состоит из времени, когда требование обслуживается на узле, и интервалов прерывания.

Предложение 5. *В эргодическом случае для требований типа i среднее время пребывания требования в системе (в стационарном режиме) равно*

$$W_i = \left(1 - \sum_{l=1}^{i-1} \frac{\lambda_l}{\mu_l}\right)^{-1} \frac{1}{\mu_i - \lambda_i},$$

а среднее фактическое время обслуживания равно

$$M_i = \left(1 - \sum_{i=1}^{i-1} \frac{\lambda_i}{\mu_i}\right)^{-1} \frac{1}{\mu_i}.$$

Замечание. Пусть W_i^0 — среднее время ожидания, а M_i^0 — среднее время обслуживания в ситуации, когда тип i имеет наивысший приоритет. Тогда $W_i^0 = 1/(\lambda_i - \mu_i)$ и $M_i^0 = 1/\mu_i$.

Посмотрим на поведение эргодичной системы в течение очень долгого времени $[0, T]$. Из приведенных выше результатов и закона больших чисел следует, что в течение времени порядка $\left(1 - \sum_{i=1}^{i-1} \frac{\lambda_i}{\mu_i}\right) T$ система будет свободна от требований типов $1, 2, \ldots, i-1$ и i-требования, как имеющие максимальный

приоритет, в течении этого периода времени могут обслуживаться. Оставшееся время (порядка $\left(\sum_{i=1}^{i-1}\frac{\lambda_i}{\mu_i}\right)T$) обслуживание i-требований блокировано требованиями с более высоким приоритетом. Это означает, что прохождение i-требований через систему в $\left(1-\sum_{i=1}^{i-1}\frac{\lambda_i}{\mu_i}\right)^{-1}$ раз медленнее, чем могло бы быть в случае, если бы i-требования имели наивысший приоритет. Естественно назвать множитель $c_i=\left(1-\sum_{i=1}^{i-1}\frac{\lambda_i}{\mu_i}\right)^{-1}$ *эффективным временем* для типа i.

3.3.2 Абсолютный приоритет без дообслуживания

В этом пункте мы считаем, что справедливо следующее предположение.

Предположение Д. *Параметры модели λ_i, μ_i, $i=1,\ldots,K$, таковы, что для всех k имеет место* $\sum_{i=1}^{k}\frac{\lambda_i}{\mu_i+\lambda_1+\cdots+\lambda_{i-1}}\neq 1$.

Мы докажем, что в случае абсолютного приоритета без дообслуживания система с K различными бесконечными источниками эргодична тогда и только тогда, когда

$$\frac{\lambda_1}{\mu_1}+\frac{\lambda_2}{\mu_2+\lambda_1}+\frac{\lambda_3}{\mu_3+\lambda_1+\lambda_2}+\cdots+\frac{\lambda_K}{\mu_k+\lambda_1+\cdots+\lambda_{K-1}}<1. \qquad (3)$$

Обозначим

$$S_k=\frac{\lambda_1}{\mu_1}+\frac{\lambda_2}{\mu_2+\lambda_1}+\frac{\lambda_3}{\mu_3+\lambda_1+\lambda_2}+\cdots+\frac{\lambda_k}{\mu_k+\lambda_1+\cdots+\lambda_{k-1}}.$$

Предложение 6. (Явная конструкция второго векторного поля.) *Второе векторное поле обладает следующими свойствами:*

1. $\Lambda_1=\{1,2,\ldots,K\}$ – *выходящая грань для всех* $(K-1)$*-мерных граней* $\Lambda_1-\{i\}$, $i\neq 1$, *и входящая для грани* $\Lambda_2=\{2,3,\ldots,K\}$ *при* $\frac{\lambda_1}{\mu_1}<1$, *но выходящая для нее при* $\frac{\lambda_1}{\mu_1}>1$.

2. *Второе векторное поле на* $(K-k+1)$*-мерной грани* $\Lambda_k=\{k,k+1,\ldots,K\}$ *равно*

$$v^{\Lambda_k}=(0,\ldots,\lambda_k-(1-S_{k-1})\,\mu_k,\lambda_{k+1},\ldots,\lambda_K).$$

Таким образом, грань Λ_k *выходящая для всех* $(K-k)$*-мерных граней* $\Lambda_k-\{i\}$, $i=k+1,\ldots,K$. *Грань* Λ_k *входящая для грани* $\Lambda_{k+1}=\{k+1,\ldots,K\}$ *при* $S_k<1$ *и выходящая для нее при* $S_k>1$.

3. *Если Λ_{k+1} эргодична, обозначим через $\sigma_{1,\dots,k}$ стационарную вероятность начала координат для индуцированной цепи $\mathcal{L}^{\Lambda_{k+1}}$. Тогда*

$$\sigma_{1,\dots,k} = 1 - S_k \equiv 1 - \sum_{i=1}^{k} \frac{\lambda_i}{\mu_i + \lambda_1 + \dots + \lambda_{i-1}}.$$

Предложение 7. *Условие* (3) *необходимо и достаточно для эргодичности.*

Доказательство. Доказательство аналогично доказательству в предыдущем пункте. Единственное отличие состоит в том, что уравнение баланса для типа k имеет вид

$$\lambda_k - (\mu_k + \lambda_1 + \dots + \lambda_{k-1})(\sigma_{1,\dots,k-1} - \sigma_{1,\dots,k}) = 0.$$

Предложение 8. *В эргодичном случае (в стационарном режиме) среднее фактическое время обслуживания равно*

$$M_i = \frac{1}{\mu_i + \lambda_1 + \dots + \lambda_{i-1}};$$

среднее время пребывания в системе требований типа i равно

$$W_i = (1 - S_{i-1})^{-1} \frac{1}{(\mu_i + \lambda_1 + \dots + \lambda_{i-1}) - \lambda_i}.$$

Замечание. Легко видеть, что в эргодичном случае стационарная вероятность того, что заданное требование типа i будет прервано до завершения обслуживания, равна

$$\frac{\lambda_1 + \dots + \lambda_{i-1}}{\mu_i + \lambda_1 + \dots + \lambda_{i-1}}.$$

Знание этих показателей имеет значение для разработчиков компьютерных архитектур, стремящихся выбрать оптимальный порядок приоритетов $P(i)$.

3.3.3 Жидкостная аппроксимация для систем с абсолютным приоритетом

Абсолютный приоритет с дообслуживанием

Мы находимся в ситуации пункта 3.3.1 и считаем, что имеет место предположение Г.

Пусть n таково, что

$$\sum_{i=1}^{n-1} \frac{\lambda_i}{\mu_i} < 1, \quad \text{но} \quad \sum_{i=1}^{n} \frac{\lambda_i}{\mu_i} > 1.$$

По предложению 3 эргодичными являются только следующие грани: $\Lambda_1, \ldots, \Lambda_n$. Мы напомним, что второе векторное поле определено только на эргодичных гранях (см. [3]).

Наша очередная цель — построить детерминированную динамическую систему в $\mathbf{R}_+^K$, которая связана с исходной случайной системой в $\mathbf{Z}_+^K$.

Введем в рассмотрение следующие грани в $\mathbf{R}_+^K$:

$$\hat{\Lambda}_k = \left\{x = (x_1, \ldots, x_K) \in \mathbf{R}_+^K\colon\ x_1 = \cdots = x_{k-1} = 0,\ x_k > 0, \ldots,\ x_K > 0\right\}.$$

Рассмотрим следующее векторное поле $v(x)$ в $\mathbf{R}_+^K$:

$$\begin{gathered} v(x) = v^{\Lambda_1},\ \text{если}\ \ x \in \hat{\Lambda}_1 \setminus \hat{\Lambda}_2, \\ \ldots \\ v(x) = v^{\Lambda_{n-1}},\ \text{если}\ \ x \in \hat{\Lambda}_{n-1} \setminus \hat{\Lambda}_n, \\ v(x) = v^{\Lambda_n},\ \text{если}\ \ x \in \hat{\Lambda}_n. \end{gathered}$$

В случае, когда $n = K + 1$, мы положим $v(0) = 0$. Это векторное поле определяет динамическую систему U_t, $t \geqslant 0$, в $\mathbf{R}_+^K$:

$$\frac{d}{dt} U_t x = v\left(U_t x\right), \quad U_0 x \equiv x.$$

Эта система является корректно определенной в любой точке $x \in \mathbf{R}_+^K$.

Пусть $x \in \mathbf{R}^K$. Предположим, что в начальный момент задана конфигурация $[xN] \stackrel{\text{def}}{=} ([x_1 N], \ldots, [x_K N])$, то есть в момент времени $s = 0$ мы имеем на узле обслуживания очередь из $[x_1 N]$ требований типа $1, \ldots,$ очередь из $[x_K N]$ требований типа K. Обозначим через $X_s\left([xN]\right)$ состояние системы в момент s.

Предложение 9 (эйлеровский предел). *Для любой точки $x \in \mathbf{R}_+^K$ существует следующий (детерминированный) предел*

$$\frac{1}{N} X_{[tN]}(xN) \to U_t x \quad (N \to \infty)\ (\textit{почти всюду}).$$

Из этого предложения следует смысл введенной выше динамической системы. Динамическая система U_t описывает эволюцию больших очередей: если в момент 0 приоритетная система находится в состоянии $[xN]$, то в момент времени tN состояние системы будет «приблизительно» равно $[(U_t x)\, N]$. Слово «приблизительно» означает, что уклонение реального состояния в момент tN от значения $[(U_t x)\, N]$ по порядку меньше, чем N.

Возможны два различных случая:

$(n = K + 1)$. Это соответствует ситуации, когда все грани $\Lambda_1, \ldots, \Lambda_K$ и $\Lambda_{K+1} = \{0\}$ эргодичны и, следовательно, вся приоритетная система эргодична. В этом случае для любой начальной точки $x \in \mathbf{R}_+^K$ траектория

$U_t x$ достигает начала координат за конечное время (см. рис. 8a). Обозначим через $\tau(x)$ время достижения нуля траекторией динамической системы, выходящей из x:

$$U_t x \neq 0 \text{ для } t < \tau(x) \quad \text{и} \quad U_t x = 0 \text{ для } t \geqslant \tau(x).$$

($n \leqslant K$). В этой ситуации грани $\Lambda_1, \ldots, \Lambda_n$ эргодичны, но все компоненты второго векторного поля на грани Λ_n строго положительны, следовательно, полная приоритетная система неэргодична. В этом случае любая траектория $U_t x$ достигает грани $\tilde{\Lambda}_n$ за конечное время $\tau_n(x)$ и остается в этой грани навсегда (не уходя в некоторую подгрань меньшей размерности). В терминах приоритетной системы это означает, что после времени $\tau_n(x)N$ длины очередей из требований типов $1, 2, \ldots, n-1$ будут по порядку меньше, чем N, а очереди из требований типов $n, n+1, \ldots, K$ будут расти пропорционально N (см. рис. 8b).

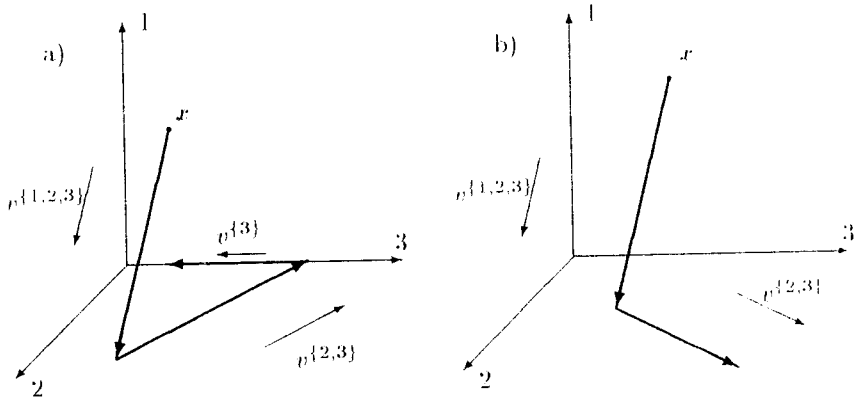

Рис. 8. Динамическая система U_t в эргодическом (a) и неэргодическом (b) случаях

Теперь мы вычислим время $\tau(x)$ достижения начала координат в эргодичном случае.

Лемма 10. *Для эргодичной системы время $\tau(x)$ достижения нуля динамической системой U_t линейно по начальной точке x. А именно,*

$$\tau(x) = (b, x) \equiv b_1 x_1 + \cdots + b_K x_K,$$

где вектор $b = (b_1, \ldots, b_K)$ имеет следующий вид

$$\begin{aligned} b_1 &= e_2(1+e_3)(1+e_4)\ldots(1+e_{K+1})/\mu_1, \\ b_2 &= e_3(1+e_4)\ldots(1+e_{K+1})/\mu_2, \\ &\ldots \\ b_{K-1} &= e_K(1+e_{K+1})/\mu_{K-1}, \\ b_K &= e_{K+1}/\mu_K. \end{aligned} \tag{4}$$

где мы обозначили

$$c_k = \left(1 - \sum_{i=1}^{k-1} \frac{\lambda_i}{\mu_i}\right)^{-1}.$$

Замечание. Заметим, что величины c_k есть в точности то, что мы назвали в пункте 3.3.1 эффективным временем для типа k.

Доказательство леммы. Пусть $x \in \hat{\Lambda}_1 \setminus \hat{\Lambda}_2$. Легко проверить, что траектория $U_t x$ достигает грани $\hat{\Lambda}_2$ за время

$$\tau_2(x) = \frac{x_1}{\mu_1 - \lambda_1} \equiv \frac{c_2}{\mu_1} x_1.$$

Предположим, что траектория достигает грань $\hat{\Lambda}_{k-1}$ за время $\tau_{k-1}(x)$. Тогда она попадет на грань $\hat{\Lambda}_k$ в момент

$$\tau_k(x) = \frac{c_k}{\mu_{k-1}} x_{k-1} + (1 + c_k)\, \tau_{k-1}(x).$$

Заметим, что $\tau(x) \equiv \tau_{K+1}(x)$. Теперь легко завершить доказательство. □

Предложенный выше анализ приводит к следующей *оптимизационной задаче*. Мы ограничимся эргодичным случаем. Представим себе, что разработчик компьютерных архитектур имеет дело с бесконечными источниками, характеристики которых λ_i, μ_i, $i = \overline{1, K}$, заданы. Мы хотим выбрать такой порядок приоритетов $i_1, i_2, \ldots, i_K$, чтобы соответствующая приоритетная система, начинающая из некоторого состояния с «длинными очередями» $\{x, N\}$ достигла равновесия за минимально возможное время. Другими словами мы ищем такую перестановку $\chi\colon (1, \ldots, K) \to (i_1, \ldots, i_K)$, чтобы обеспечить наименьшее время $\tau_\chi(x)$ достижения начала из точки x. Легко видеть, что при заданном x мы должны выбрать такой порядок приоритетов χ, для которого вектор b_χ имеет минимальную проекцию на направление x.

Тем самым, мы приходим к следующему определению. Зафиксируем некоторую конечную меру ψ на следующей $(K-1)$-мерной поверхности:

$$\Delta_K = \{x \in \mathbf{R}_+^K\colon\ x_1^2 + \cdots + x_K^2 = 1,\ x_1 \geqslant 0, \ldots, x_K \geqslant 0\}.$$

Эта мера выбирается разработчиком компьютерных архитектур сообразно некоторым внешним соображениям, которые мы не обсуждаем здесь. Если у нас нет причин отдавать предпочтение каким-либо направлениям $\gamma \in \Delta_K$, нам следует выбрать равномерное распределение на Δ_K в качестве меры ψ.

Определение. Порядок приоритетов χ называется *оптимальным* для заданной меры ψ, если

$$\tau_\chi(\psi) \stackrel{\text{def}}{=} \int_{\Delta_K} \tau_\chi(\gamma)\, d\psi(\gamma) = \min_\sigma\ \tau_\sigma(\psi),$$

где минимум берется по всем перестановкам $\sigma\colon (1, \ldots, K) \to (i_1, \ldots, i_K)$.

Предложение 11. *Если в качестве меры ψ взято равномерное распределение на Δ_K, то порядок приоритетов χ оптимальный в том и только том случае, когда вектор b_χ имеет наименьшую проекцию на направление* $(1, 1, \ldots, 1)$.

Абсолютный приоритет без дообслуживания

Сейчас мы в ситуации пункта 3.3.2 и считаем, что предположение $\mathcal{A}$ имеет место. Жидкостный анализ в этом случае напоминает анализ в случае абсолютного приоритета с дообслуживанием. Все различия связаны с различной формой второго векторного поля в обоих случаях.

Прежде всего мы зафиксируем целое число n, такое что

$$\sum_{i=1}^{n-1} \frac{\lambda_i}{\mu_i + \lambda_1 + \cdots + \lambda_{i-1}} < 1, \quad \text{но} \quad \sum_{i=1}^{n} \frac{\lambda_i}{\mu_i + \lambda_1 + \cdots + \lambda_{i-1}} > 1.$$

Из предложения 6 следует, что эргодичными являются только следующие грани: $\Lambda_1, \ldots, \Lambda_n$. На этих гранях определено второе векторное поле. Так же, как и в случае абсолютного приоритета с дообслуживанием, мы определим векторное поле $v(x)$ на $\mathbf{R}_+^K$, которое, в свою очередь, задает динамическую систему U_t, $t \geqslant 0$, на $\mathbf{R}_+^K$:

$$\frac{d}{dt} U_t x = v(U_t x), \quad U_0 x \equiv x.$$

Обозначим через $X_s([xN])$ состояние системы в момент s при условии, что мы вышли в момент $s = 0$ из конфигурации очередей $[xN] = ([x_1 N], \ldots, [x_K N])$.

Предложение 12 (эйлеровский предел). *Для любого $x \in \mathbf{R}_+^K$ существует следующий детерминированный предел:*

$$\frac{1}{N} X_{[tN]}(xN) \to U_t x \quad (N \to \infty) \ (\textit{почти всюду}).$$

Весь последующий анализ абсолютно аналогичен случаю абсолютного приоритета с дообслуживанием.

Единственное, что мы здесь отметим, это выражение для подсчета времени $\tau(x)$ достижения нуля траекториями $U_t x$, $t \geqslant 0$, в эргодическом случае.

Лемма 13. *Для эргодичной системы время $\tau(x)$ достижения начала динамической системой U_t равно*

$$\tau(x) = (b, x),$$

где вектор $b = (b_1, \ldots, b_K)$ имеет вид (4), *где величины e_k определены следующим образом:*

$$e_k = \left(1 - \sum_{i=1}^{k-1} \frac{\lambda_i}{\mu_i + \lambda_1 + \cdots + \lambda_{i-1}}\right)^{-1}.$$

3.3.4 Относительный приоритет

Как мы отмечали выше этот случай является более деликатным из-за более сложного пространства состояний. Пространство состояний в случае K бесконечных источников с относительным приоритетом может быть выбрано в следующем виде:

$$\mathcal{S} = \{\varnothing\} \cup \{1, \ldots, K\} \times \mathbf{Z}_+^K.$$

Здесь состояние $\{\varnothing\}$ означает, что в системе нет ни одного требования, состояние $\{i\} \times (n_1, \ldots, n_K)$ означает, что есть очередь из n_1 требований типа $1, \ldots,$ очередь из n_K требований типа K, а узел занят требованием типа i, таким образом, общее число требований в системе равно $n_1 + \cdots + n_K + 1$. Легко выписать соответствующие интенсивности переходов. Этот случай также может быть проанализирован методом вложенных цепей (см. пример пункта 3.2.3).

3.4 Необходимое условие эргодичности для специального класса приоритетных дисциплин

Мы ограничимся системой бесконечных источников типов $1, \ldots, K$. Мы предполагаем, что дисциплина такова, что имеет место следующее.

Условие 1. Время обслуживания требования типа k, поступающего на узел обслуживания, определяется до начала обслуживания и распределено экспоненциально со средним μ_k^{-1}.

Условие 2. Требование не может покинуть систему до того, как истечет время, предписанное ему для обслуживания.

Замечание. Для дисциплин абсолютный приоритет с дообслуживанием и относительный приоритет условия 1 и 2 выполнены. Для абсолютного приоритета без дообслуживания условие 2 не имеет места.

Предложение 14. (Необходимое условие эргодичности.) *Предположим, что условия 1 и 2 выполнены и система эргодична. Тогда имеет выполнено следующее:*

$$\sum_{k=1}^{K} \frac{\lambda_k}{\mu_k} < 1.$$

Замечание. Условие 2 в предложении существенно. Чтобы увидеть это, следует рассмотреть дисциплину абсолютный приоритет без дообслуживания в случае

$$K = 2, \ \lambda_1 = \lambda_2 = 1, \ \mu_1 = \mu_2 = 2.$$

Эта система эргодична, но $\dfrac{\lambda_1}{\mu_1} + \dfrac{\lambda_1}{\mu_2} = 1$.

Доказательство предложения. Пусть a — тип требования, обслуживаемый на узле в заданный момент времени. Событие $\{a = \varnothing\}$ означает, что

в системе нет требований. Предположим, что система эргодична и $\mathsf{P}_{\mathrm{st}}\{\cdot\}$ — ее стационарное распределение. Очевидно, что

$$\mathsf{P}_{\mathrm{st}}\{a=\varnothing\}+\mathsf{P}_{\mathrm{st}}\{a=1\}+\cdots+\mathsf{P}_{\mathrm{st}}\{a=K\}=1.$$

Уравнение баланса для типа k имеет следующую форму

$$\lambda_k-\mu_k\mathsf{P}_{\mathrm{st}}\{a=k\}=0.$$

Итак, мы получили

$$\sum_{k=1}^{K}\frac{\lambda_k}{\mu_k}=1-\mathsf{P}_{\mathrm{st}}\{a=\varnothing\},$$

и предложение доказано.

3.5 Система с несколькими бесконечными и конечными источниками

3.5.1 Абсолютный приоритет с дообслуживанием

Мы рассмотрим систему с требованиями следующих типов

$$\underbrace{1,\ldots,K_1}_{I_1},\underbrace{K_1+1,\ldots,K_1+L_1}_{F_1},K_1+L_1+1,\ldots,\sum_{l=1}^{r}(K_l+L_l). \tag{5}$$

Требования типа i имеют абсолютный приоритет с дообслуживанием над требованиями типа j, если $i<j$. Мы предполагаем, что типы

$$\underbrace{1,\ldots,K_1}_{I_1},\underbrace{K_1+L_1+1,\ldots,K_1+L_1+K_2}_{I_2},\ldots$$

соответствуют бесконечным источникам требований, а типы

$$\underbrace{K_1+1,\ldots,K_1+L_1}_{F_1},\ldots,\underbrace{\sum_{l=1}^{r-1}(K_l+L_l)+K_r+1,\sum_{l=1}^{r}(K_l+L_l)}_{F_r}$$

соответствуют конечным источникам требований. Мы предполагаем также, что для бесконечного типа $k=\sum\limits_{i=1}^{l-1}(K_i+L_i)+K_l+v\in I_l$ интенсивность входного потока равна $\lambda_v^{(l)}$, а время обслуживания экспоненциально со средним $\left(\mu_l^{(l)}\right)^{-1}$.

Рассмотрим каждую группу F_l конечных типов как отдельную приоритетную систему. Это неприводимая конечная цепь Маркова, которая всегда

эргодична. Обозначим через $\pi_0^{(l)}$ стационарную вероятность события «на узле нет требований» (все требования в буфере).

Предположим, что полная система эргодична. Наша цель — получить необходимые условия эргодичности. Обозначим через $\Pi^{(l)}$ стационарную вероятность события «в системе нет требований из групп $I_1, I_2, \ldots, I_l$». Определим «полную нагрузку» $R^{(l)}$ группы I_l как

$$R^{(l)} = \sum_{i=1}^{K_l} \frac{\lambda_i^{(l)}}{\mu_i^{(l)}}.$$

Имеют место следующие соотношения:

$$\Pi^{(l)} = \Pi^{(l-1)}\pi_0^{(l-1)} - R^{(l)}, \ l \geqslant 2, \quad \Pi^{(1)} = 1 - R^{(1)}. \tag{6}$$

Из предположения эргодичности следует, что все вероятности (6) положительны. Это приводит к следующим *необходимым условиям эргодичности*:

$$\begin{cases} R^{(1)} < 1, \\ R^{(l)} < \Pi^{(l-1)}\pi_0^{(l-1)}, \ l = 2, 3, \ldots, r. \end{cases} \tag{7}$$

Эти условия оказываются также достаточными для эргодичности.

Предложение 15. *Составная система из различных конечных и бесконечных источников с дисциплиной обслуживания абсолютный приоритет с дообслуживанием эргодична тогда и только тогда, когда имеют место условия* (7).

Наша цель — выписать выражения для среднего фактического времени обслуживания и для среднего времени пребывания в системе для требования бесконечного типа k в эргодичном случае. Предположим, что система находится в стационарном режиме. Для каждого $k \leqslant K_{l+1}$ рассмотрим

$$\sigma_{I^1,F^1,\ldots,I^l,F^l,1,\ldots,k} \stackrel{\text{def}}{=}$$

$$\stackrel{\text{def}}{=} \mathsf{P}\left\{\begin{array}{c} \text{в системе нет требований} \\ \text{с типами, принадлежащими группам } I^1, F^1, \ldots, I^l, F^l, \\ \text{и с первыми } k \text{ типами из группы } I^{l+1} \end{array}\right\}.$$

Лемма 16.

$$\sigma_{I^1,F^1,\ldots,I^l,F^l,1,\ldots,k} = \Pi^{(l)}\pi_0^{(l)} - \sum_{i=1}^{k} \frac{\lambda_i^{(l+1)}}{\mu_i^{(l+1)}}.$$

Предложение 17. *Среднее время пребывания для бесконечного источника типа k (см. обозначения выше) равно*

$$W_k^{(l)} = \left(\sigma_{I^1,F^1,\ldots,I^l,F^l,1,\ldots,k}\right)^{-1} \frac{1}{\lambda_k^{(l+1)} - \mu_k^{(l+1)}}.$$

а среднее фактическое время обслуживания равно

$$M_k^{(l)} = \left(\sigma_{l,r^1,\ldots,l^{r^l},1,\ldots,k}\right)^{-1} \frac{1}{\mu_k^{(l+1)}}.$$

Аналогичные формулы можно написать для требований конечного типа.

3.5.2 Абсолютный приоритет без дообслуживания для бесконечных источников

Рассмотрим систему с требованиями различных типов (5) со следующими дисциплинами: если хотя бы один из типов i или j соответствует *конечному* источнику, то дисциплина обслуживания между i и j есть абсолютный приоритет *с дообслуживанием*; если оба типа соответствуют *бесконечным* источникам, то дисциплина между i и j есть абсолютный приоритет *без дообслуживания*. Как обычно, i имеет приоритет над j, если $i < j$.

В этом случае *условия эргодичности* также имеют вид (7), в котором $\pi_0^{(l)}$ и $\Pi^{(l)}$ определены так же, как и раньше, но «нагрузки» $R^{(l)}$ определяются следующим образом

$$R^{(l)} = \frac{\lambda_1^{(l)}}{\mu_1^{(l)} + \sum\limits_{i=1}^{l-1}\left(\lambda_1^{(i)} + \cdots + \lambda_{K_i}^{(i)}\right)} + \frac{\lambda_2^{(l)}}{\mu_2^{(l)} + \lambda_1^{(l)} + \sum\limits_{i=1}^{l-1}\left(\lambda_1^{(i)} + \cdots + \lambda_{K_i}^{(i)}\right)} + \cdots$$
$$\cdots + \frac{\lambda_{K_l}^{(l)}}{\mu_{K_l}^{(l)} + \lambda_1^{(l)} + \cdots + \lambda_{K_l-1}^{(l)} + \sum\limits_{i=1}^{l-1}\left(\lambda_1^{(i)} + \cdots + \lambda_{K_i}^{(i)}\right)}. \tag{8}$$

Предложение 18. *Приоритетная система эргодична тогда и только тогда, когда выполнены условия* (7), *где* $\Pi^{(l)}$ *определяются по формуле* (6), *а* $R^{(l)}$ *определяются по формуле* (8).

Наша следующая цель — выписать выражения для среднего фактического времени обслуживания и для среднего времени ожидания. Предположим, что система эргодична и находится в стационарном режиме.

Лемма 19.

$$\sigma_{l,r^1,\ldots,l^{r^l},1,\ldots,k} = \Pi^{(l)}\pi_0^{(l)} -$$
$$- \sum_{i=1}^{k} \frac{\lambda_i^{(l+1)}}{\mu_i^{(l+1)} + \lambda_1^{(l+1)} + \cdots + \lambda_{i-1}^{(l+1)} + \sum\limits_{i=1}^{l}\left(\lambda_1^{(i)} + \cdots + \lambda_{K_i}^{(i)}\right)}.$$

Среднее фактическое время обслуживания для бесконечного типа, который имеет k-й порядковый номер в группе I_{l+1}, равно

$$M_k^{(l)} = \left(\mu_k^{(l+1)} + \lambda_1^{(l+1)} + \cdots + \lambda_{i-1}^{(l+1)} + \sum_{i=1}^{l}\left(\lambda_1^{(i)} + \cdots + \lambda_{K_i}^{(i)}\right)\right)^{-1}$$

и среднее время пребывания в системе равно

$$W_k^{(l)} = \left(\sigma_{I^1,F^1,\ldots,I^l,F^l,1,\ldots,k}\right)^{-1} \frac{1}{1/M_k^{(l)} - \lambda_k^{(l+1)}}.$$

Аналогичные формулы можно выписать для требований конечного типа.

3.5.3 Второе векторное поле и жидкостная аппроксимация для составной системы

Мы рассмотрим составную модель пункта 3.5.1, состоящую из многих бесконечных и конечных источников (5) с дисциплиной абсолютный приоритет с дообслуживанием. Обозначим эту приоритетную систему через $\mathcal{L}$. Состояниями системы $\mathcal{L}$ служат строки следующего вида

$$\left(n_1, \ldots, n_{K_1}, m_{K_1+1}, \ldots, m_{K_1+L_1}, \ldots, m_{\sum_{i=1}^{r-1}(K_i+L_i)+K_r+1}, \ldots, m_{\sum_{i=1}^{r}(K_i+L_i)}\right),$$

где m_i — число требований конечного типа i, находящихся в очереди к узлу обслуживания, n_j — длина очереди из требований бесконечного типа j, $m_i \in \mathbf{Z}_{N_i} \equiv \{0, \ldots, N_i\}$, $n_j \in \mathbf{Z}_+$. Таким образом, пространством состояний системы является множество

$$\mathcal{S} = \mathbf{Z}_+^{K_1} \times \underset{i=K_1+1}{\overset{K_1+L_1}{\times}} \mathbf{Z}_{N_i} \times \cdots \times \mathbf{Z}_+^{K_r} \times \underset{i=\sum_{l=1}^{r-1}(K_l+L_l)+K_r+1}{\overset{\sum_{l=1}^{r}(K_l+L_l)}{\times}} \mathbf{Z}_{N_i}.$$

Индуцированное случайное блуждание в $\mathbf{Z}_+^{K_1+\cdots+K_r}$

В этом пункте мы имеем дело с *индуцированной* цепью Маркова $\mathcal{L}'$ с пространством состояний

$$\mathcal{S}' = \mathbf{Z}_+^{K_1+\cdots+K_r}, \tag{9}$$

которую мы определим ниже.

Пусть $\pi_0^{(l)}$ — стационарная вероятность 0 для конечной цепи Маркова, соответствующей группе конечных источников F_l (см. стр. 291). Рассмотрим вначале вспомогательную цепь Маркова $\mathcal{L}^{\text{aux}}$ с пространством состояний (9)

и интенсивностями переходов $\lambda^{\mathrm{aux}}_{\alpha\beta}$, $\alpha, \beta \in \mathcal{S}'$. Эта цепь соответствует приоритетной системе, состоящей из бесконечных источников

$$\underbrace{1, \ldots, K_1}_{I_1}, \underbrace{K_1 + L_1 + 1, \ldots, K_1 + L_1 + K_2}_{I_2}, \ldots$$

с дисциплиной абсолютный приоритет с дообслуживанием между ними. Заметим, что если $\lambda^{\mathrm{aux}}_{\alpha\beta} > 0$, то лишь одна компонента вектора $\beta - \alpha$ может быть отрицательной (и автоматически равной -1).

Определим теперь переходные вероятности $\lambda'_{\alpha\beta}$ для индуцированной цепи Маркова $\mathcal{L}'$. Для $\alpha \neq \beta$ положим

$$\lambda'_{\alpha\beta} = \lambda^{\mathrm{aux}}_{\alpha\beta}, \quad \text{если} \quad \beta_j - \alpha_j \geqslant 0 \text{ для всех } j,$$

$$\lambda'_{\alpha\beta} = \left(\pi_0^{(1)} \pi_0^{(2)} \ldots \pi_0^{(k)}\right) \lambda^{\mathrm{aux}}_{\alpha\beta}, \quad \text{если} \quad \beta_j - \alpha_j = -1 \text{ для некоторого } j, \quad \sum_{l=1}^{k} (K_l + L_l) < j \leqslant \sum_{l=1}^{k} (K_l + L_l) + K_{k+1}.$$

Оказывается, что качественное поведение исходной цепи Маркова $\mathcal{L}$ тесно связано с качественным поведением индуцированной цепи Маркова $\mathcal{L}'$.

Следующее утверждение объясняет полезность индуцированной цепи $\mathcal{L}'$.

Предложение 20. *Цепь Маркова $\mathcal{L}$ эргодична тогда и только тогда, когда цепь Маркова $\mathcal{L}'$ эргодична.*

Напомним, что цепь Маркова $\mathcal{L}$ соответствует *составной* системе из нескольких *бесконечных* и *конечных* источников, в то время как цепь $\mathcal{L}'$ соответствует приоритетной системе с несколькими *бесконечными* источниками, которая была изучена в пунктах 3.3.1 и 3.3.3. Таким образом, изучение исходной приоритетной системы $\mathcal{L}$ можно свести к анализу цепи Маркова $\mathcal{L}'$, который может быть дан в духе пунктов 3.3.1 и 3.3.3. В частности, поведение на больших временах составной системы $\mathcal{L}$ с длинными очередями из требований бесконечных типов определяется свойствами второго векторного поля для цепи Маркова $\mathcal{L}'$.

Второе векторное поле

В этом пункте мы рассматриваем индуцированную цепь Маркова $\mathcal{L}'$, состояниями которой являются целочисленные векторы

$$(\underbrace{n_1, \ldots, n_{K_1}}_{I_1}, \underbrace{n_{K_1+1}, \ldots, n_{K_1+K_2}}_{I_2}, \ldots) \in \mathbf{Z}_+^{K_1+\ldots+K_r}.$$

Рассмотрим следующие грани в $\mathbf{Z}_+^{K_1+\ldots+K_r}$

$$\Lambda_{l,k} = \left\{ \sum_{i=1}^{l} K_i + k, \sum_{i=1}^{l} K_i + k + 1, \ldots, \sum_{i=1}^{r} K_i \right\}.$$

Заметим, что $\Lambda_{l,K_l+1} \equiv \Lambda_{l+1,1}$.

В следующем предложении мы даем полное описание второго векторного поля для цепи $\mathcal{L}'$.

Предложение 21. (Явный вид ВВП для индуцированной цепи.)

1. *Грань $\Lambda_{0,1}$ всегда эргодична (по определению). Второе векторное поле на этой грани равно*

$$v^{\Lambda_{0,1}} = \left(\lambda_1^{(1)} - \mu_1^{(1)}, \lambda_2^{(1)}, \ldots, \lambda_{K_1}^{(1)}, \lambda_1^{(2)}, \ldots, \lambda_{K_r}^{(r)}\right).$$

Таким образом, если $\lambda_1^{(1)} > \mu_1^{(1)}$, то грань $\Lambda_{0,1}$ выходящая для любой грани коразмерности 1; если $\lambda_1^{(1)} < \mu_1^{(1)}$, то грань $\Lambda_{0,1}$ входящая для $\Lambda_{0,2}$ и выходящая для любой другой грани коразмерности 1; следовательно, грань $\Lambda_{0,2}$ эргодична, если $\lambda_1^{(1)} < \mu_1^{(1)}$.

2. *В том случае, когда грань $\Lambda_{0,2}$ эргодична, стационарная вероятность начала для эргодичной цепи $\mathcal{L}^{\Lambda_{0,2}}$ равна $\sigma_1 = 1 - \dfrac{\lambda_1^{(1)}}{\mu_1^{(1)}}$.*

3. *Если грань $\Lambda_{0,k}$ эргодична, то ВВП на $\Lambda_{0,k}$ имеет вид*

$$v^{\Lambda_{0,k}} = \left(0, \ldots, \lambda_k^{(1)} - \mu_k^{(1)}\sigma_{1,\ldots,k-1}, \lambda_{k+1}^{(1)}, \ldots, \lambda_{K_r}^{(r)}\right).$$

Таким образом, если $\lambda_k^{(1)}/\mu_k^{(1)} > \sigma_{1,\ldots,k-1}$ $\left(\Longleftrightarrow \sum_{i=1}^{k}(\lambda_i^{(1)}/\mu_i^{(1)}) > 1\right)$, то грань $\Lambda_{0,k}$ выходящая для всех подграней коразмерности 1; если $\sum_{i=1}^{k}(\lambda_i^{(1)}/\mu_i^{(1)}) < 1$, то грань $\Lambda_{0,k}$ входящая для $\Lambda_{0,k+1}$ и выходящая для любой другой подграни коразмерности 1; следовательно, в последнем случае грань $\Lambda_{0,k+1}$ эргодична.

4. *В случае, когда $\Lambda_{0,k}$ входящая для $\Lambda_{0,k+1}$, стационарная вероятность нуля в эргодичной цепи Маркова $\mathcal{L}^{\Lambda_{0,k+1}}$ равна*

$$\sigma_{1,\ldots,k} = 1 - \sum_{i=1}^{k} \frac{\lambda_i^{(1)}}{\mu_i^{(1)}}.$$

5. *Предположим, что грань $\Lambda_{l,1}$ эргодична. Тогда ВВП на $\Lambda_{l,1}$ равно*

$$v^{\Lambda_{l,1}} = \left(0, \ldots, \lambda_1^{(l+1)} - \mu_1^{(l+1)}\sigma_{I_1,\ldots,I_l}\pi_0^{(l)}, \lambda_2^{(l+1)}, \ldots, \lambda_{K_r}^{(r)}\right)$$

(здесь $\sigma_{I_1,\ldots,I_l} \stackrel{\text{def}}{=} \sigma_{1,2,\ldots,\sum_{i=1}^{l} K_i}$).

Таким образом, если $\lambda_1^{(l+1)}/\mu_1^{(l+1)} > \sigma_{I_1,\dots,I_l}\pi_0^{(l)}$, *то грань* $\Lambda_{l,1}$ *выходящая для всех подграней коразмерности* 1; *если* $\lambda_1^{(l+1)}/\mu_1^{(l+1)} < \sigma_{I_1,\dots,I_l}\pi_0^{(l)}$, *то грань* $\Lambda_{l,1}$ *входящая для грани* $\Lambda_{l,2}$ *и выходящая для любой другой подграни коразмерности* 1; *следовательно, в случае* $\lambda_1^{(l+1)}/\mu_1^{(l+1)} < \sigma_{I_1,\dots,I_l}\pi_0^{(l)}$ *грань* $\Lambda_{l,2}$ *эргодична.*

6. *В случае, когда* $\Lambda_{l,1}$ *входящая для* $\Lambda_{l,2}$, *стационарная вероятность нуля для эргодичной цепи* $\mathcal{L}^{\Lambda_{l,2}}$ *равна*

$$\sigma_{I_1,\dots,I_l,1} = \sigma_{I_1,\dots,I_l}\pi_0^{(l)} - \frac{\lambda_1^{(l+1)}}{\mu_1^{(l+1)}}.$$

7. *Если грань* $\Lambda_{l,k}$ *эргодична, то ВВП на* $\Lambda_{l,k}$ *равно*

$$v^{\Lambda_{l,k}} = \left(0,\dots,\lambda_k^{(l+1)} - \mu_k^{(l+1)}\sigma_{I_1,\dots,I_l,1,\dots,k-1},\lambda_{k+1}^{(l+1)},\dots,\lambda_K^{(r)}\right)$$

(*здесь* $\sigma_{I_1,\dots,I_l,1,\dots,k-1} \stackrel{\text{def}}{=} \sigma_{1,2,\dots,\sum_{i=1}^{l}K_i,\dots,\sum_{i=1}^{l}K_i+k-1}$).

Таким образом, если $\lambda_k^{(l+1)}/\mu_k^{(l+1)} > \sigma_{I_1,\dots,I_l,1,\dots,k-1}$, *то грань* $\Lambda_{l,k}$ *выходящая для всех подграней коразмерности* 1; *если* $\lambda_k^{(l+1)}/\mu_k^{(l+1)} < \sigma_{I_1,\dots,I_l,1,\dots,k-1}$, *то грань* $\Lambda_{l,k}$ *входящая для* $\Lambda_{l,k+1}$ *и выходящая для любой другой подграни коразмерности* 1; *следовательно, в последнем случае грань* $\Lambda_{l,k+1}$ *эргодична.*

8. *В случае, когда* $\Lambda_{l,k}$ *входящая для* $\Lambda_{l,k+1}$, *стационарная вероятность нуля в эргодичной цепи* $\mathcal{L}^{\Lambda_{l,k+1}}$ *равна*

$$\sigma_{I_1,\dots,I_l,1,\dots,k} = \sigma_{I_1,\dots,I_l,1,\dots,k-1} - \frac{\lambda_k^{(l+1)}}{\mu_k^{(l+1)}} \equiv \sigma_{I_1,\dots,I_l}\pi_0^{(l)} - \sum_{i=1}^{k}\frac{\lambda_i^{(l+1)}}{\mu_i^{(l+1)}}.$$

Жидкостная аппроксимация

Рассмотрим *детерминированную* динамическую систему U_t, $t \in \mathbf{R}_+$, на $\mathbf{R}_+^{K_1+\dots+K_r}$, порождаемую вторым векторным полем цепи Маркова $\mathcal{L}'$ (см. пункт 3.3.3).

Пусть $Y_s = (y_1(s), y_2(s), \dots) \in \mathcal{S}$ есть состояние цепи $\mathcal{L}$ в момент времени s. Рассмотрим случайный процесс $X_s = (y_j(s),\ j \in I_1,\dots,I_r) \in \mathcal{S}'$, являющийся «проекцией» Y_s. Заметим, что процесс X_s не является марковским. Наблюдение за процессом X_s означает, что мы интересуемся только бесконечными источниками и не обращаем внимания на конечные источники.

Выберем следующую начальную конфигурацию цепи $\mathcal{L}$:

$$Y_0 = (y_i^0,\ i \in F_1,\dots,F_r,\ [x_jN],\ j \in I_1,\dots,I_r).$$

Другими словами, мы берем *любую* конфигуацию очередей конечных источников

$$y^0 = (y_i^0,\ i \in I_1, \ldots, I_r)$$

и конфигурацию длинных очередей требований из бесконечных источников

$$([x_j N],\ j \in I_1, \ldots, I_r),\ x \in \mathbf{R}_+^{K_1+\cdots+K_r}.$$

Обозначим через $X_s([xN]; y^0)$ «спроектированный процесс» X_s, соответствующий такому выбору начальной конфигурации.

Предложение 22. (Эйлеровский предел для составной системы.) *Для любого выбора y^0 следующая сходимость имеет место для всех $x \in \mathbf{R}_+^{K_1+\cdots+K_r}$.*

$$\frac{1}{N} X_{[tN]}([xN]; y^0) \to U_t x \ (N \to \infty) \quad (\textit{почти наверное}).$$

Как и в пункте 3.3.3, мы можем теперь рассмотреть для динамической системы U_t время $\tau(x)$ достижения нуля из точки x и вычислить его явно (это имеет смысл, конечно, лишь в случае, когда приоритетная система приводима). Это дает эффективных способ решения *оптимизационной задачи*, состоящей в наилучшем выборе порядка приоритетов между различными типами требований (см. пункт 3.3.3).

Аналогичный анализ возможен также для составной системы пункта 3.5.2.

A Идеология индуцированных цепей и второго векторного поля

В этом добавлении мы приводим краткую сводку обозначений, определений и некоторых результатов, которые активно используются в статье. Подробное изложение содержится в работах [3.5, 8].

A.1 Сети с очередями как цепи Маркова в $\mathbf{Z}_+^\nu$

Рассмотрим класс сетей с ν очередями. Все требования в каждой очереди предполагаются идентичными. Таким образом, состояние сети в момент t определяется вектором

$$\alpha = (n_1(t), \ldots, n_\nu(t)),$$

где n_i — число требований в очереди i, $n_i = 0, 1, 2, \ldots$.

Предположим, что стохастическая динамика, описывающая эволюцию состояний, является марковской. Мы имеем цепь Маркова с непрерывным временем с пространством состояний $\mathbf{Z}_+^\nu$. Обозначим $\lambda_{\alpha\beta}$ интенсивность перехода

$$\alpha = (n_1(t), \ldots, n_\nu(t)) \ \longrightarrow\ \beta = (n_1(t) + i_1, \ldots, n_\nu(t) + i_\nu).$$

Для многих сетей, в том числе и для тех, которые мы собираемся рассматривать в дальнейшем, эти цепи Маркова представляют собой специальный класс случайных блужданий, который мы описываем ниже.

А.2 Максимально однородные случайные блуждания в $\mathbf{Z}_+^\nu$

Для любых $1 \leqslant i_1 \leqslant \cdots \leqslant i_k \leqslant \nu$ *гранью* в положительном октанте $\mathbf{R}_+^\nu = \{(r_1, \ldots, r_\nu)\colon r_i \geqslant 0\}$ мы назовем следующее множество:

$$\Lambda(i_1, \ldots i_k) \equiv \{i_1, \ldots i_k\} \stackrel{\text{def}}{=} \stackrel{\text{def}}{=} \{(r_1, \ldots, r_\nu)\colon r_i > 0,\ i \in \{i_1, \ldots i_k\};\ r_i = 0,\ i \notin \{i_1, \ldots i_k\}\}.$$

Важно обратить внимание на то, что грань не содержит своей границы.

Ниже мы будем рассматривать класс случайных блужданий в $\mathbf{Z}_+^\nu$, удовлетворяющий следующим условиям.

Условие $\mathbf{A_0}$ (*максимальная однородность*). Для любой грани Λ и для всех $a \in \Lambda \cap \mathbf{Z}_+^\nu$ интенсивности скачков удовлетворяют равенству

$$\lambda_{\alpha\beta} = \lambda_{\alpha+a,\beta+a}\ \forall \alpha \in \Lambda \cap \mathbf{Z}_+^\nu\ \forall \beta \in \mathbf{Z}_+^\nu.$$

Условие $\mathbf{A_1}$ (*ограниченность скачков*).

$$\lambda_{\alpha\beta} = 0 \quad \text{для } \|\alpha - \beta\| > d,$$

где $d > 0$ и $\|\alpha\| = \max\limits_i |\alpha_i|$, $\alpha = (\alpha_1, \ldots, \alpha_\nu)$.

Для любой точки $\alpha \in \mathbf{Z}^\nu$ определим *средний снос* следующим образом:

$$M(\alpha) = \sum_\beta (\beta - \alpha)\lambda_{\alpha\beta}.$$

Теперь определим *первое векторное поле* на $\mathbf{R}^\nu$, положив его постоянным на каждой грани и равным

$$M^\Lambda = M(\alpha), \quad \alpha \in \Lambda.$$

А.3 Индуцированные цепи и второе векторное поле

Для любой грани $\Lambda \neq \{1, \ldots, \nu\}$ мы выберем произвольную точку $a \in \Lambda \cap \mathbf{Z}_+^\nu$ и проведем через эту точку плоскость C^Λ размерности $\nu - |\Lambda|$ перпендикулярно Λ. Определим *индуцированную цепь* Маркова с пространством состояний C^Λ и интенсивностями переходов

$${}_\Lambda\lambda_{\alpha\beta} = \lambda_{\alpha\beta} + \sum_{\gamma \notin \beta} \lambda_{\alpha\gamma}\ \forall \alpha, \beta \in C^\Lambda,$$

где сумма взята по всем $\gamma \in \mathbf{Z}_+^\nu$, таким что прямая, соединяющая γ и β, перпендикулярна C^Λ. Для этой индуцированной цепи мы будем использовать обозначение $\mathcal{L}^\Lambda$. Важно отметить, что эта конструкция не зависит от выбора a.

Предположение Е. *Для любой грани* Λ *индуцированная цепь* $\mathcal{L}^\Lambda$ *неприводима.*

Назовем грань Λ *эргодичной*, если цепь $\mathcal{L}^\Lambda$ эргодична. В том случае, когда $\mathcal{L}^\Lambda$ эргодична, обозначим через π^Λ ее стационарное распределение вероятностей.

Введем в рассмотрение вектор $v^\Lambda = (v_1^\Lambda, \ldots, v_\nu^\Lambda)$, полагая

$$v_i^\Lambda = 0, \quad i \neq\in \Lambda,$$
$$v_i^\Lambda = \sum_{\gamma \in \mathcal{C}^\Lambda} \pi^\Lambda(\gamma) M_i(\gamma), \quad i \in \Lambda.$$

Интуитивный смысл вектора v^Λ — средний снос вдоль грани Λ.

Максимальная грань $\Lambda = \{1, \ldots, \nu\}$ по определению считается эргодичной, а вектор v^Λ, соответствующий ей, полагается равным

$$v^\Lambda \equiv M(\alpha), \ \alpha \in \Lambda \cap \mathbf{Z}_+^\nu.$$

Предположение Ж. $v_i^\Lambda \neq 0$ *для всех* $i \in \Lambda$.

Зафиксируем две грани Λ и Λ_1, такие что $\Lambda \supset \Lambda_1$, $\Lambda \neq \Lambda_1$. Предположим, что грань Λ эргодична. В этом случае определен вектор v^Λ. Для направленности компонент вектора v^Λ существуют три возможности:

все координаты v_i^Λ для $i \in \Lambda - \Lambda_1$ отрицательны; в этом случае мы назовем грань Λ *входящей* для Λ_1;

все координаты v_i^Λ, $i \in \Lambda - \Lambda_1$, положительны; в этом случае мы назовем грань Λ *выходящей* для Λ_1;

в иных случаях мы назовем грань Λ *нейтральной* для Λ_1.

Доказательства следующих утверждений можно найти в [3].

Утверждение 1. *Грань* Λ *размерности* $\nu - 1$ *эргодична в том и только том случае, когда* $v_i^{\{1,\ldots,\nu\}} < 0$ *для* $i \equiv \{1, \ldots, \nu\} - \Lambda$.

Утверждение 2. *Если все грани* Λ, *такие что* $\Lambda \supset \Lambda_1$, *эргодичны и входящие для* Λ_1, *то* Λ_1 *эргодична.*

Второе векторное поле.

Каждой точке $x \in \mathbf{R}_+^\nu$ припишем вектор $v(x)$ и назовем эту функцию вторым векторным полем. Оно может быть многозначным на некоторых неэргодичных гранях.

Для эргодичных граней Λ положим

$$v(x) \equiv v^\Lambda, \quad x \in \Lambda.$$

Если грань Λ_1 неэргодична, то в каждой точке $x \in \Lambda_1$ функция $v(x)$ принимает все такие значения v^Λ, которые соответствуют граням Λ, выходящим для Λ_1. Те точки, в которых второе векторное поле определено неоднозначно, мы будем называть *точками вспоения.*

Утверждение 3. *Если для некоторой эргодичной грани* Λ *все компоненты* v^Λ *положительны, то случайное блуждание неэргодично.*

А.4 Вероятностный смысл второго векторного поля

Для произвольной грани Λ введем обозначение Λ_{R_1,R_2} для следующего множества:

$$\Lambda_{R_1,R_2} = \{(r_1,\dots,r_\nu):\ r_i > R_1,\ i \in \Lambda;\ r_i \leqslant R_2,\ i \notin \Lambda\}.$$

Пусть ξ_t обозначает состояние случайного блуждания в момент t.

В [3,8] доказано следующее утверждение.

Утверждение 4. *Пусть грань* Λ *эргодична. Тогда для любых чисел* $R_2, \epsilon, \sigma > 0$ *можно указать такие* $t_0 > 0$ *и* $R_1 > 0$, *что для любой точки* $\alpha \in \Lambda_{R_1,R_2} \cap \mathbf{Z}_+^\nu$ *и всех* $t \geqslant t_0$

$$\mathsf{P}\{\|\xi_t - (\alpha + tv^\Lambda)\| > t\epsilon\} < \sigma; \quad \xi_0 = \alpha.$$

Другими словами, если начальное состояние близко к эргодической грани Λ, но достаточно далеко от других граней Λ', $\Lambda \not\subset \Lambda'$, то по истечении некоторого времени (достаточно долгого, но меньшего, чем расстояние до ближайшей из упомянутых граней Λ') в индуцированной цепи $\mathcal{L}^\Lambda$ установится режим, близкий к стационарному. Тогда v^Λ — это просто средний сдвиг вдоль грани Λ для этого установившегося режима.

Литература

[1] Baskett F., Chandy K. M., Muntz R. K., Palacios F. G. Open, closed, and mixed networks of queues with different classes of customers // Journal of Association for Computing Machinery. — 1975. — V. 22, № 2. — P. 248–260.

[2] Botvich D. D., Zamyatin A. A. On fluid approximations for conservative networks // Markov Processes Relat. Fields. — 1995. — V. 1, № 1.

[3] Fayolle G., Malyshev V. A., Menshikov M. V. Topics in constructive theory of countable Markov chains. — Cambridge Univ. Press, 1994.

[4] Filin A. V., Malyshev V. A., Manita A. D. Probabilistic models of computer architectures. INRIA, Rapport de Recherche № 2419, 1994.

[5] Malyshev V. A. Networks and dynamical systems // Adv. Appl. Prob. — 1993. — V. 25. — P. 140–175.

[6] Волковинский М. И., Кабалевский А. Н. Анализ приоритетных очередей. — М.: Энергоиздат, 1981.

[7] Джейсуол Н. К. Очереди с приоритетами. — М.: Мир, 1973.

[8] Малышев В. А., Меньшиков М. В. Эргодичность, непрерывность и аналитичность счетных цепей Маркова // Труды Московского математического общества. — 1979. — Т. 39. — С. 3–48.

[9] Вероятностный анализ сетей связи. Препринт № 1. — Франко-русский центр МГУ, 1996.

Статья поступила в редакцию в январе 1996 г.