МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО»

Кафедра системного анализа и
автоматического управления

ИССЛЕДОВАНИЕ МОДЕЛИ СИСТЕМЫ ОБЛАЧНЫХ ВЫЧИСЛЕНИЙ

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студента 4 курса 481 группы
направления 27.03.03 — Системный анализ и управление
факультета компьютерных наук и информационных технологий
Зайцева Ильи Сергеевича

Научный руководитель
доцент, к. ф.-м. н. _____ Е. С. Рогачко

Заведующий кафедрой
к. ф.-м. н., доцент _____ И. Е. Тананко

Саратов 2025

ВВЕДЕНИЕ

Актуальность темы. Современные облачные вычисления стали неотъемлемой частью информационной инфраструктуры организаций, обеспечивая ее гибкость, масштабируемость и экономическую эффективность. Однако, с ростом сложности облачных систем и увеличением требований к качеству обслуживания (Quality of Service-QoS) возникают новые задачи, связанные с оптимизацией производительности и управлением ресурсами. В последнее время наблюдается значительное увеличение количества исследований, посвященных моделированию облачных систем, но с комплексным анализом их характеристик с использованием сетей массового обслуживания связано только небольшое число работ. Существующие исследования можно разделить на две категории: работы, посвященные общим принципам облачных вычислений [1-3], и исследования, фокусирующиеся на применении теории массового обслуживания для оптимизации облачных систем [4-7]. Настоящая работа относится ко второй категории.

Цель бакалаврской работы — разработка математической модели системы облачных вычислений на основе результатов теории массового обслуживания и проведение комплексного анализа её характеристик.

Поставленная цель определила **следующие задачи**:

— анализ существующих подходов к моделированию облачных систем, включая методы управления ресурсами и оптимизации производительности;

— разработка математической модели системы облачных вычислений в виде открытой сети массового обслуживания;

— создание алгоритма для вычисления характеристик сети Джексона, включая расчет стационарных вероятностей состояний сети и показателей производительности;

— реализация программы для моделирования работы системы облачных вычислений и визуализации результатов;

— проведение численных экспериментов, анализ полученных результатов и оптимизация параметров системы для улучшения её эффективности.

Методологические основы исследования систем облачных вычислений и анализа их характеристик представлены в работах R. Buyya, J. Broberg, A. Goscinski[1], Y. Zhang, Y. Zhou[2], M. Gribaudo, P. Piazzolla[3], D. Kliazovich,

P. Bouvry, S. Khan[4], R. Ghosh, S. Naik[5], S. Panda, P. Jana[6], H. Fernandez, G. Pierre, T. Kielmann[7].

Практическая значимость бакалаврской работы. Полученные в работе результаты имеют практическую ценность для IT-компаний, предоставляющих облачные услуги. Результаты работы могут быть использованы:

— для проектирования облачных инфраструктур с заданными характеристиками качества обслуживания;
— анализа узких мест существующих систем;
— оценки эффективности различных стратегий распределения ресурсов.

Структура и объем работы. Бакалаврская работа состоит из введения, 3 разделов, заключения, списка использованных источников и приложения. Общий объем работы — 51 страница, из них 45 страниц — основное содержание, включая 15 рисунков и 2 таблицы, список использованных источников информации — 20 наименований.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Первый раздел «Математическая модель системы облачных вычислений» посвящен описанию модели системы облачных вычислений, основанной на результатах теории массового обслуживания.

В подразделе 1.1 приведено описание системы облачных вычислений. Система облачных вычислений представляет собой комплекс аппаратных и программных ресурсов, предоставляемых пользователям через Интернет. Основная цель таких систем — обеспечить доступ к вычислительным мощностям и хранению данных по требованию. Важным аспектом функционирования систем облачных вычислений является их способность динамически адаптироваться к изменяющимся потребностям пользователей, обеспечивая при этом высокую производительность, масштабируемость и надежность. На рисунке 1 изображена парадигма облачных вычислений.

Подраздел 1.2 посвящен описанию математической модели системы облачных вычислений в виде сети массового обслуживания. Модель системы облачных вычислений может быть представлена как открытая сеть массового обслуживания (СеМО), состоящая из нескольких узлов, каждый из которых представляет собой систему массового обслуживания (СМО) (рисунок 2). Опишем компоненты системы облачных вычислений и соответствующие им элементы модели.

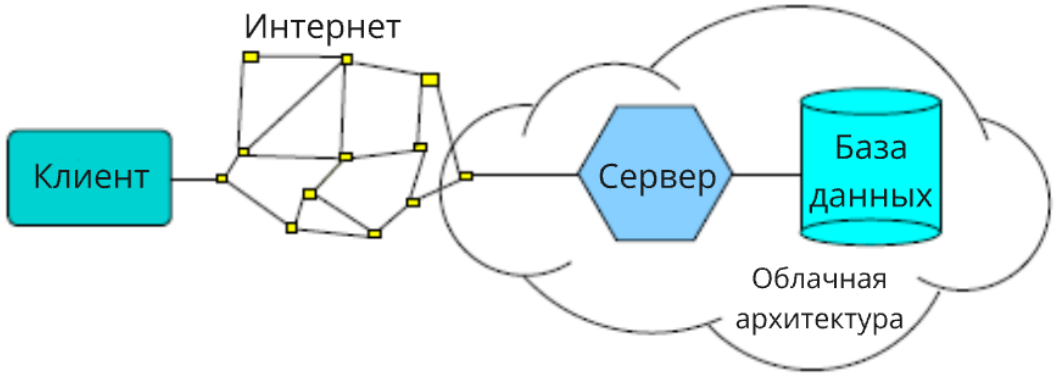

Рисунок 1 – Парадигма облачных вычислений

— Entering Server (ES) - Входной сервер:
  – Функция: Балансировка нагрузки и распределение требований между серверами для обработки.
  – Модель: СМО M/M/1 с параметрами λ(интенсивность поступления) и L (интенсивность обслуживания).
— Processing Server (PS) - Узел обработки требований:
  – Функция: Обработка требований пользователей. Узел обработки представляет собой физические или виртуальные серверы облачной инфраструктуры.
  – Модель: СМО M/M/m, где m - количество приборов, μ - интенсивность обслуживания одного прибора.
— Database Server (DS) - Сервер базы данных:
  – Функция: Доступ к данным, хранящимся в облаке. Сервер базы данных отвечает за операции чтения и записи.
  – Модель: СМО M/M/1 с параметрами $\delta\gamma$ (интенсивность поступления требований) и D (интенсивность обслуживания), здесь δ - вероятность обращения к серверу базы данных, а γ - интенсивность потока требований к базе данных.
— Output Server (OS) - Выходной сервер:
  – Функция: Передача обработанных данных пользователю.
  – Модель: СМО M/M/1 с интенсивностью обслуживания O/F.
— Client Server (CS) - Клиентский сервер:
  – Функция: Получение требований и отправка ответов пользователю.
  – Модель: СМО M/M/1 с интенсивностью обслуживания C/F.

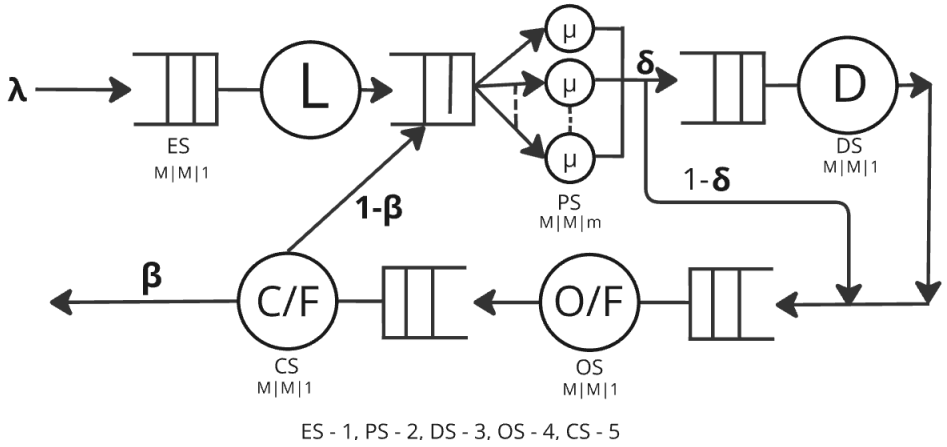

Рисунок 2 – Модель в виде сети массового обслуживания

Опишем параметры СМО:

1) Интенсивность поступления требований (λ): Это среднее число требований, поступающих в систему за единицу времени. В контексте облачных вычислений λ может зависеть от числа пользователей и их активности.

2) Интенсивность обслуживания (μ): Это среднее число требований, которые могут быть обработаны прибором за единицу времени. Чем выше значение μ, тем более производителен прибор.

3) Число приборов (m): Это количество приборов, которые работают параллельно в узле обработки. Увеличение числа приборов может повысить общую пропускную способность системы.

4) Параметры сервера базы данных (δ, γ, D):

— δ - вероятность обращения к серверу базы данных.

— γ - интенсивность потока требований к базе данных.

— D - интенсивность обслуживания требований.

5) Параметры выходного сервера (O, F):

— O - скорость передачи данных.

— F - средний размер данных ответа.

6) Параметры клиентского сервера (C, F, β):

— C - скорость передачи данных клиентского сервера.

— F - средний размер данных ответа.

— β - вероятность выхода из системы.

Эта модель позволяет детально анализировать производительность системы и выявлять узкие места. Знание времени отклика и других показателей по-

могает оптимизировать распределение ресурсов, обеспечивая требуемое качество обслуживания (QoS) для пользователей.

Описанная модель системы облачных вычислений по сути является сетью Джексона[8]. Приведем соответствие обозначений в модели системы облачных вычислений обозначениям для сети Джексона.

Для наглядности изобразим нашу СеМО в виде графа интенсивностей потоков между системами(рисунок 3).

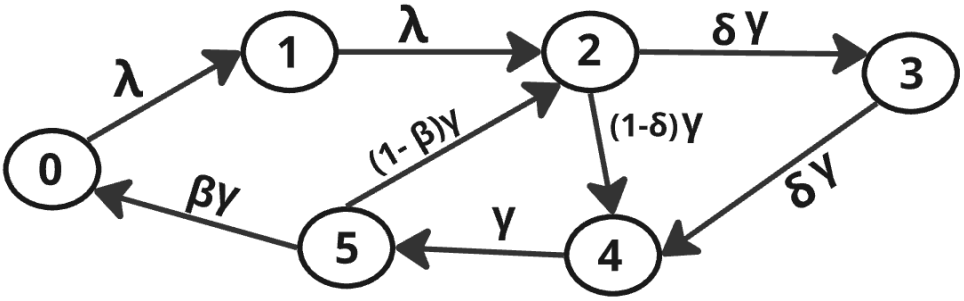

Рисунок 3 – Граф интенсивностей потоков между системами в СеМО

Модели на рисунке 2 сопоставим сеть массового обслуживания, определяемую набором $\langle L, 1, \lambda_0, M, \Theta, \overline{\kappa}, \overline{\mu}, FCFS\rangle$, где $L = 5$, $\overline{\mu} = (\mu_1, \mu_2, \mu_3, \mu_4, \mu_5)$ и $\overline{\kappa} = (\kappa_1, \kappa_2, \kappa_3, \kappa_4, \kappa_5)$.

Установим соответствие:

1. ES - СМО типа $M|M|1$, где $\kappa_1 = 1$, а $\mu_1 = L$;
2. PS - СМО типа $M|M|m$, где $\kappa_2 = m$, а $\mu_2 = \mu$;
3. DS - СМО типа $M|M|1$, где $\kappa_3 = 1$, а $\mu_3 = D$;
4. OS - СМО типа $M|M|1$, где $\kappa_4 = 1$, а $\mu_4 = \frac{O}{F}$;
5. CS - СМО типа $M|M|1$, где $\kappa_5 = 1$, а $\mu_5 = \frac{C}{F}$.

Интенсивность поступления требований в сеть массового обслуживания $\lambda_0 = \lambda$. Переходы между системами определяются маршрутной матрицей:

$$\Theta = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & \delta & 1-\delta & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ \beta & 0 & 1-\beta & 0 & 0 & 0 \end{bmatrix}.$$

Подраздел 1.3 посвящен описанию сетей Джексона и метода их анализа.

Введены обозначения, связанные с СеМО, также рассмотрена теорема Джексона, ее доказательство и алгоритм, реализующий метод анализа однородных открытых экспоненциальных СеМО.

Второй раздел «Описание алгоритма и программы для анализа системы облачных вычислений» содержит описание алгоритма и программы для анализа модели рассмотренной системы облачных вычислений.

В подразделе 2.1 представлен алгоритм программы для анализа системы облачных вычислений и таблица идентификаторов, используемых в алгоритме.

Алгоритм программы состоит из 8 блоков.

В блоках 1 и 2 программа запрашивает у пользователя ввод основных параметров системы. После ввода данных программа проверяет их корректность. Инициализируются характеристики, необходимые для дальнейших вычислений.

В блоке 3 решается система уравнений для нахождения вектора ω относительных интенсивностей потоков требований в СМО:

$$A\omega = b,$$

где $A = \Theta^T - I$, I — единичная матрица, а b — вектор, последний элемент которого равен 1, остальные — 0.

В блоке 4 вычисляются интенсивности λ_i потоков требований для каждой СМО по формуле:

$$\lambda_i = \frac{\lambda_0 \omega_i}{\omega_0}, \quad i = 1, \ldots, L.$$

В блоке 5 проверяется условие существования стационарного режима. Для этого для каждой СМО вычисляется параметр ψ_i:

$$\psi_i = \frac{\lambda_i}{\kappa_i \mu_i}.$$

Если $\psi_i \geq 1$, это означает, что СМО перегружена, и условие существования стационарного режима нарушено. В этом случае программа выводит предупреждение. Иначе выполняются блоки 6-8.

В блоке 6 вычисляются вероятности P_{i0} — это вероятности того, что в СМО нет требований, по формуле:

$$P_{i0} = \frac{1}{\sum_{n=0}^{\kappa_i - 1} \frac{(\kappa_i \psi_i)^n}{n!} + \frac{(\kappa_i \psi_i)^{\kappa_i}}{\kappa_i!(1-\psi_i)}}.$$

В блоке 7 вычисляются следующие характеристики систем:

— Средняя длина очереди

$$b_i = \frac{P_{i0} \kappa_i^{\kappa_i} \psi_i^{\kappa_i + 1}}{\kappa_i!(1-\psi_i)^2}.$$

— Среднее число занятых приборов $h_i = \psi_i \kappa_i$.

— Среднее число свободных приборов $g_i = \kappa_i - h_i$.

— Среднее число требований в СМО $n_i = b_i + h_i$.

— Среднее время пребывания требований в системе $u_i = n_i / \lambda_i$.

— Среднее время ожидания требований $w_i = b_i / \lambda_i$.

В блоке 8 вычисляется среднее время отклика по формуле:

$$\overline{\tau} = \frac{\sum_{i=1}^{L} n_i}{\lambda_0}.$$

Подраздел 2.2 посвящен описанию программы для анализа систем облачных вычислений. Программа была написана в среде PyCharm на языке программирования Python. В программе реализован алгоритм метода анализа открытых сетей обслуживания, основанного на композиции известных расчетных формул для систем обслуживания типа $M/M/\kappa_i$.

Подраздел 2.3 посвящен примеру использования программы. Разработанная программа имеет оконный интерфейс. При запуске программы открывается окно, где пользователю предлагается ввести параметры системы. Далее необходимо нажать на кнопку «Рассчитать», для того чтобы программа проверила корректность введенных параметров. После ввода параметров вывод рассчитанного среднего времени отклика системы облачных вычислений на запрос пользователя выполняется в отдельное окно, вывод остальных характеристик осуществляется в текстовый файл.

Третий раздел «Результаты исследования модели системы облачных вычислений» содержит результаты исследования зависимостей

различных характеристик функционирования системы облачных вычислений от изменения исходных параметров.

В примере 1 рассматривается базовый эксперимент, в котором система имеет следующие параметры: $\lambda_0 = 5$, $\delta = 0.3$, $\beta = 0.4$, $m = 2$, $\mu_1 = 10$, $\mu_2 = 8$, $\mu_3 = 12$, $\mu_4 = 15$, $\mu_5 = 20$. В примере описаны результаты расчета характеристик функционирования системы облачных вычислений. Было определено, что узких мест в системе нет и она справляется с нагрузкой.

В примере 2 рассматривается система как в примере 1 при увеличении значения вероятности δ обращения к серверу базы данных DS до 0,7. Было определено, что при увеличении δ среднее время отклика системы увеличилось. Это связано с тем, что увеличение δ приводит к увеличению нагрузки на систему, что увеличивает время обработки запросов.

В примере 3 описаны результаты расчета характеристик функционирования системы облачных вычислений при увеличении числа m приборов (серверов) в узле обработки запросов в 2 раза. Было определено, что при увеличении m среднее время отклика системы уменьшилось. Это связано с тем, что увеличение числа серверов в узле обработки запросов позволяет системе обрабатывать больше запросов одновременно, что снижает время отклика.

В примере 4 описаны результаты расчета характеристик функционирования системы облачных вычислений в зависимости от изменения интенсивности поступления запросов при изменении числа приборов (серверов) в узле обработки запросов. На рисунке 4 представлен график зависимости среднего времени T отклика системы от изменения интенсивности λ_0 поступления запросов при изменении m. Было определено, что при значении $m = 2$ система хорошо справляется с возрастанием нагрузки, а время отклика T растёт постепенно. Это говорит о стабильной работе узлов, пока нагрузка не достигает порогового значения, соответствующего нарушению условия стационарного режима функционирования системы. При значении $m = 4$, при тех же значениях λ_0, среднее время отклика T в начальных точках становится меньше, а рост значений T — более медленный. Это показывает, что ресурсы второго узла теперь лучше компенсируют рост нагрузки.

В примере 5 в рассмотренной системе есть узкое место из-за перегрузки во втором узле обработки запросов. Была проведена оптимизация системы и определено, что при увеличении числа m приборов во втором узле в 2-6

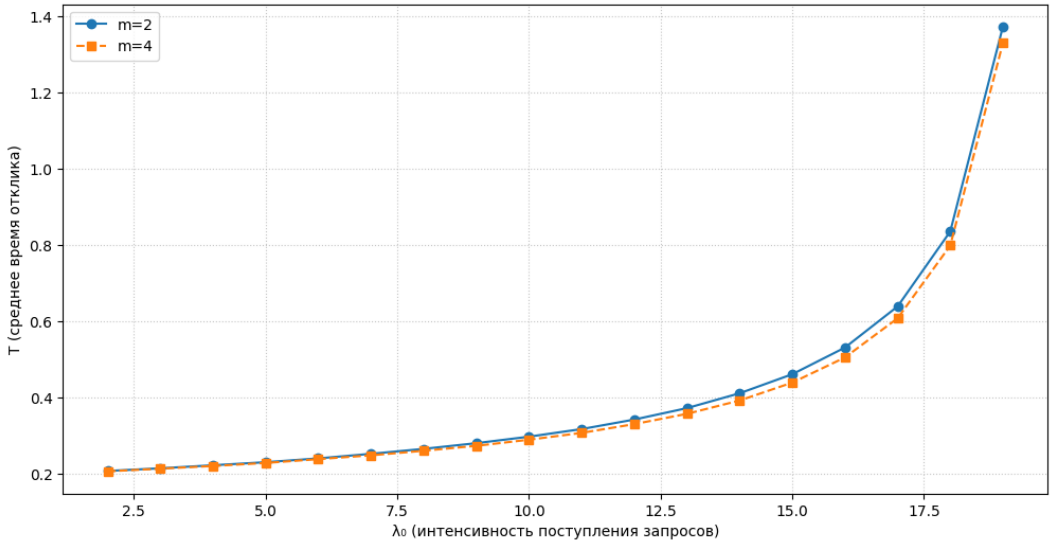

Рисунок 4 – График зависимости T от λ_0

раз(начальное $m = 2$), значение среднего времени отклика системы уменьшится, а также система стабилизируется. Было определено, что при увеличении m среднее время отклика T значительно снижается при увеличении количества приборов с 2 до 6. После $m = 6$ наблюдается плато — дальнейшее увеличение числа приборов даёт незначительное улучшение. Прирост эффективности снижается с каждым новым прибором (затраты растут, а выгода нет).

В примере 6 рассмотрена система из примера 5. Оптимизация происходит путем увеличения интенсивности обработки запросов во втором узле (μ_2). Было определено, что увеличение интенсивности обслуживания с 2 до 4 резко снижает T, особенно в диапазоне μ_2 от 2 до 2.7. Даже при фиксированном количестве приборов во втором узле $m = 2$, увеличение μ_2 показывает высокую эффективность.

После сравнения результатов примеров 5 и 6 было установлено, что изменение μ_2 — более эффективная мера на начальных этапах оптимизации, так как достигается больший эффект при меньших затратах. Однако, когда μ_2 достигнет предела, уменьшение T можно обеспечить посредством увеличения m.

В примере 7 в рассмотренной системе обнаружено узкое место из-за перегрузки пятого узла. Чтобы уменьшить среднее время отклика системы и разгрузить узел 5, была увеличена интенсивность обработки требований в клиентском сервере (μ_5) в 2 раза. Было определено, что даже небольшое уве-

личение μ_5 (на 0.1) приводит к значительному уменьшению среднего времени отклика T. Дальнейшее увеличение μ_5 обеспечивает быстрое уменьшение величины T. Это указывает на высокую чувствительность отклика системы к изменению производительности узла, когда он находится в состоянии перегрузки.

ЗАКЛЮЧЕНИЕ

В бакалаврской работе была разработана математическая модель системы облачных вычислений на основе результатов теории массового обслуживания. В качестве модели использовалась сеть Джексона - однородная открытая экспоненциальная сеть массового обслуживания.

Был разработан алгоритм для вычисления характеристик системы облачных вычислений, включающий расчет ключевых показателей производительности: среднего времени отклика, коэффициентов загрузки узлов. Программная реализация алгоритма выполнена на языке Python с использованием современных библиотек для научных вычислений. Разработанное программное обеспечение позволяет:

— моделировать работу системы облачных вычислений при различных ее параметрах;
— анализировать стационарные характеристики системы;
— визуализировать полученные результаты для принятия управленческих решений.

Проведенные численные эксперименты подтвердили адекватность разработанной модели и позволили получить практические рекомендации по оптимизации параметров системы облачных вычислений.

Основные источники информации:

1 Buyya, R. Cloud Computing: Principles and Paradigms / R. Buyya, J. Broberg, A. Goscinski. — Hoboken : Wiley, 2018. — 664 p.
2 Zhang, Y. Jackson Network Modeling for Cloud Performance Analysis / Y. Zhang, Y. Zhou // IEEE Transactions on Parallel and Distributed Systems. — 2020. — Vol. 31, No. 5. — P. 1124–1136.
3 Gribaudo, M. Performance Modeling of Cloud Centers Using Queueing Networks / M. Gribaudo, P. Piazzolla // The Journal of Supercomputing. — 2015. — Vol. 71. — P. 492–507.

4 Kliazovich, D. Energy-Efficient Cloud Computing Architectures / D. Kliazovich, P. Bouvry, S. Khan // Sustainable Computing: Informatics and Systems. — 2016. — Vol. 10. — P. 1–15.

5 Ghosh, R. QoS Optimization in Cloud Platforms / R. Ghosh, S. Naik // ACM Computing Surveys. — 2021. — Vol. 54, No. 3. — P. 1–35.

6 Panda, S. Task Scheduling Algorithms for Cloud Computing / S. Panda, P. Jana // Journal of Network and Computer Applications. — 2017. — Vol. 80. — P. 96–118.

7 Fernandez, H. SLA-Aware Resource Management in Clouds / H. Fernandez, G. Pierre, T. Kielmann // Future Generation Computer Systems. — 2019. — Vol. 91. — P. 540–552.

8 Митрофанов, Ю. И. Анализ сетей массового обслуживания: учебное пособие / Ю. И. Митрофанов. — Саратов : Научная книга, 2005. — 175 с.