МЕТОДЫ ОБРАБОТКИ ЗАПРОСОВ В СИСТЕМАХ УПРАВЛЕНИЯ БАЗАМИ ДАННЫХ ДЛЯ МНОГОПРОЦЕССОРНЫХ СИСТЕМ С ИЕРАРХИЧЕСКОЙ АРХИТЕКТУРОЙ

05.13.11 – математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей

Диссертация на соискание ученой степени кандидата физико-математических наук

А.В. Лепихов

Научный руководитель:
СОКОЛИНСКИЙ Леонид Борисович,
доктор физ.-мат. наук, профессор

Работа выполнена при поддержке Российского фонда фундаментальных исследований (проект 06-07-89148)

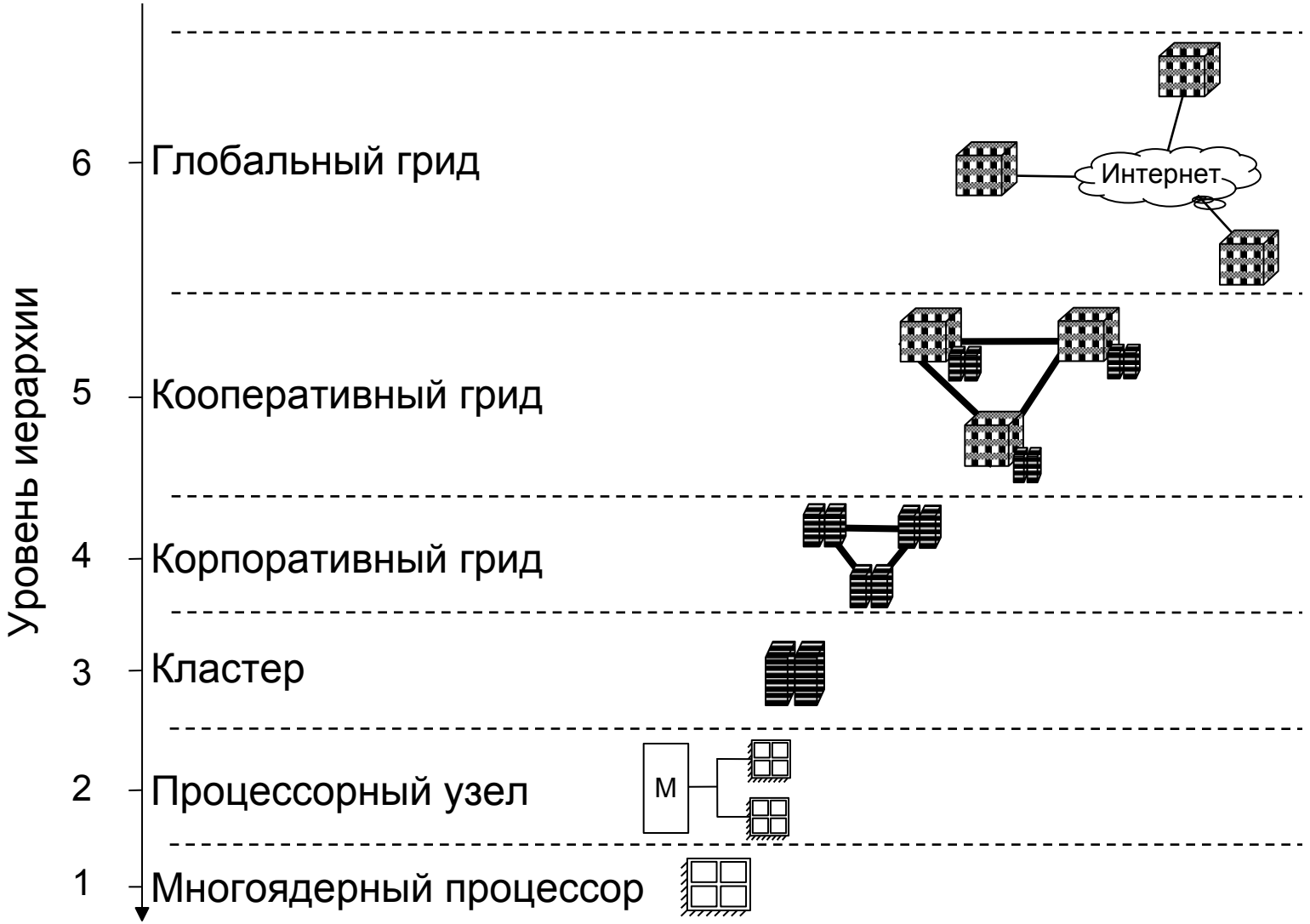
Цель диссертационной работы

Цель работы: разработка эффективных методов и алгоритмов параллельной обработки запросов, размещения данных и балансировки загрузки, ориентированных на многопроцессорные системы с иерархической архитектурой и их реализация в прототипе СУБД для многопроцессорных иерархий

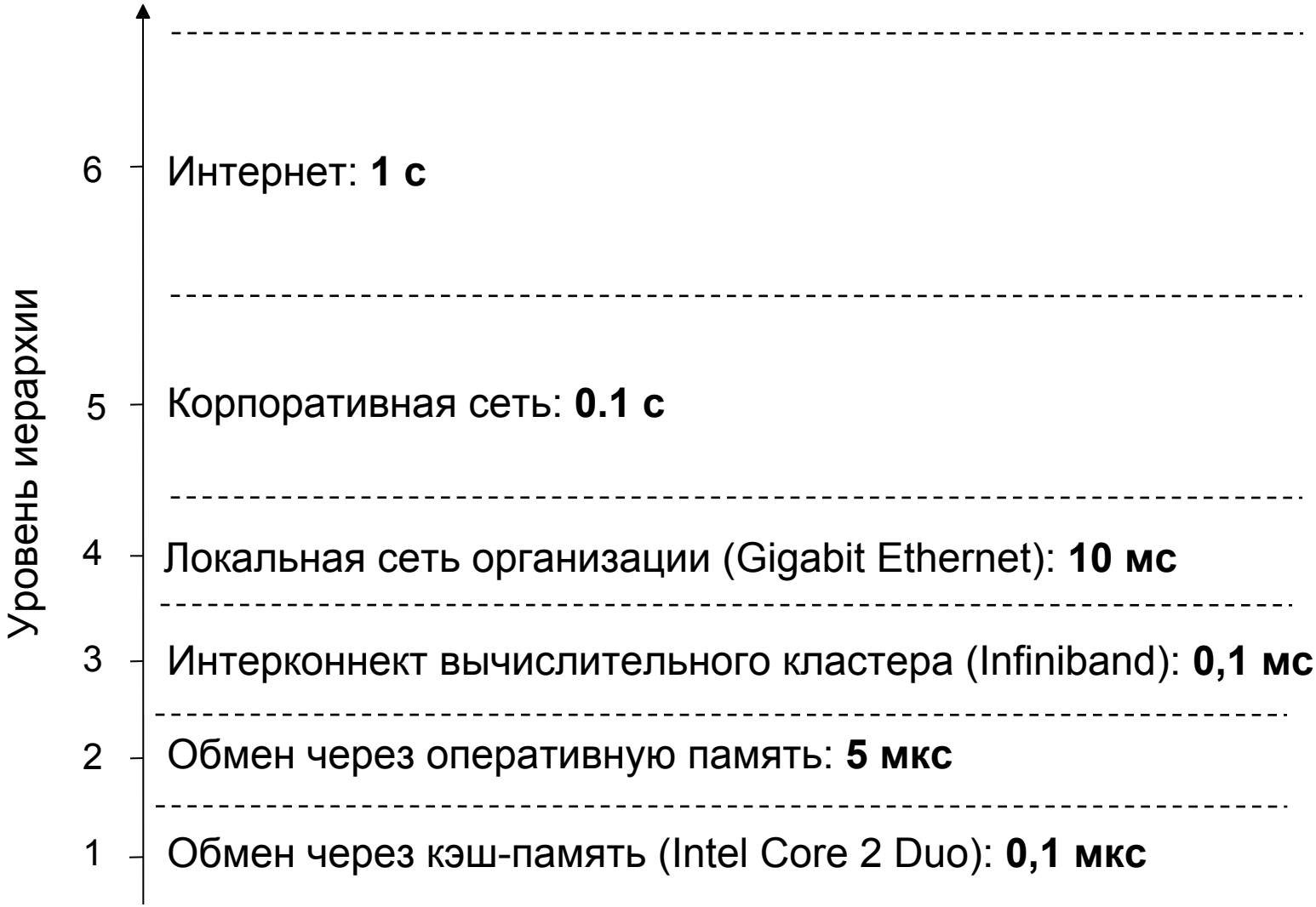
Основные задачи

1. Разработать и аналитически исследовать стратегию размещения и репликации базы данных для многопроцессорных иерархических систем
2. Разработать эффективный алгоритм динамической балансировки загрузки на основе предложенной стратегии размещения данных
3. Разработать метод параллельной обработки запросов для многопроцессорных иерархий, использующий предложенные стратегию размещения данных и алгоритм балансировки загрузки
4. Реализовать разработанные методы и алгоритмы в прототипе иерархической СУБД «Омега»
5. Провести вычислительные эксперименты для оценки эффективности предложенных решений

Структура многопроцессорной иерархии



Время передачи сообщения размером 1 МБ



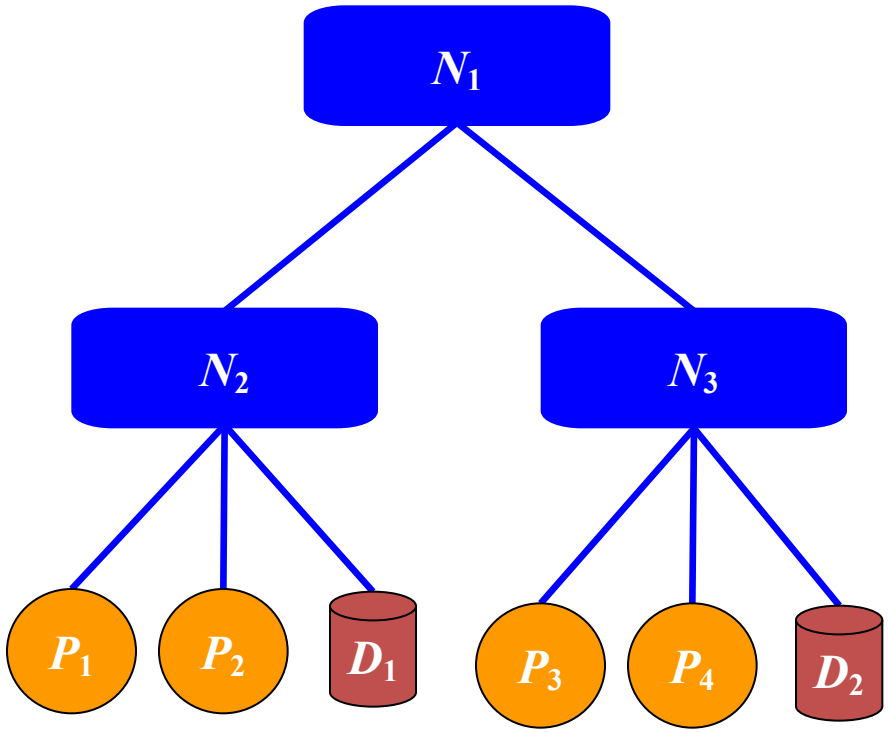
Многопроцессорная иерархическая система

Многопроцессорная иерархическая система – это многопроцессорная система, в которой процессоры объединяются в единую систему с помощью соединительной сети, имеющей иерархическую структуру и обладающей свойствами однородности по горизонтали и неоднородности по вертикали

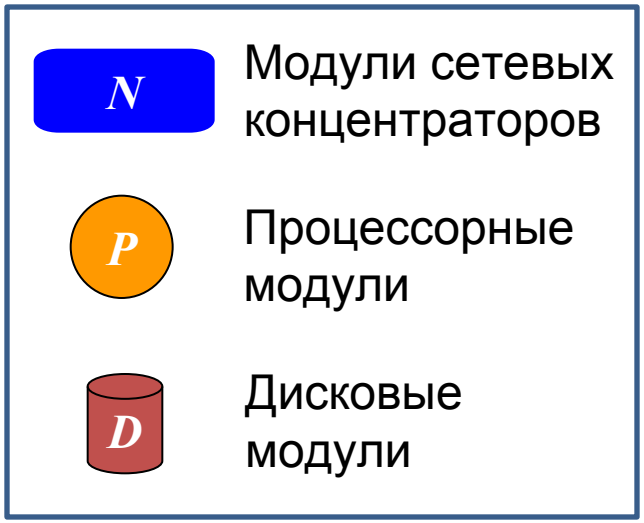
Свойства многопроцессорной иерархии

- Однородность по горизонтали:
  в пределах одного уровня иерархии скорость обменов между двумя процессорами является постоянной, независимо от того в каком поддереве иерархии эти процессоры находятся
- Неоднородность по вертикали:
  скорость обменов на разных уровнях иерархии существенно различается

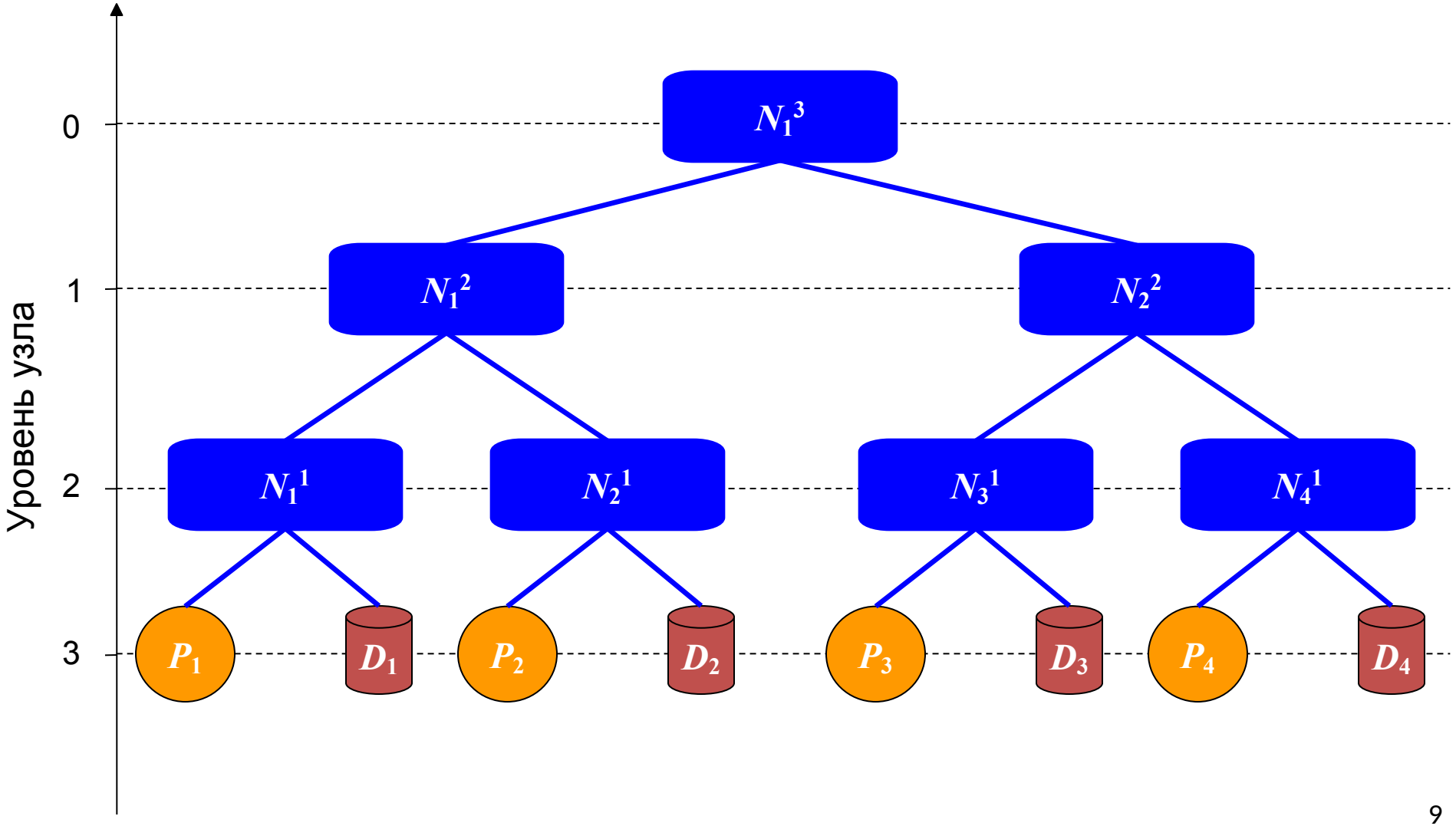
DM (Database Machine) модель



DM-дерево



Симметричная модель многопроцессорной иерархии



Иерархическая СУБД

Иерархическая СУБД - система управления *реляционными* базами данных для многопроцессорных систем с симметричной иерархической архитектурой

Сравнительный анализ

	Параллельная СУБД	Распределенная СУБД	Иерархическая СУБД
Контекстная независимость узла	+	–	+
Одноранговость соединительной сети	+	–	–
Фрагментный параллелизм	+	–	+
Репликация данных	–	+	+
Балансировка загрузки	+	–	+

Выводы

- Иерархическая СУБД совмещает в себе свойства как параллельной, так и распределенной СУБД
- Методы и алгоритмы обработки запросов параллельных и распределенных СУБД *не могут* быть прямо перенесены в иерархические СУБД
- Для иерархических СУБД необходимо разрабатывать *новые* методы и алгоритмы обработки запросов

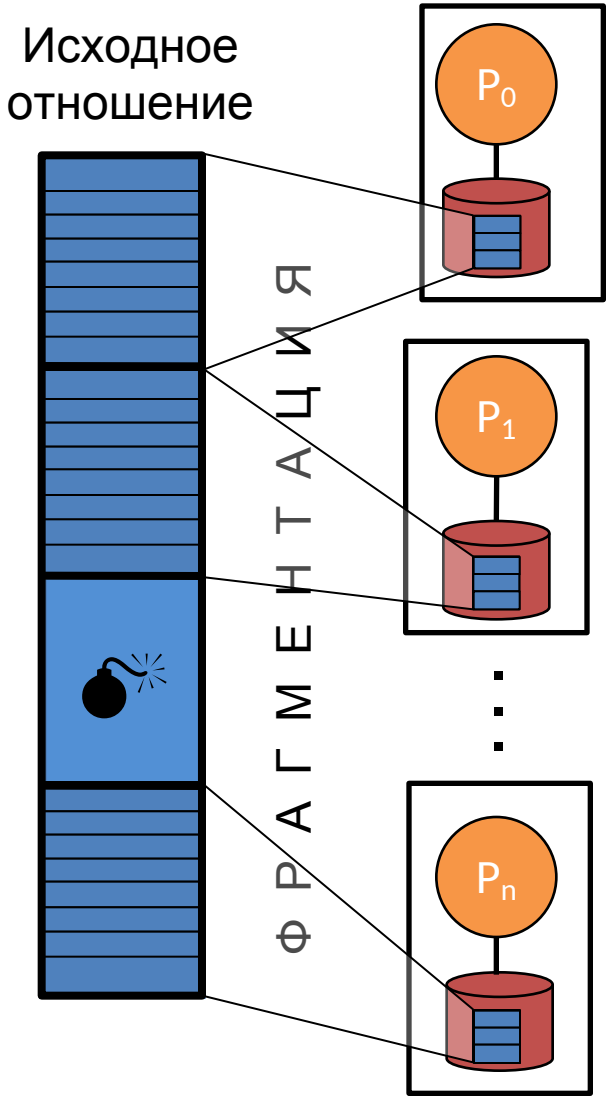
Метод обработки запросов в иерархических СУБД

- Метод частичного зеркалирования
- Метод балансировки загрузки

Метод частичного зеркалирования

- Стратегия распределения
- Стратегия репликации

Стратегия распределения

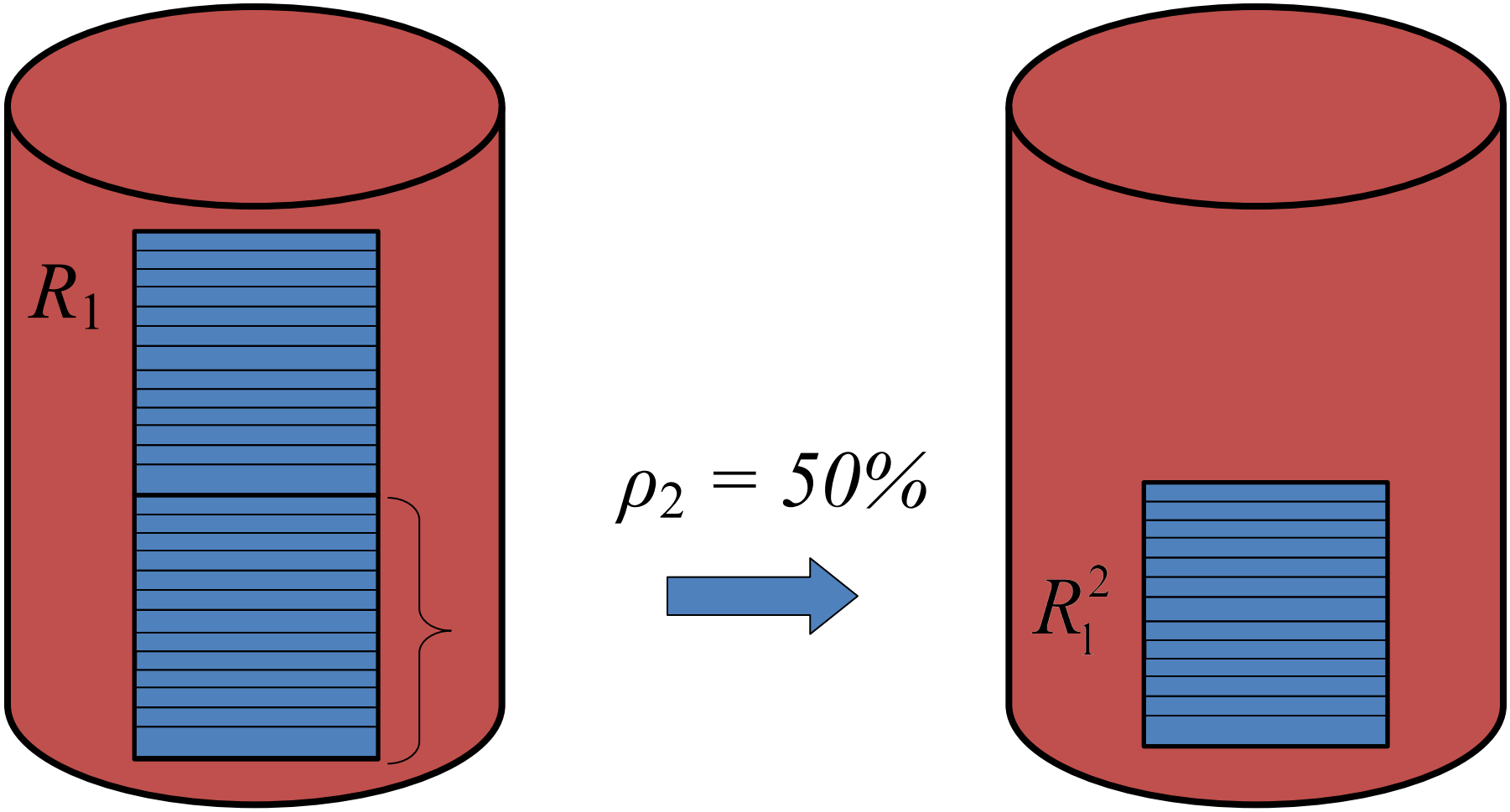


- Отношение разбивается на фрагменты, располагающиеся на различных дисках
- Фрагмент делится на логические сегменты, между которыми определено отношение порядка
- Сегмент является наименьшей единицей репликации

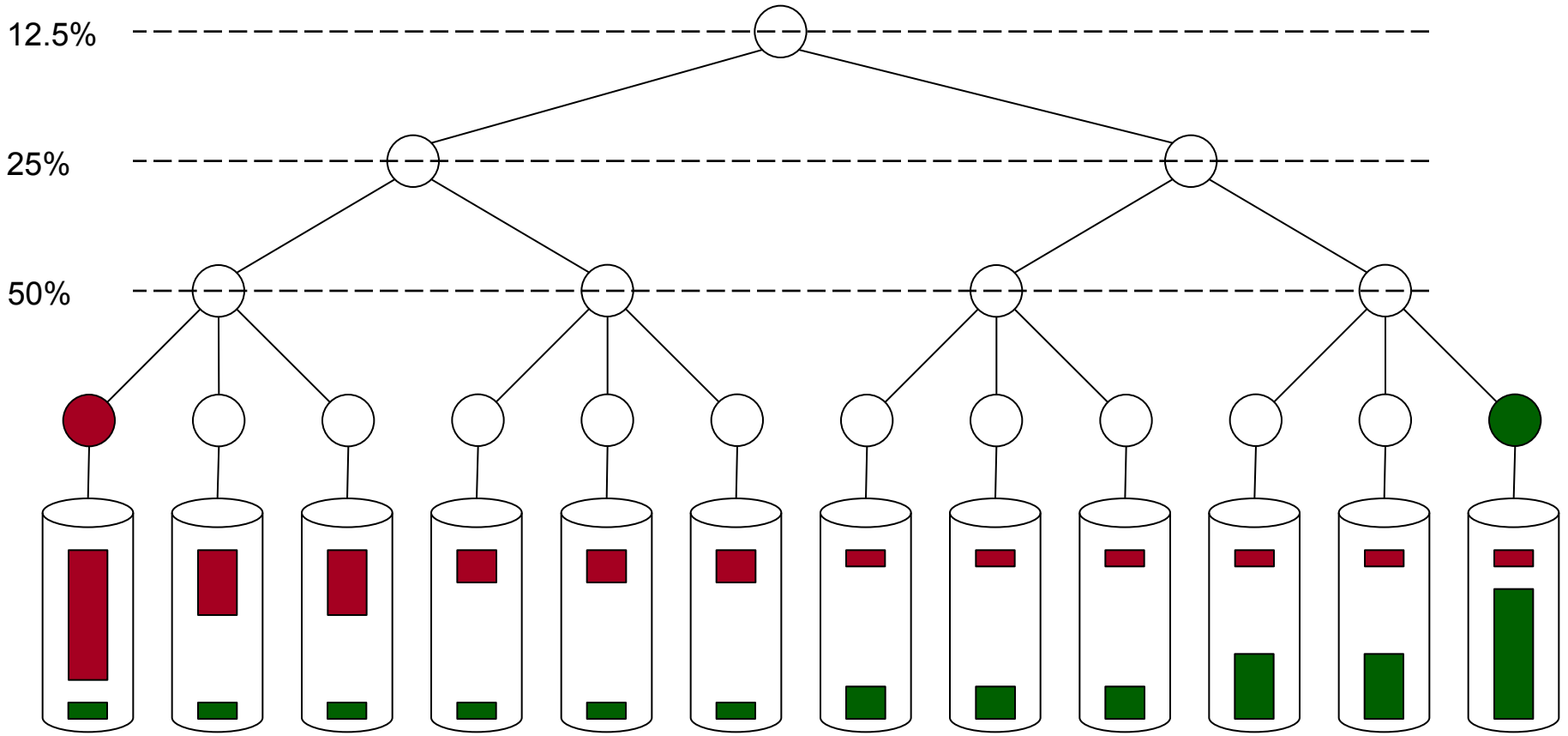
Стратегия репликации

- Фрагмент может иметь несколько (возможно неполных) зеркальных копий, называемых репликами, которые располагаются на других дисках
- На каждом диске может находиться не более одной реплики данного фрагмента
- Содержимое реплики однозначно определяется коэффициентом репликации ρ, назначенным диску, на котором хранится реплика

Построение реплики



Коэффициент репликации



Оценка трудоемкости покортежного формирования реплики

Теорема. Пусть T – регулярное *DM*-дерево высоты $H > 0$. Пусть фрагмент F_0 располагается на диске $d_0 \in D(T)$. Пусть M поддерево дерева T такое, что $1 \le l(M) \le H - 2$ и $d_0 \in D(M)$. Пусть M' – произвольное смежное с M поддерево дерева T; F_i – реплика фрагмента F_0, размещенная на диске $d_i \in D(M')$. Обозначим $t(F_i)$ – трудоемкость покортежного формирования реплики F_i при отсутствии помех. Тогда

$$t(F_i) = h\left(l(M) - 1\right) \times r\left(l(M) - 1\right) T\left(F_0\right) + O\left(h_0\right),$$

где:

$h(i)$ – трудоемкость узлов на уровне i

$r(i)$ – функция репликации для уровня i

$T(F)$ – количество кортежей во фрагменте F

$l(M)$ – уровень поддерева M

Нормальная функция репликации

1) для $l = H - 2$: $r(H-2) = \dfrac{1}{h(H-2)(d_{H-2} - 1)}$;

2) для $0 \le l < H - 2$: $r(l) = \dfrac{h(l+1)(d_{l+1} - 1) r(l+1)}{h(l)(d_l - 1) d_{l+1}}$.

$l(M)$ – уровень поддерева M

d_i – степень узла на уровне i

$h(i)$ – трудоемкость узла на уровне i

$r(i)$ – функция репликации для уровня i

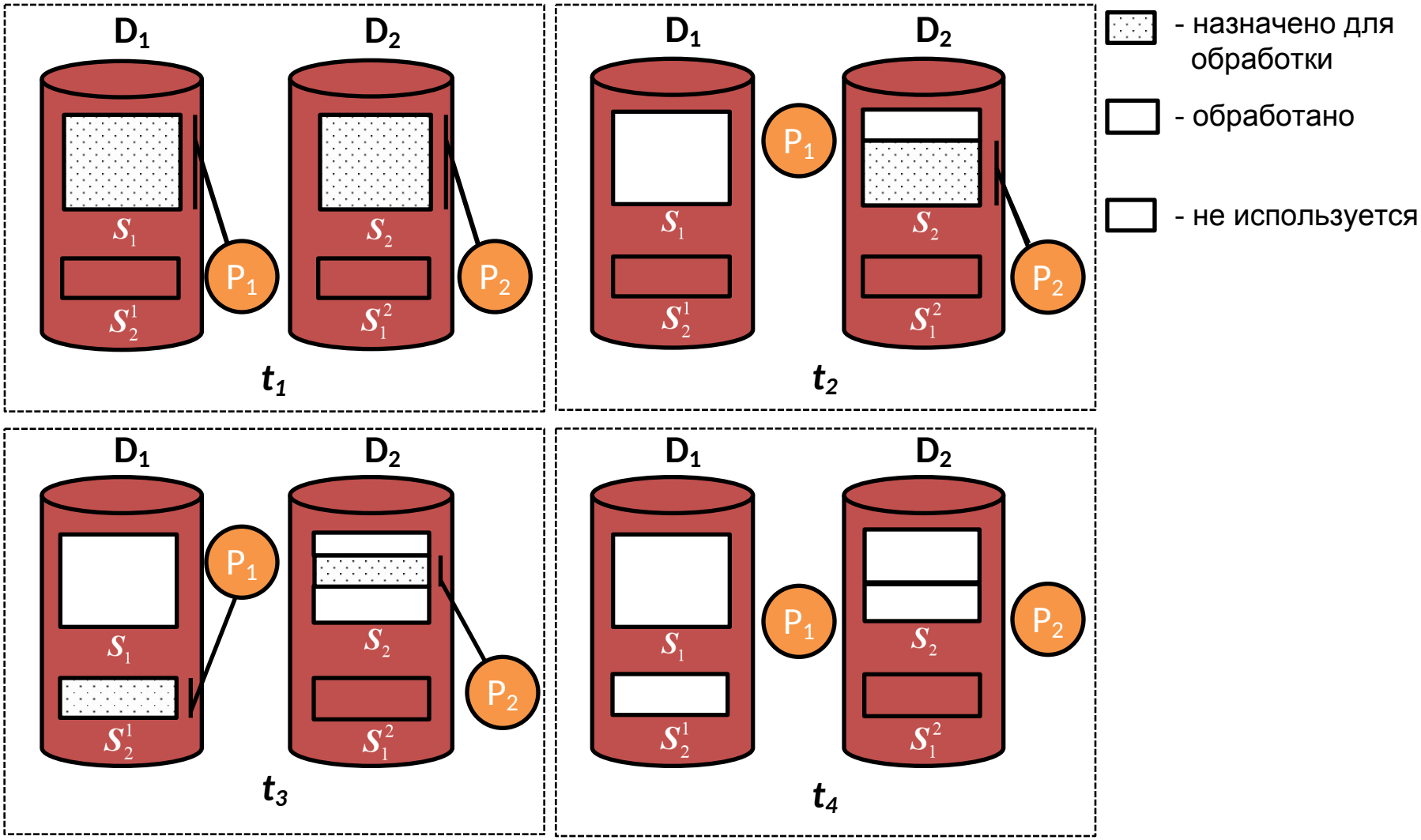
Оценка трудоемкости формирования реплик (без учета помех)

Теорема. Пусть T – регулярное DM-дерево высоты $H \geq 2$. Пусть $\mathbf{F}$ – множество фрагментов, составляющих базу данных. Пусть $\mathbf{R}$ – множество всех реплик всех фрагментов из множества $\mathbf{F}$, построенных с использованием нормальной функции репликации. Пусть $T(\mathbf{F})$ – размер базы данных в кортежах (здесь мы предполагаем, что все кортежи имеют одинаковую длину в байтах), $t(\mathbf{R})$ – суммарная трудоемкость покортежного формирования всех реплик без учета помех. Тогда

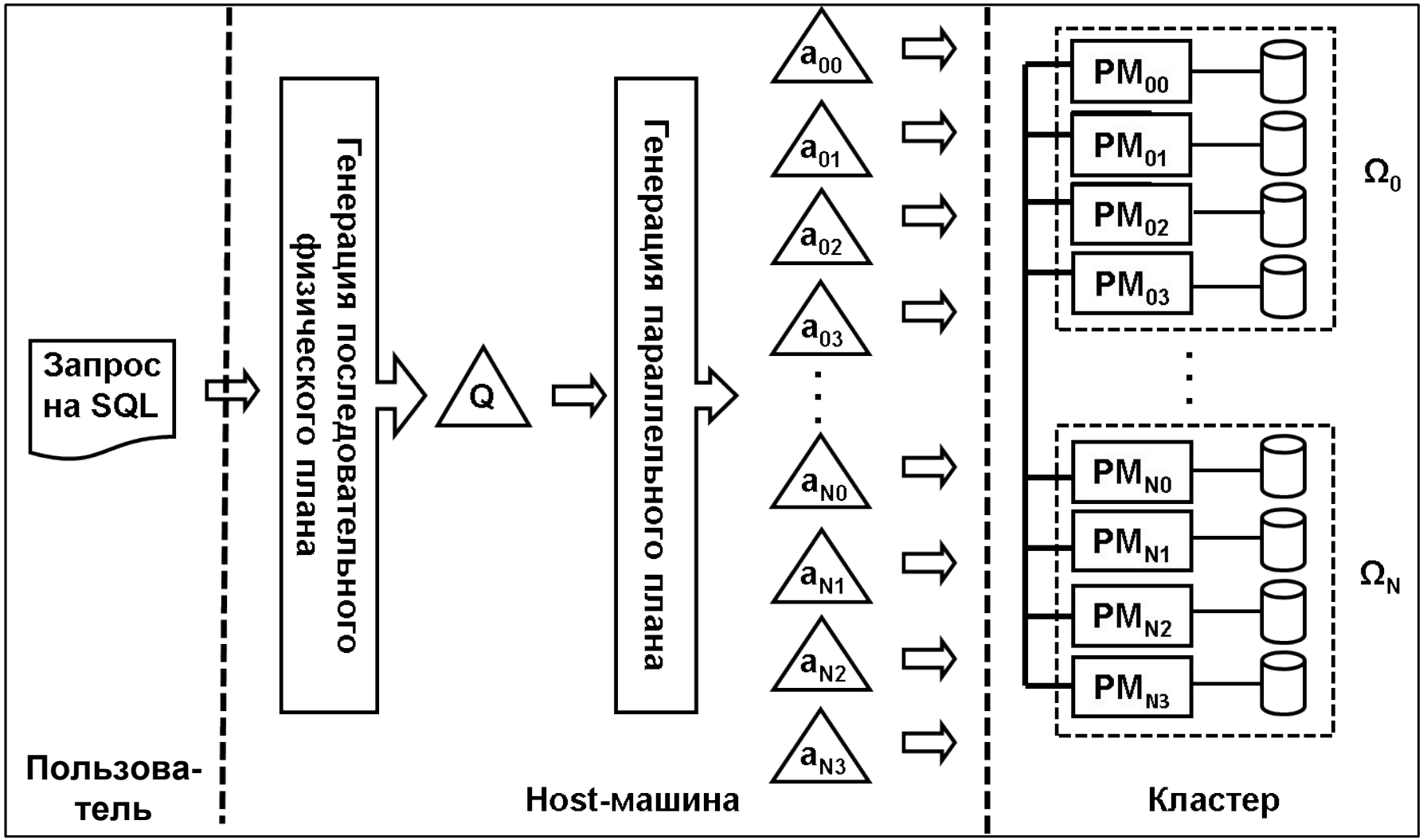
$$t(\mathbf{R}) \approx k\,T(\mathbf{F}),$$

где k – некоторая константа, не зависящая от $\mathbf{F}$.

Метод балансировки загрузки

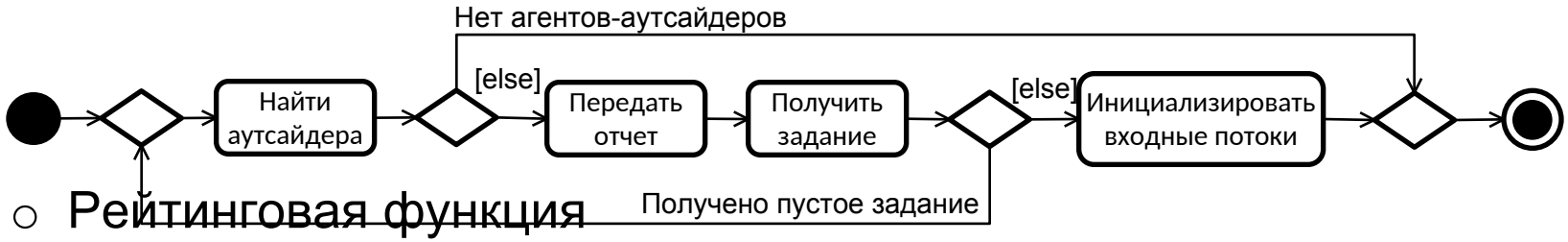


Обработка запросов



Механизмы балансировки

- Стратегия выбора аутсайдера



  - Рейтинговая функция

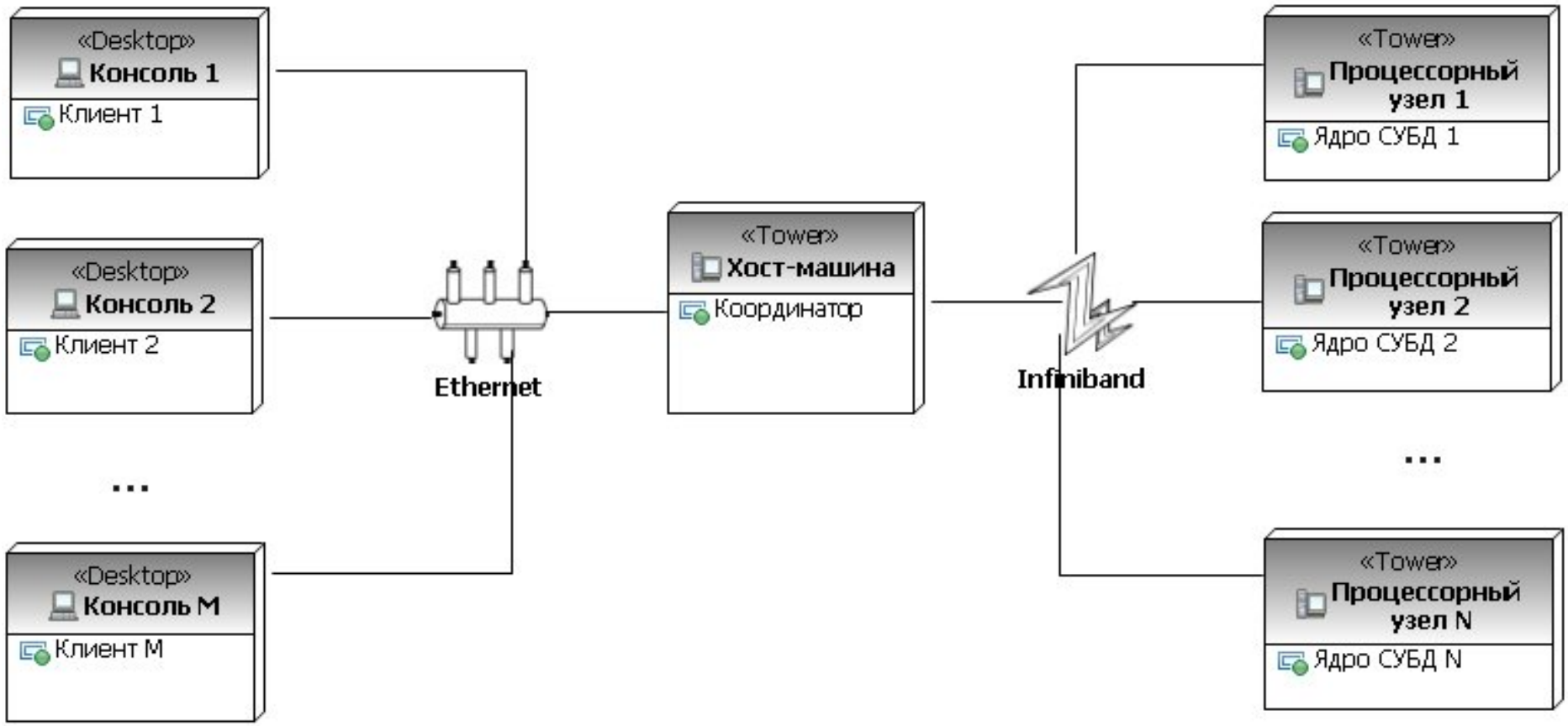
- Функция балансировки

$$g(Q) = a_i \operatorname{sgn}\left(\max_{1 \le i \le n}(q_i) - B\right) \lambda\, r\left(l(M)\right) \sum_{i=1}^{n} q_i$$

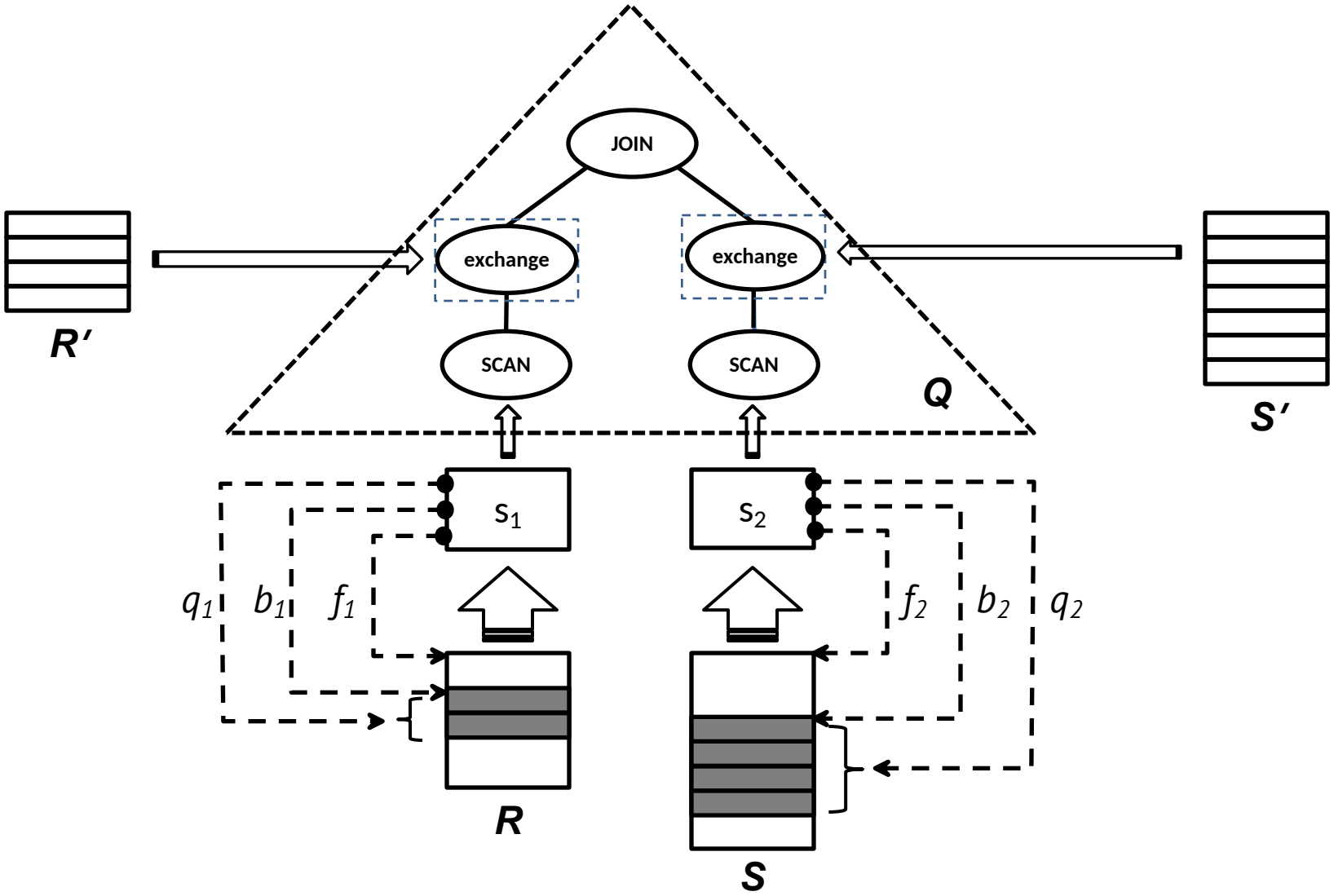
$$D(s_i) = \min\left(\lfloor q_i / 2 \rfloor, r\left(l(M)\right) S\left(f_i\right)\right)$$

λ - весовой коэффициент
B – минимальное количество сегментов, для балансировки
a_i – индикатор балансировки
q_i –количество сегментов в обрабатываемом отрезке
n – количество потоков параллельного агента
$S(f_i)$ – количество сегментов во фрагменте f_i.

Структура иерархической СУБД «Омега»



Параллельный агент



Параметры экспериментов

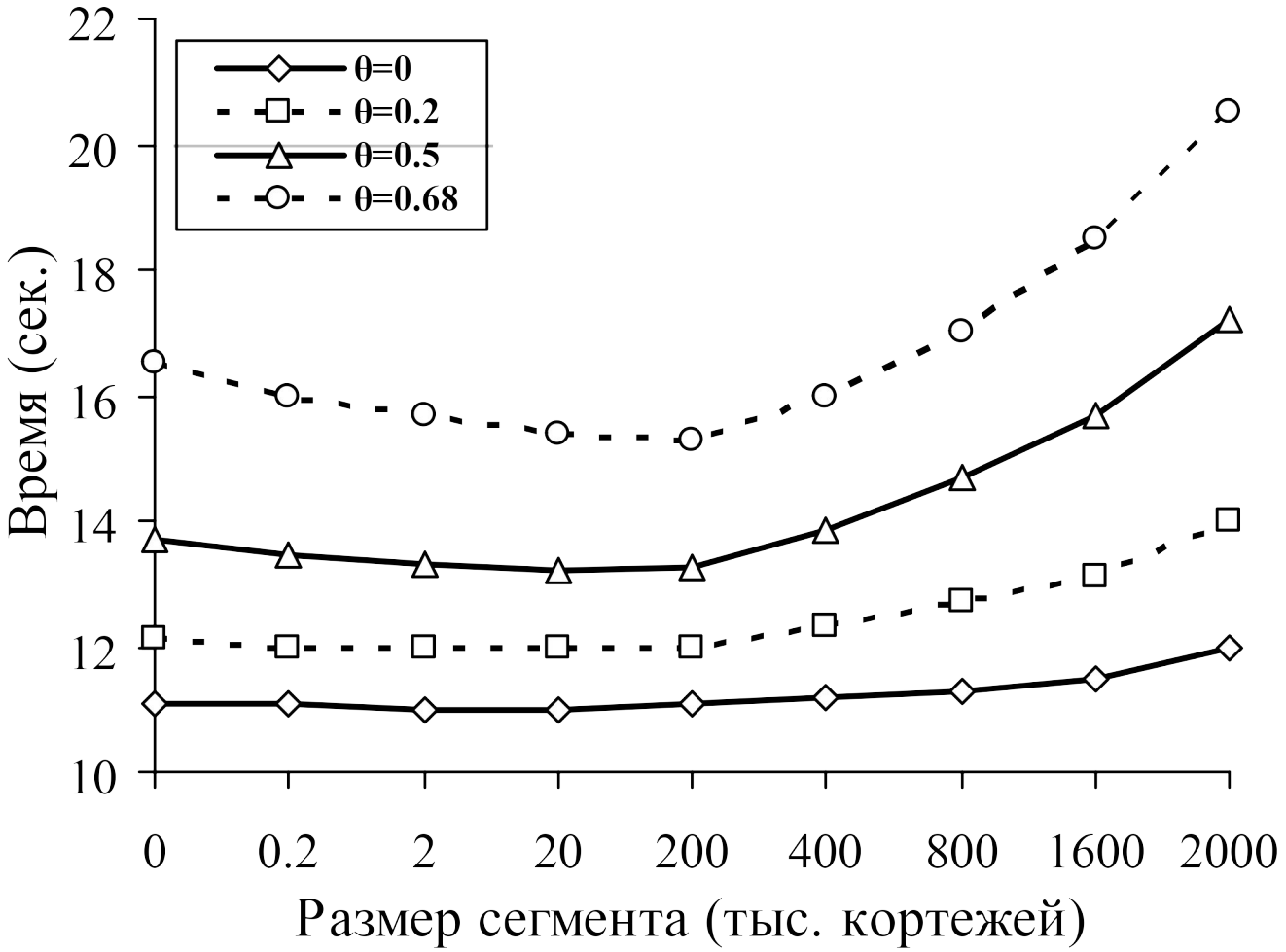
Параметры базы данных:

- Запрос $Q = R \not\bowtie S$
- R – опорное отношение (1 500 000 записей)
- S – тестируемое отношение (60 000 000 записей)
- Отношения R и S фрагментированы не по атрибуту соединения

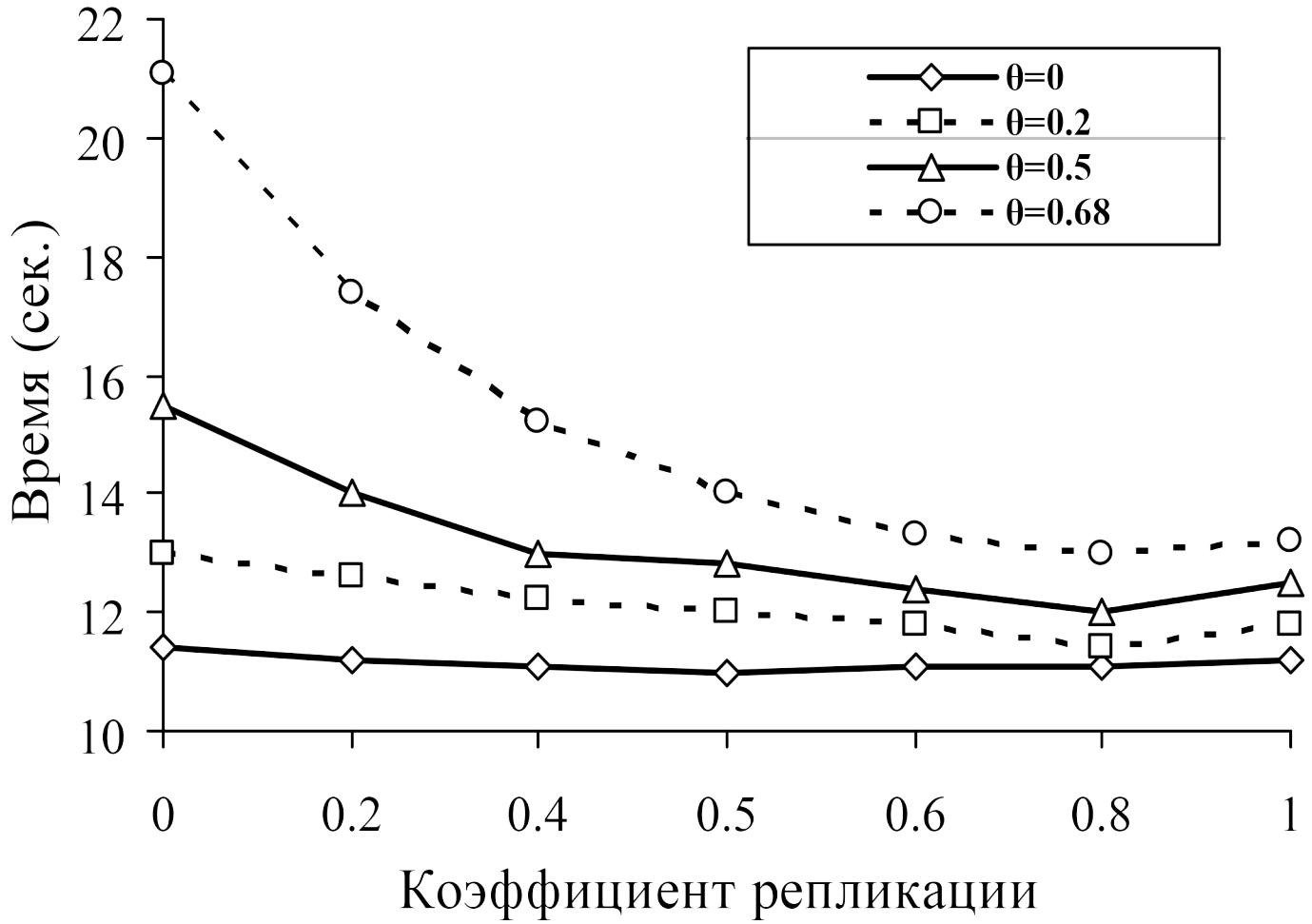
Параметры эксперимента:

- μ – коэффициент перекоса по значению атрибута соединения (в %)
- θ – коэффициент перекоса по значению атрибута фрагментации
- ρ – коэффициент репликации
- n – количество процессорных узлов

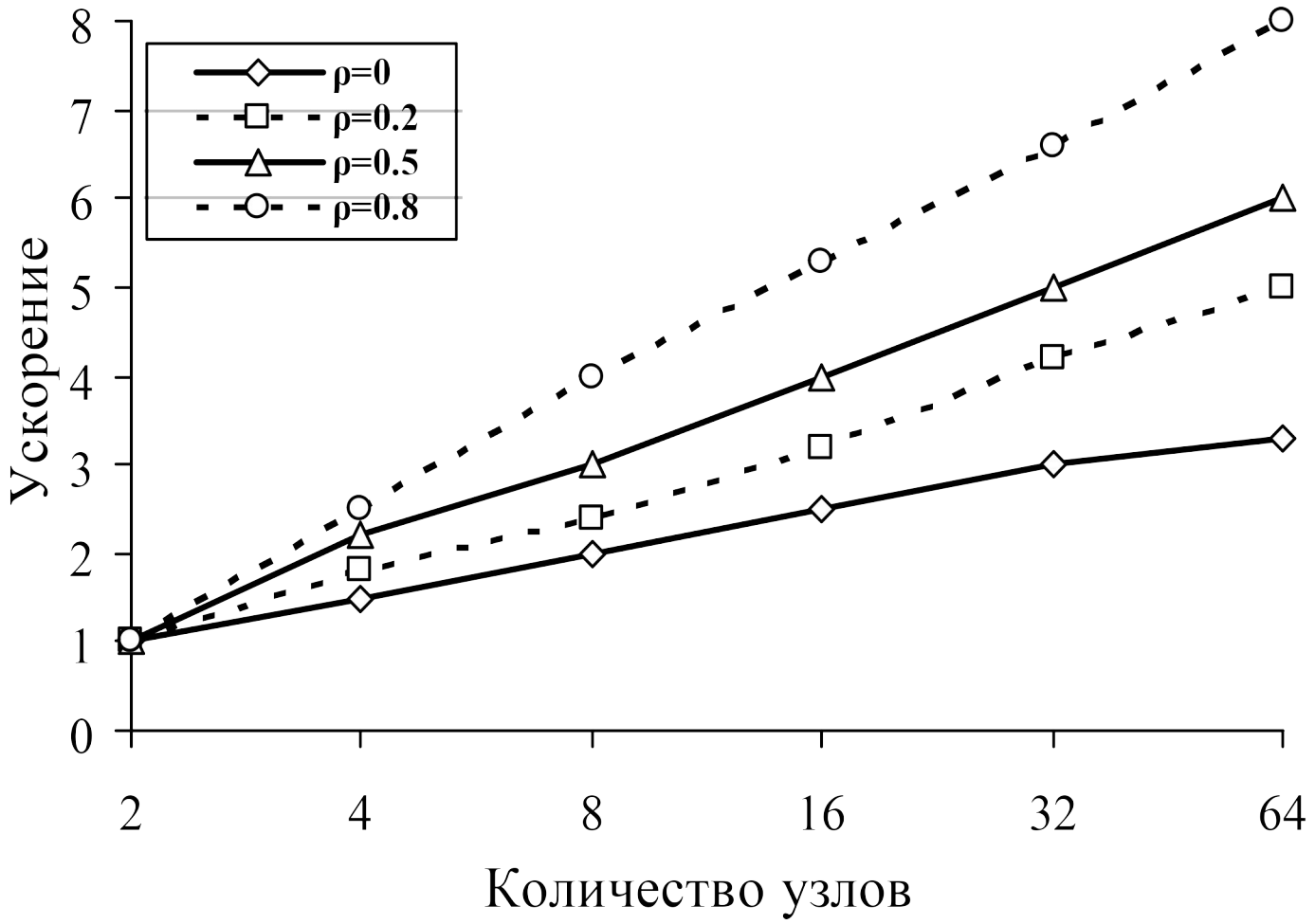
Исследование размера сегмента (*n=64 , μ=50%*)



Влияние коэффициента репликации на эффективность балансировки (*n=64 , μ=50%*)



Исследование масштабируемости метода балансировки загрузки (μ=50%, θ=0.5)



РОССИЙСКАЯ ФЕДЕРАЦИЯ

СВИДЕТЕЛЬСТВО

о государственной регистрации программы для ЭВМ

№ 2008614996

Параллельная СУБД «Омега» для кластерных систем

Правообладатель(ли): ***Государственное образовательное учреждение высшего профессионального образования «Южно-Уральский государственный университет», ГОУ ВПО «ЮУрГУ» (RU)***

Автор(ы): ***Лепихов Андрей Валерьевич, Соколинский Леонид Борисович, Цымблер Михаил Леонидович (RU)***

Заявка № **2008614484**

Дата поступления **3 октября 2008 г.**

Зарегистрировано в Реестре программ для ЭВМ

16 октября 2008 г.

Руководитель Федеральной службы по интеллектуальной собственности, патентам и товарным знакам

Б.П. Симонов

Публикации

1. *Лепихов А.В.*Технологии параллельных систем баз данных для иерархических многопроцессорных сред / *Лепихов А.В., Соколинский Л.Б., Костенецкий П.С.* // Автоматика и телемеханика. –2007. –Том 68, №5. -С. 847–859.
2. *Лепихов А.В.* Модель вариантов использования параллельной системы управления базами данных для грид // Вестник ЮУрГУ. Серия «Математическое моделирование и программирование». –Челябинск : ЮУрГУ, 2008 г. –№ 15 (115). –Вып. 1. –С. 42–53.
3. *Лепихов А.В.* Балансировка загрузки при выполнении операций соединения в параллельных СУБД для кластерных систем // Научный сервис в сети Интернет: решение больших задач. Труды Всероссийской научной конференции (22–27 сентября 2008 г., г. Новороссийск). –М.: Изд-во МГУ, 2008. –С. 292–295.
4. *Лепихов А.В.* Стратегия размещения данных в многопроцессорных системах с симметричной иерархической архитектурой / *А.В. Лепихов, Л.Б. Соколинский* // Научный сервис в сети Интернет: технологии параллельного программирования. Труды Всероссийской научной конференции (18–23 сентября 2006 г., г. Новороссийск). –М.: Изд-во МГУ, 2006. –С. 39-42.
5. *Lepikhov A.V.* Data Placement Strategy in Hierarchical Symmetrical Multiprocessor Systems / *A.V. Lepikhov, L.B. Sokolinsky* // Proceedings of Spring Young Researchers' Colloquium in Databases and Information Systems (SYRCoDIS'2006), June 1-2, 2006. -Moscow, Russia: Moscow State University. -2006. -С. 31-36.
6. *А.В. Лепихов* Свидетельство Роспатента об официальной регистрации программы для ЭВМ «Параллельная СУБД «Омега» для кластерных систем» / *А.В. Лепихов, Л.Б. Соколинский, М.Л. Цымблер*; -№2008614996 от 03.10.2008.

Основные результаты, выносимые на защиту

1. Построена математическая модель многопроцессорной иерархии. На основе этой модели разработан метод частичного зеркалирования, который может быть использован для динамической балансировки загрузки. Доказаны теоремы, позволяющие получить аналитическую оценку трудоемкости формирования и обновления реплик при использовании метода частичного зеркалирования.
2. Предложен метод параллельной обработки запросов в иерархических многопроцессорных системах, позволяющий осуществлять эффективную динамическую балансировку загрузки на базе техники частичного зеркалирования.
3. Разработан прототип иерархической СУБД «Омега», реализующий предложенные методы и алгоритмы. Проведены тестовые испытания СУБД «Омега» на вычислительных кластерах, входящих в грид-систему «СКИФ-Полигон», подтвердившие эффективность предложенных алгоритмов, методов и подходов.

СПАСИБО ЗА ВНИМАНИЕ!

Изоморфизм

DM-деревья A и B называются *изоморфными*, если существуют взаимно однозначное отображение f множества M (A) на множество M (B) и взаимно однозначное отображение g множества E (A) на множество E (B) такие, что:

1) узел v является конечным узлом дуги e в дереве A тогда и только тогда, когда узел f(v) является конечным узлом дуги g(e) в дереве B;

2) узел w является начальным узлом дуги e в дереве A тогда и только тогда, когда узел f(w) является начальным узлом дуги g(e) в дереве B;

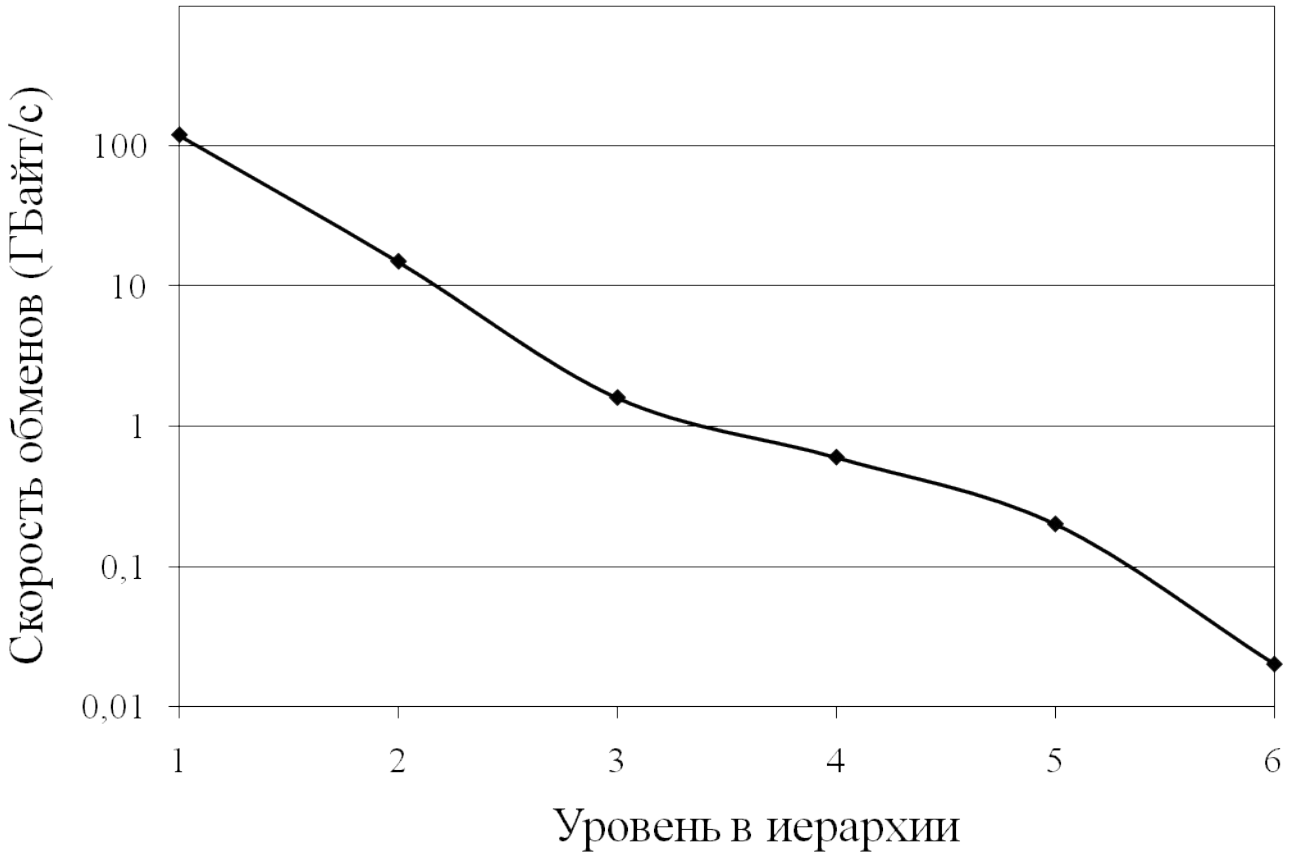
3) pÎ P (A) Û f(p)Î P (B);

4) dÎ D(A) Û f(d)Î D(B);

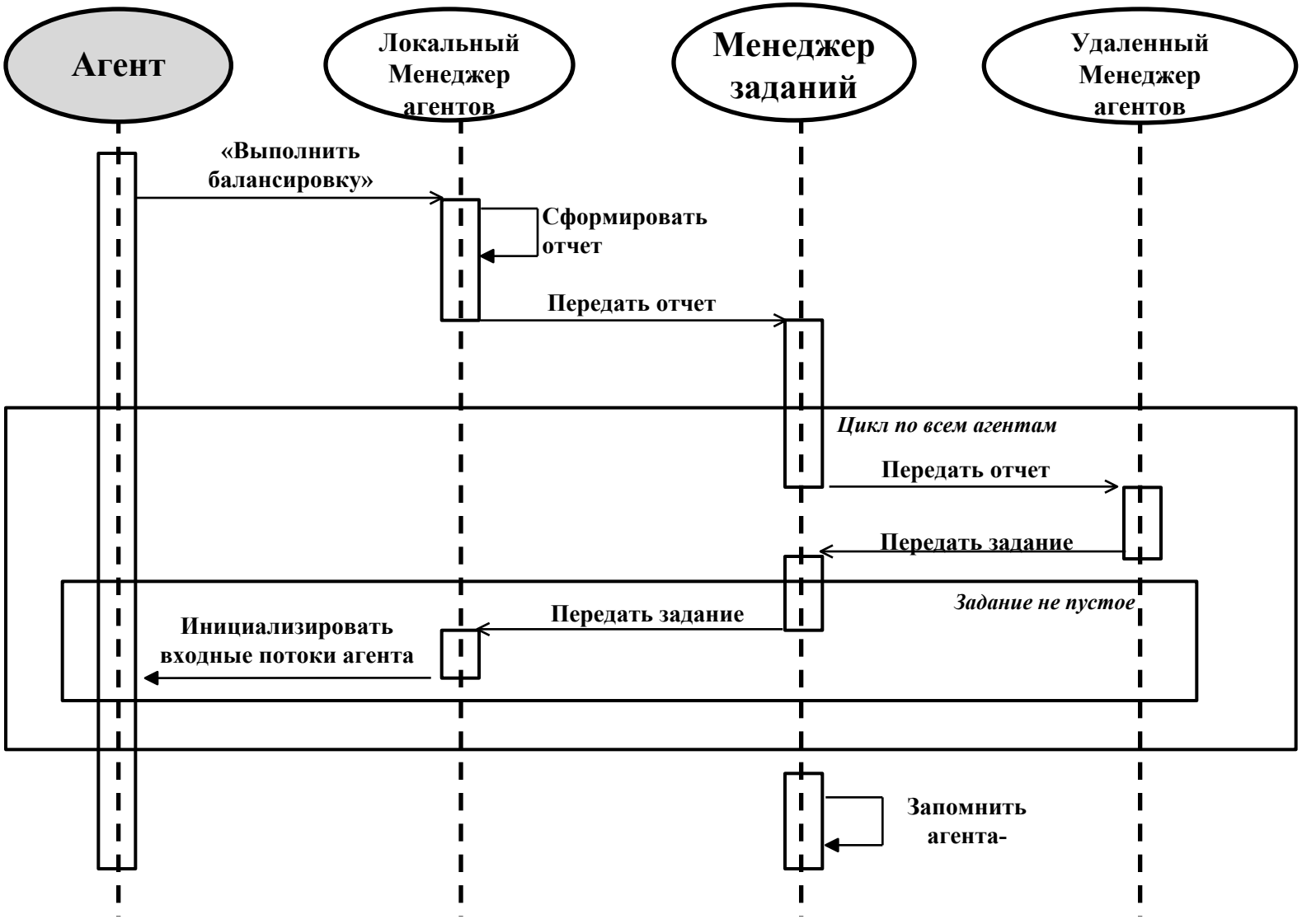
5) nÎ N(A) Û f(n)Î N(B);

6) h(f(v)) =h(v).

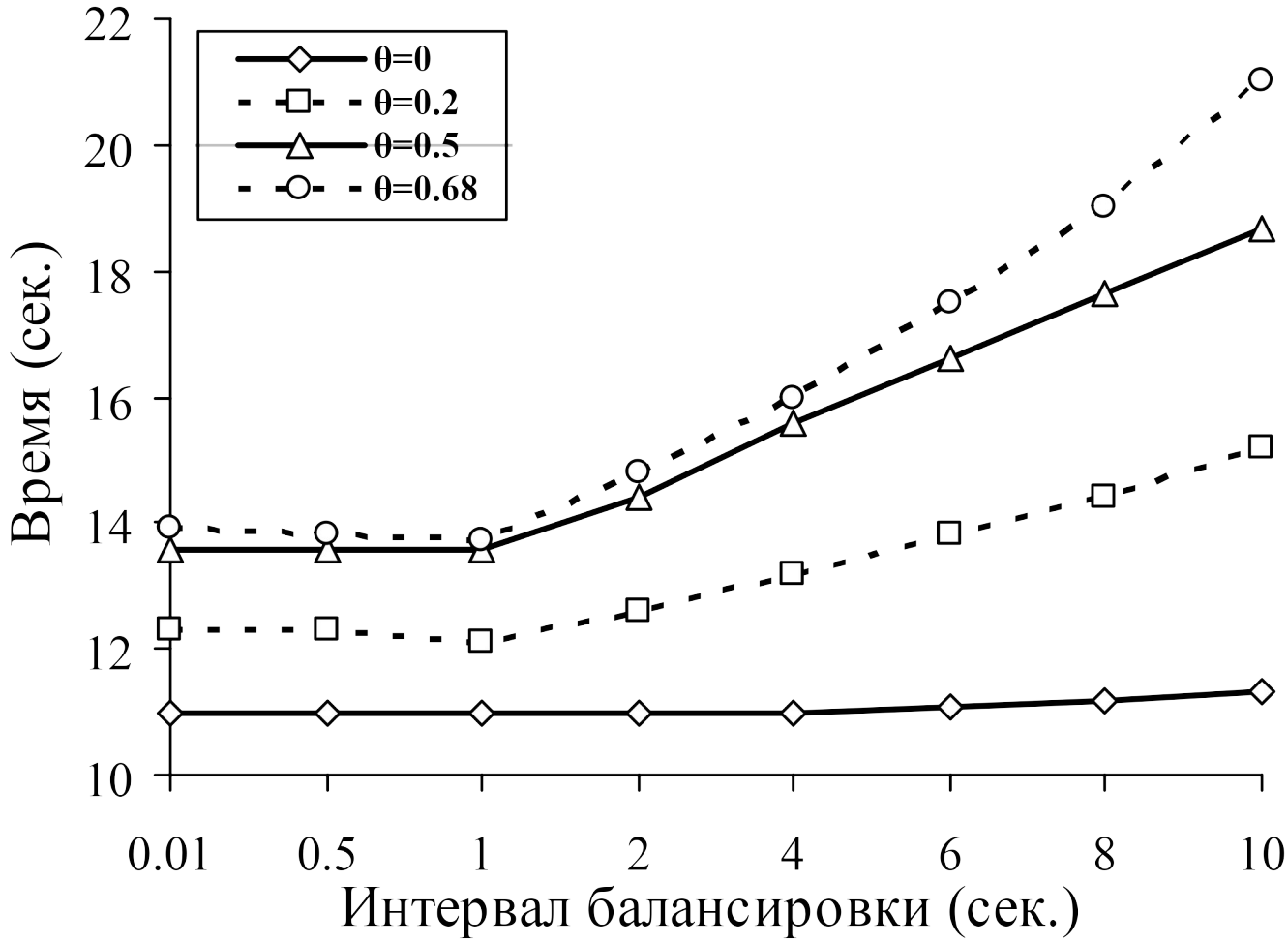
Скорость межпроцессорных коммуникаций иерархии



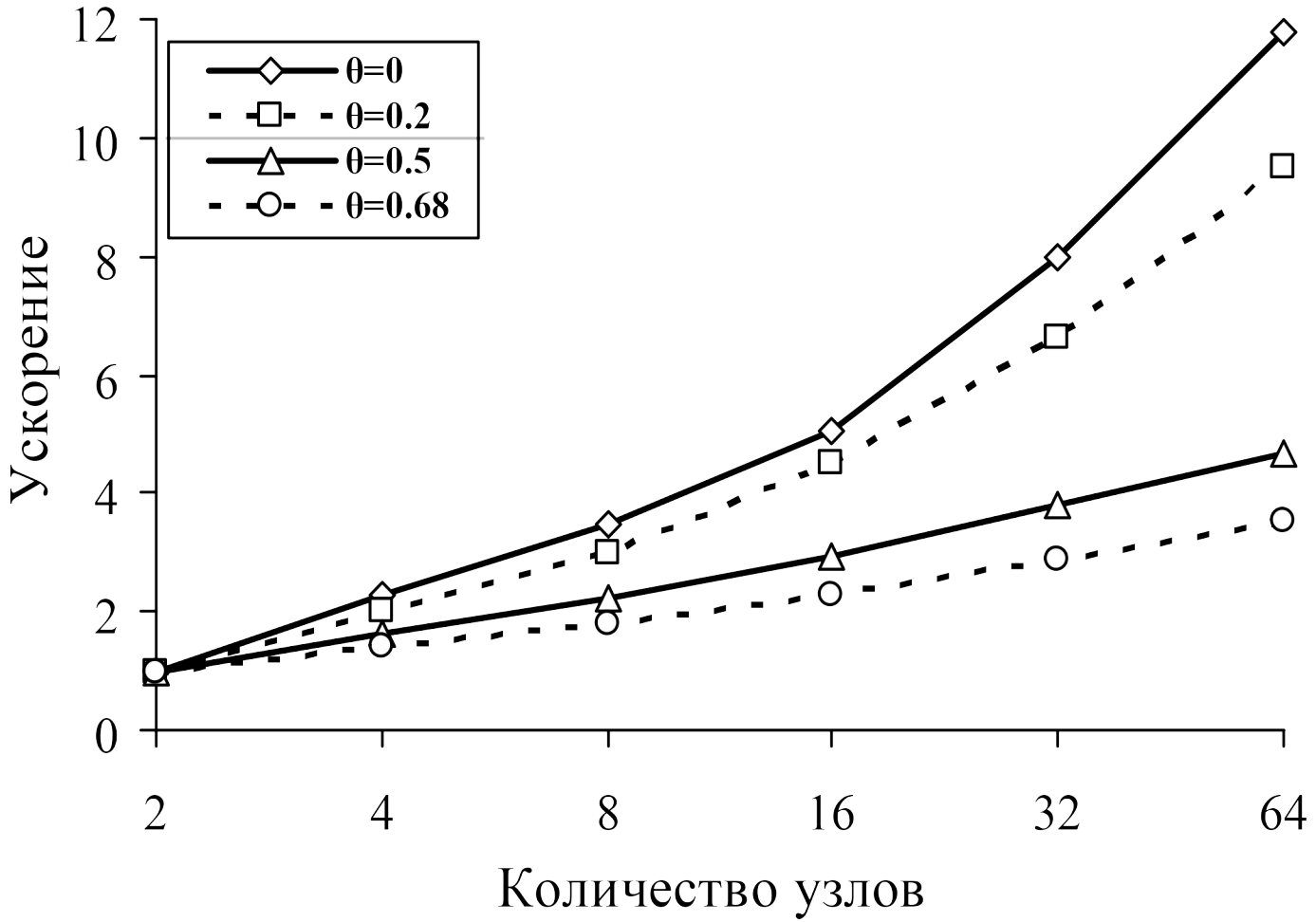
Механизм балансировки загрузки



Зависимость времени выполнения запроса от интервала балансировки (*n=64 , μ=50%)*



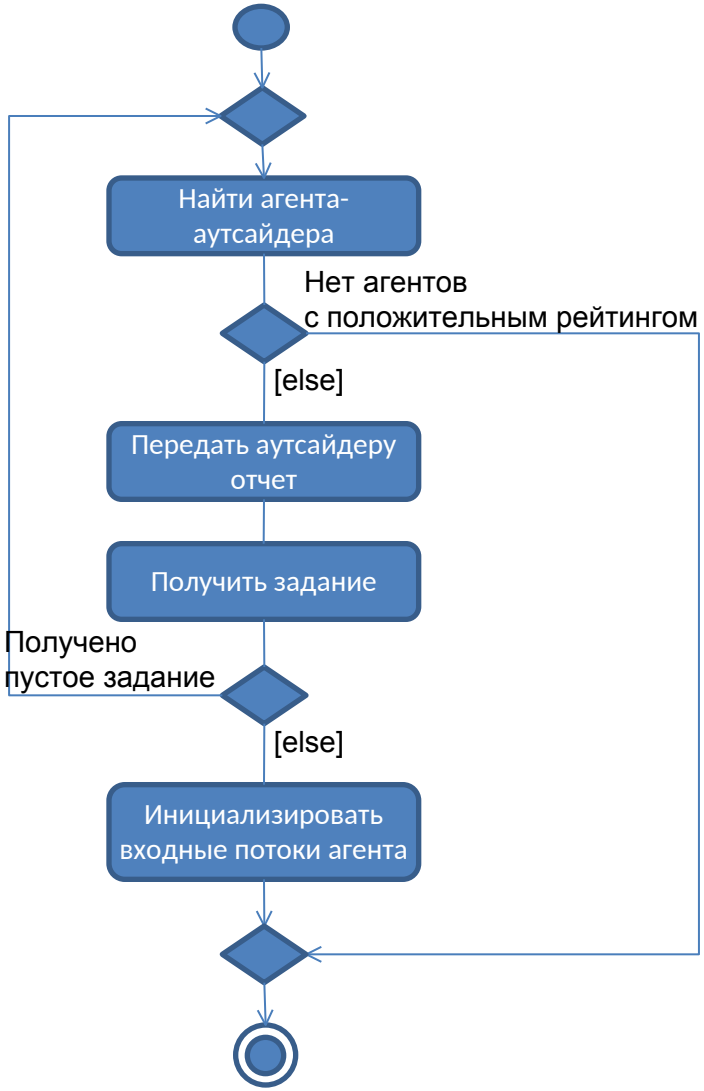
Влияние перекоса по данным на ускорение (μ=50%, ρ=0.50)



Определение

Иерархическая многопроцессорная система – это многопроцессорная система, в которой процессоры объединяются в единую систему с помощью соединительной сети, имеющей иерархическую структуру и обладающую свойствами однородности по горизонтали и неоднородности по вертикали

Стратегия выбора аутсайдера



Рейтинговая функция

$$g(Q) = a_i \operatorname{sgn}(\max_{1 \le i \le n}(q_i) - B) \cdot r(l(M)) \sum_{i=1}^{n} q_i$$

- λ - весовой коэффициент
- B – минимальное количество сегментов, для балансировки
- a_i – индикатор балансировки
- q_i –количество сегментов в обрабатываемом отрезке
- n – количество потоков параллельного агента

Иерархическая СУБД «Омега»

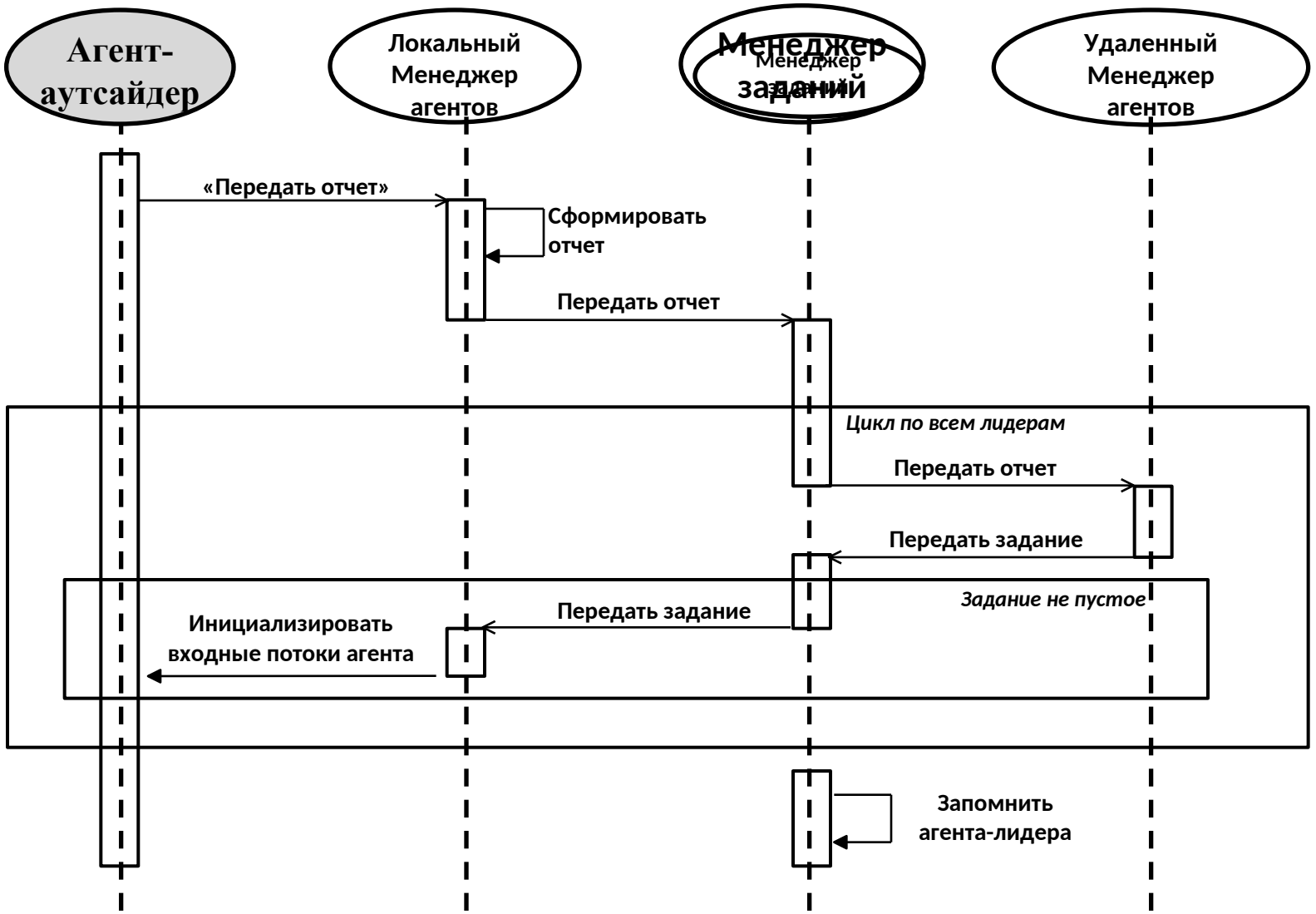
- Проектирование в среде Rational Software Architect 7.0.0.7
- Реализация прототипа иерархической СУБД «Омега» на языке Си с использованием пакета MPI.
- Вычислительные эксперименты

Функция балансировки

$$D(s_i) = \min\left(\left\lceil \frac{q_i}{2} \right\rceil, r(l(M)) \times S(f_i)\right)$$

- s_i – входной поток;
- $S(f_i)$ – количество сегментов во фрагменте f_i.

Механизм балансировки загрузки



Вычислительные эксперименты

Параметр	Значение
Число процессоров:	**64**
Тип процессора:	**Intel Xeon E5472 (4 ядра по 3.0 GHz)**
Оперативная память:	**8 ГБ/диск**
Дисковая память:	**120 ГБ/диск**
Тип системной сети:	**InfiniBand (20Gbit/s, макс. задержка 2 мкс)**
Тип управляющей (вспомогательной) сети:	**Gigabit Ethernet**
Операционная система:	**SUSE Linux Enterprise Server 10**

Влияние перекоса по атрибуту соединения на ускорение (*n=65, ρ=0.5*)