МОДЕЛИРОВАНИЕ И ОПТИМИЗАЦИЯ РАСПРЕДЕЛЕННЫХ БАЗ ДАННЫХ

Телятников А.О.
Донецкий национальный технический университет,
кафедра автоматизированных систем управления
E-mail: Alexander.Telyatnikov@gmail.com

Abstract

Telyatnikov A.O. Modelling and optimization of the distributed databases. *The results of modelling and optimization of the distributed database(DDB) for a large-scale concern on production of pastry goods are described in the article. The results of research of the DDB functioning descriptions are resulted, conducted by the DDB object model. Influence of hardware parameters of and data allocation on efficiency of the DDB work is explored. The DDB optimization with the use of genetic algorithms is executed.*

Введение

Современный этап развития компьютерных информационных систем можно охарактеризовать как переход от автоматизации отдельных задач к построению корпоративных информационных систем, составной частью которых являются распределенные базы данных (РБД). РБД представляет собой сложную динамическую систему, в которой выполняется множество запросов к распределенным данным и производится обновление множества копий, размещенных на разных узлах компьютерной сети. Скорость выполнения запросов и распространения обновлений определяется параметрами технических средств РБД и размещением данных на узлах сети.

Для проведения анализа функционирования РБД, оценки влияния параметров технических средств и распределения данных по узлам сети на производительность системы, разработана объектная модель РБД [1]. Данная модель построена на основе объектных моделей ее типовых компонентов: узел, канал передачи данных, приложение, запрос, таблица. В качестве критерия эффективности работы РБД предложено использовать суммарное среднее время выполнения запросов и распространения обновлений [2]. Для оптимизации распределения данных по узлам сети разработана новая модификация генетического алгоритма, которая используется совместно с объектной моделью РБД [3]. В качестве критерия эффективности РБД предложено использовать суммарное среднее время выполнения запросов и распространения обновлений, порожденных функционированием системы в течении заданного интервала времени, которое определяется следующим выражением:

$$T = \frac{1}{N_q}\sum_{s=1}^{N_q} t'_s + \frac{1}{N_u}\sum_{e=1}^{N_u} t''_e . \qquad (1)$$

где N_q — количество запросов в системе; N_u — количество обновлений в системе; t'_s — время выполнения s-го запроса, $s \in \left[1, N_q\right]$; t''_e — время распространения e-го обновления данных, $e \in \left[1, N_u\right]$.

Описание объекта экспериментальных исследований

Для проведения экспериментальных исследований с использованием разработанной модели и алгоритма оптимизации в качестве объекта выбрана компьютерная информационная система ЗАО ПО “Киев-Конти”. Данное предприятие является крупным производителем кондитерской продукции, входит в тройку лидеров отечественного кондитерского рынка и занимает первое место по темпам роста объёмов производства (по

информации сайта http://www.kiev-konti.com). В состав компании “Киев-Конти” входят 4 фабрики: — три в Украине (Донецкая, Константиновская, Горловская) и одна в России (Курская). Компания имеет распределенную систему сбыта, состоящую из пяти филиалов (складов продукции), из них четыре в Украине — в г. Донецк, Киев, Львов, Николаев и один филиал в России в г. Воронеж, а также несколько региональных представительств.

Из-за территориальной распределенности структуры компании “Киев-Конти” ее компьютерная информационная система также имеет распределенную архитектуру, одним из элементов которой является РБД, в состав которой входит 10 узлов: центральный узел (корпоративный сервер), по одному узлу на каждой фабрике и в каждом филиале. Данная распределенная база данных имеет размер и характеристики, позволяющие использовать ее в качестве объекта экспериментальных исследований с помощью разработанных объектной модели РБД и оптимизационного алгоритма.

Структура сети передачи данных компьютерной информационной системы ЗАО ПО “Киев-Конти” показана на рис. 1.

Корпоративный сервер, расположенный в центральном офисе компании в г. Донецке, подключен к центральному коммутатору с помощью высокоскоростного соединения (1 Гбит/сек.). К этому же концентратору подключено сетевое оборудование, обеспечивающее соединения с узлами филиалов и фабрик.

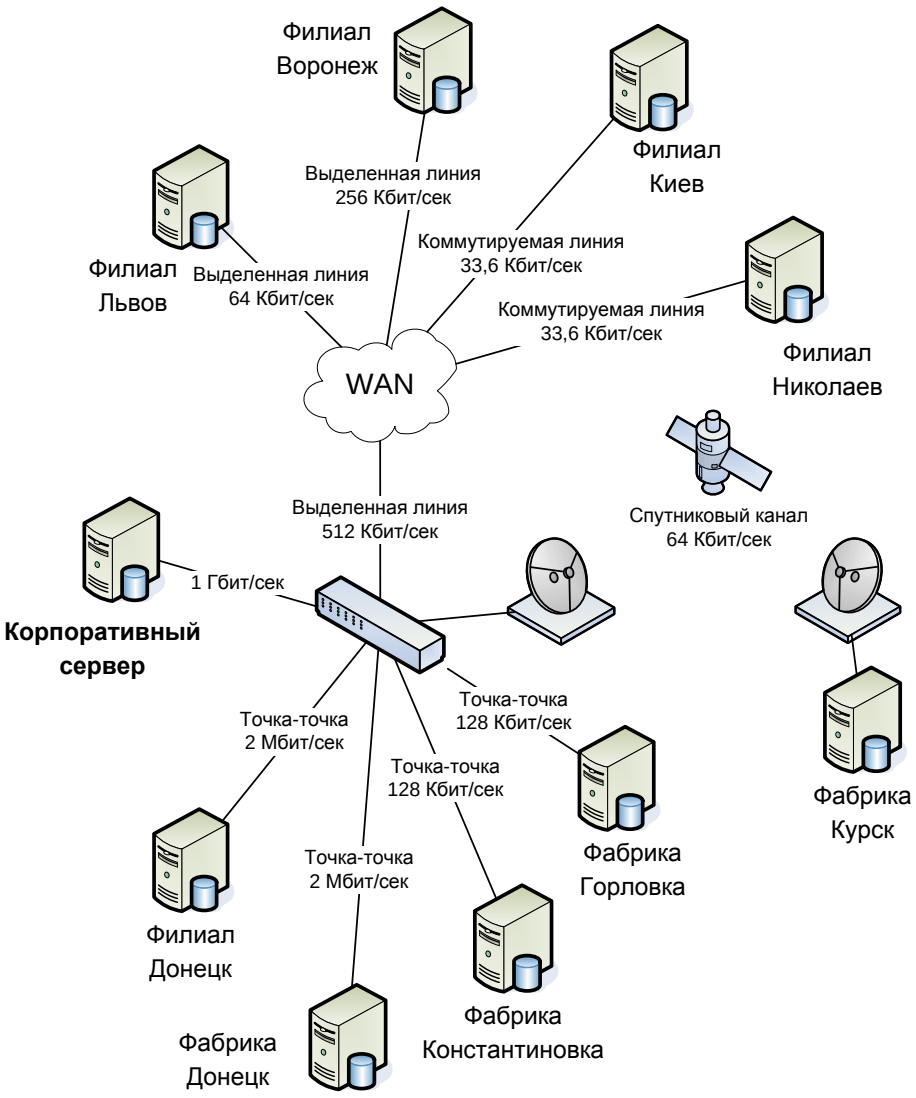

Рисунок 1 — Структура информационной сети ПО “Киев-Конти”

В РБД компьютерной информационной системы компании “Киев-Конти” хранятся следующие данные:

1) информация о выработке продукции на фабриках;
2) информация о заявках на продукцию от филиалов;
3) информация о приходе/расходе продукции по филиалам;
4) рецептура изготовления продукции.

Исследование на модели характеристик функционирования РБД

С помощью разработанной модели были проведены вычислительные эксперименты с целью анализа функционирования РБД и выявления “узких мест” системы [4]. Среднее время выполнения запросов и распространения обновлений в системе, рассчитанное с помощью модели, составило 111,77 с.

Для проверки адекватности модели РБД проводилось сравнение длительностей выполнения запросов и распространения обновлений, полученных с ее помощью, с реальными показателями, полученными с использованием утилиты SQL Profiler. Расхождение результатов моделирования с реальными данными составило менее 10%.

На основании результатов моделирования был проведен анализ трех показателей: длительность выполнения запросов и распространения обновлений, загруженность каналов передачи данных, загруженность узлов обработки данных.

При вычислении оценок длительностей выполнения запросов (обновлений) определялись средние значения времени их выполнения в различных разрезах (этапы выполнения, приложения и др.) с учетом удельного веса этих запросов (обновлений) по отношению к общему количеству запросов (обновлений).

На диаграмме (рис. 2) представлено среднее время выполнения запросов (объектов). Анализ этих данных показал, что наибольшее время выполнения имеют запросы, инициируемые на узлах “Филиал. Киев” и “Филиал. Николаев” или запросы, которые обращаются к таблицам, хранящимся на этих узлах. Это связано с тем, что при передаче данных используются медленные коммутируемые каналы, со скоростью передачи 33,6 Кбит/с, а также — с нерациональным распределением данных по узлам системы.

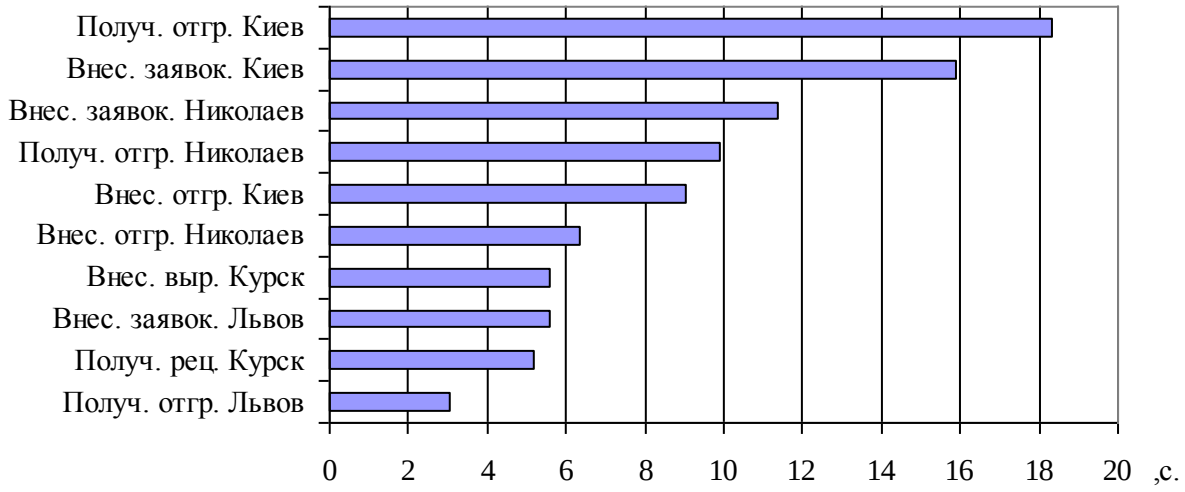

Рисунок 2 — Среднее время выполнения запросов (обновлений)

Результаты проведения анализа длительности выполнения запросов и распространения обновлений позволяют выявить приложения и конкретные запросы, которые выполняются наиболее долго. Оптимизация этих запросов обеспечивает повышение эффективности функционирования РБД в целом. Кроме этого, для анализа представляет

интерес загруженность каналов передачи данных, зная которую можно определить какие из каналов являются “узкими местами” системы.

На диаграмме (рис. 3) приведены коэффициенты загрузки каналов передачи данных. Наибольший коэффициент загрузки имеют каналы, соединяющие Киевский и Николаевский филиалы с Центральным узлом. Это объясняется тем, что по ним передаются достаточно большие объемы данных, но при этом они являются коммутируемыми и имеют маленькую пропускную способность 33,6 Кбит/с.

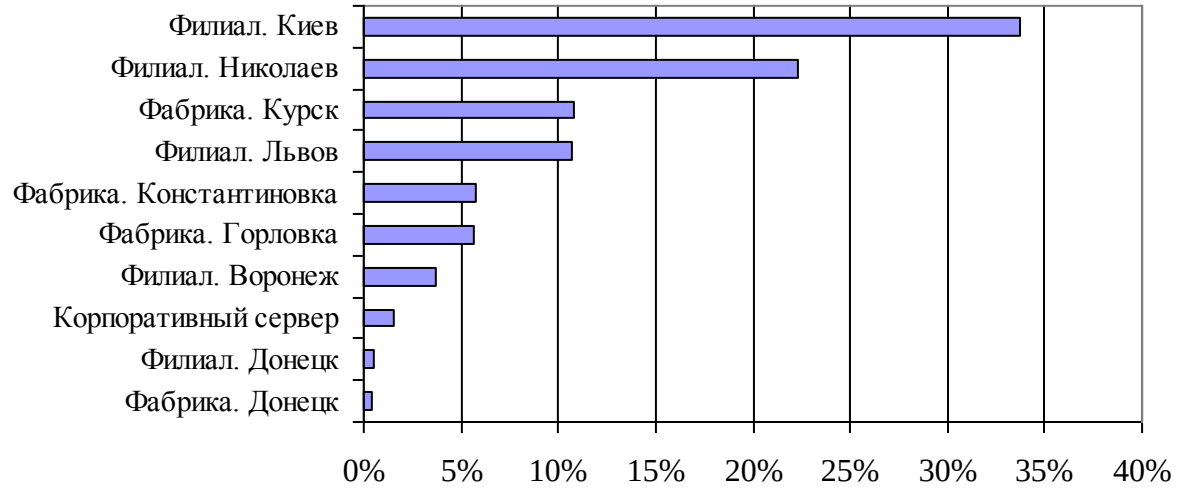

Рисунок 3 — Коэффициенты загрузки каналов передачи данных

Несмотря на то, что обработка запросов узлами РБД имеет небольшую долю в общем времени выполнения запросов, информация о том, на сколько загружены узлы обработки, может представлять интерес при анализе.

С использованием разработанной модели был проведен анализ коэффициентов загрузки узлов обработки запросов (обновлений). Анализ показал, что загрузка узлов моделируемой РБД незначительна, поэтому увеличение их производительности не приведет к существенному улучшению критерия эффективности.

На основании анализа результатов вычислительных экспериментов, проведенных с помощью разработанной модели РБД, были выявлены так называемые “узкие места” системы. Ими являются каналы передачи данных Киевского и Николаевского филиалов. Повышение пропускной способности указанных каналов позволит повысить эффективность функционирования системы в целом. Также были выявлены запросы и обновления с наибольшей длительностью выполнения — это запросы и обновления, инициируемые на узлах Киевского и Николаевского филиалов. Причиной их большой длительности выполнения является:

а) низкая пропускная способность каналов передачи данных;

б) нерациональное размещение данных по узлам компьютерной сети.

Исследование влияния параметров технических средств и размещения данных на эффективность работы РБД

Моделирование РБД применялось для исследования влияния изменения пропускной способности каналов передачи данных и перераспределения фрагментов данных по узлам компьютерной сети.

С помощью разработанной модели проведен анализ влияния увеличения пропускной способности каналов передачи данных Киевского и Николаевского филиалов до 64 Кбит/с. Среднее время выполнения запросов и распространения обновлений при этом изменилось на 15,12% и составило 94,87 с.

В табл. 1 представлено среднее время выполнения отдельных запросов (обновлений) до изменения пропускной способности каналов передачи данных и после ее увеличения (время выполнения остальных запросов (обновлений) не изменилось).

Таблица 1 — Среднее время выполнения запросов (обновлений)

Запрос	Среднее время выполнения, с.	
	До изменения	После изменения
Получение отгрузки. Николаев	9,90	6,37
Получение отгрузки. Киев	18,35	13,32
Внесение отгрузки. Николаев	6,35	4,35
Внесение отгрузки. Киев	9,05	7,48
Внесение заявок на отгрузку. Николаев	11,36	9,35
Внесение заявок на отгрузку. Киев	15,88	13,11

Увеличение пропускной способности каналов передачи данных филиалов в г. Киев и г. Николаев привело к уменьшению времени выполнения запросов (обновлений), инициируемых на узлах этих филиалов, а, также осуществляющих доступ к данным, хранящимся на них. Среднее время выполнения этих запросов в среднем уменьшилось на 24,51%.

В табл. 2 представлены коэффициенты загрузки каналов передачи данных, значения которых изменилось после увеличения пропускной способности каналов Киевского и Николаевского филиалов.

Таблица 2 — Коэффициенты загрузки каналов передачи данных

Канал передачи данных	Коэффициент загрузки, %	
	До изменения	После изменения
Корпоративный сервер	1,55	1,54
Филиал. Киев	33,73	19,77
Филиал. Николаев	22,32	12,68

Коэффициент загрузки канала передачи данных филиала г. Киев уменьшился на 41,39%, филиала г. Николаев — на 43,19%.

Рациональное перераспределение таблиц и фрагментов данных к которым обращаются указанные запросы (обновления) потенциально приводит к улучшению работы РБД, в смысле уменьшения времени отклика системы. Рациональное перераспределение данных означает их размещение с учетом спроса на них со стороны запросов и обновлений. Данные должны быть приближены к местам их наиболее интенсивного использования. Моделирование РБД показало, что после удаления копий фрагментов данных о приходе-расходе из узлов филиалов г. Киева и г. Николаева, когда эта информация остается только на корпоративном сервере, среднее время выполнения запросов и распространения обновлений уменьшилось еще на 18,19% и составило 77,61 с.

На рис. 4 изображены значения критерия эффективности работы РБД (T), полученные с помощью модели.

При первоначальном варианте параметров технических средств РБД и распределения данных, значение критерия эффективности $T = 111,77$ с, после увеличения пропускной способности каналов киевского и николаевского филиалов до 64 Кбит/с $T = 94,87$ с, что на 15,12% меньше первоначального. После перераспределения данных значение критерия эффективности уменьшилось еще на 18,19% и составило $T = 77,61$ с.

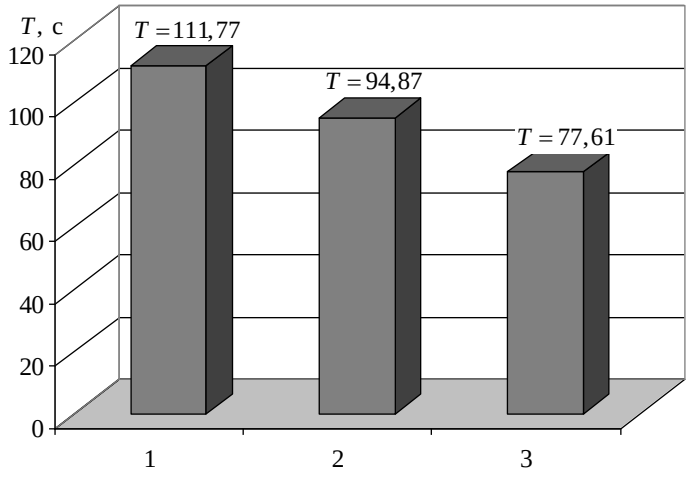

Рисунок 4 — Значения критерия эффективности при различных параметрах РБД
1 — первоначальный вариант; 2 — после увеличения пропускной способности каналов передачи данных; 3 — после перераспределения данных.

Таким образом, можно сделать вывод, что разработанная модель может быть использована для исследования влияния изменения параметров технических средств и размещения данных на эффективность работы РБД.

Оптимизация РБД с использованием генетических алгоритмов

Эффективность работы разработанного оптимизационного алгоритма проверена путем сравнения полученных с его помощью субоптимальных решений, с оптимумом, полученным полным перебором при ограничении пространства поиска с учетом особенностей рассматриваемой РБД.

Процедура полного перебора заключается в последовательном переборе всего пространства поиска задачи. Количество возможных решений определяется выражением:

$$M = \left(2^m - 1\right)^n. \qquad (2)$$

где m — число узлов РБД, n — число фрагментов данных, которые необходимо распределить по узлам сети.

Для выбранного объекта оптимизации полное пространство поиска содержит $M = \left(2^{10} - 1\right)^{18} \approx 1{,}5 \cdot 10^{54}$ вариантов решений. Такое большое количество вариантов невозможно просчитать с помощью модели за приемлемое время.

Уменьшить пространство поиска можно путем исключения из рассмотрения заведомо неоптимальных вариантов. Очевидно, что если приложения, функционирующие на определенном узле РБД, не инициируют запросов, осуществляющих доступ к некоторым фрагментам данных, то схемы данных, в которых данный фрагмент находится на этом узле, не являются оптимальными. Тогда логично будет ограничить пространство поиска вариантами, в которых фрагменты данных размещаются на узлах, из которых осуществляется доступ к ним. Количество вариантов P_i размещения i-го фрагмента данных будет определяться выражением:

$$P_i = 2^{m_i} - 1 \qquad (3)$$

где m_i — количество узлов, из которых осуществляется доступ к i-му фрагменту данных.

При этом размер пространства поиска оптимизационной задачи будет определяться следующим выражением:

$$P = \prod_{i=1}^{n} \left(2^{m_i} - 1\right) \quad (4)$$

где n — количество распределяемых фрагментов данных.

Для рассматриваемого объекта оптимизации $P = 387\,420\,489$. Такое количества вариантов можно просчитать за приемлемый промежуток времени (несколько суток).

С помощью процедуры полного перебора было получено минимальное значение критерия эффективности (1), которое составило 79,1 с. Время поиска оптимального решения с использованием процедуры полного перебора на ПЭВМ с процессором Intel Celeron 2,8 GHz составило примерно 17 суток.

Так как разработанный алгоритм оптимизации РБД, построенный на основе ГА, представляет собой случайный направленный поиск, то решения, получаемые с его помощью, являются субоптимальными. Поэтому необходимо проанализировать величину отклонения получаемых решений от глобального оптимума, которая зависит от следующих параметров:

1) размер популяции — количество точек пространства поиска, просматриваемых за одну итерацию алгоритма;
2) количество поколений — число итераций работы оптимизационного алгоритма;
3) вероятности применения операторов генетических алгоритмов (отбора, скрещивания, мутации, рекомбинации) — определяют сбалансированность процессов отбора и изменчивости.

Для проверки эффективности разработанного алгоритма оптимизации РБД, с использованием объектной модели и модифицированного генетического алгоритма, проведен ряд вычислительных экспериментов и выполнена статистическая обработка полученных результатов.

Проведено исследование влияния величины размера популяции и количества поколений на значение критерия эффективности работы РБД (рис. 5). Анализ полученных зависимостей показал, что целесообразно в качестве значений, обеспечивающих наибольшее приближение к оптимуму, принять размер популяции $N_p = 60$ и количество поколений $G = 20$.

При зафиксированных значениях параметров $N_p = 60$ и $G = 20$, был проведен поиск рациональных значений вероятностей операторов рекомбинации $P_{рек}$ и скрещивания $P_{скрещ}$. Зависимость критерия эффективности РБД (T) от вероятностей применения оператора рекомбинации $P_{рек}$ и скрещивания $P_{скрещ}$ приведена на рис. 6. На графике видно, что значения, при которых наблюдается наиболее близкое приближение получаемых субоптимальных решений к глобальному минимуму $T_{опт}$ составляют: $P_{рек} = 0,5$ и $P_{скрещ} = 0,6$.

Для определения рационального значения вероятности применения оператора мутации $P_{мут}$ было проанализировано ее влияние на критерий эффективности РБД T совместно с вероятностью применения оператора рекомбинации $P_{рек}$ (рис. 7) и скрещивания $P_{скрещ}$ (рис. 8).

Для установленных ранее значений $N_p = 60$, $G = 20$ и $P_{скрещ} = 0,6$ получены зависимости критерия эффективности РБД T от вероятностей рекомбинации $P_{рек}$ и мутации $P_{мут}$. На графике (рис. 7) видно, что наилучшее значение критерия T достигается при вероятности мутации $P_{мут} = 0,07$ и вероятности рекомбинации $P_{рек} = 0,5 \ldots 0,55$.

Был также проведен анализ влияния значения вероятностей мутации $P_{мут}$ и скрещивания $P_{скрещ}$ на работу оптимизационного алгоритма, при вероятности рекомбинации $P_{рек} = 0,5$. На графике (рис. 8) видно, что наибольшее приближение к глобальному оптимуму имеют решения оптимизационного алгоритма при $P_{мут} = 0,07$ и $P_{скрещ} = 0,6$.

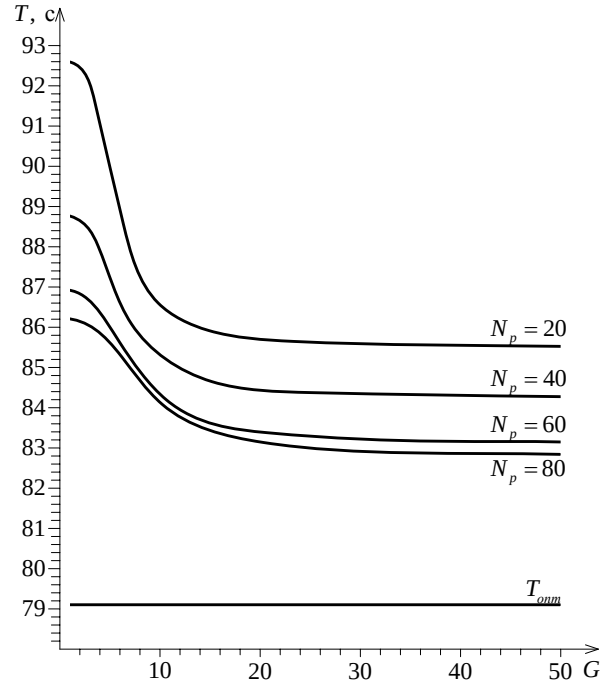

Рисунок 5 — Зависимость критерия эффективности РБД от размера популяции N_p и количества поколений G

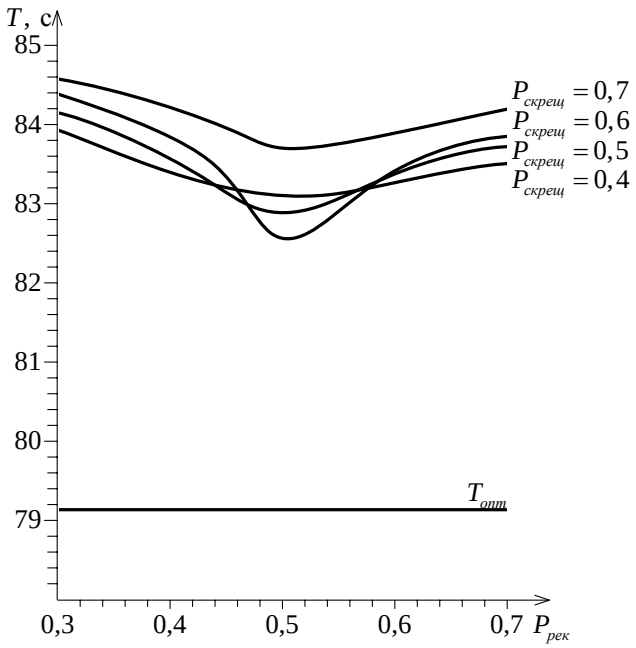

Рисунок 6 — Зависимость критерия эффективности РБД от вероятностей применения оператора рекомбинации $P_{рек}$ и скрещивания $P_{скрещ}$

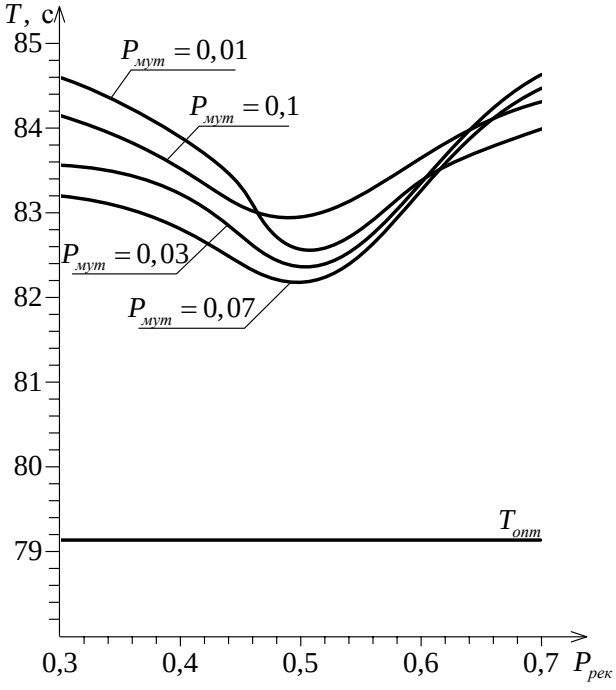

Рисунок 7 — Зависимость критерия эффективности РБД от вероятностей применения операторов мутации $P_{мут}$ и рекомбинации $P_{рек}$

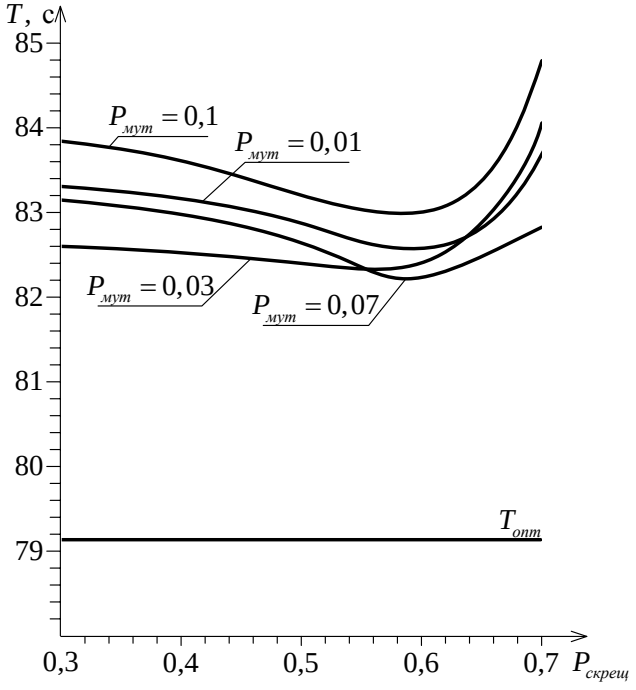

Рисунок 8 — Зависимость критерия эффективности РБД от вероятностей применения операторов мутации $P_{мут}$ и скрещивания $P_{скрещ}$

Таким образом, для разработанного алгоритма оптимизации рациональными значениями параметров, с точки зрения наилучшего приближения получаемых решений к глобальному минимуму, являются: $G = 20$; $N_p = 60$; $P_{рек} = 0,5$; $P_{скрещ} = 0,6$; $P_{мут} = 0,07$.

Наилучшее субоптимальное значение полученных критериев составляет $T = 82,19$ с. Абсолютное отклонение данной величины от глобального минимума $T_{опт}$ составляет 3,09 с, относительное отклонение — 3,76%. Уменьшение суммарного среднего времени выполнения запросов и распространения обновлений от первоначального варианта составило 29,58 с или 26,47% .

Время поиска субоптимального решения с использованием ГА на ПЭВМ с процессором Intel Celeron 2,8 GHz составило примерно 1 мин.

Заключение

В результате моделирования работы распределенной базы данных ЗАО ПО "Киев-Конти" установлено, что "узкими местами", снижающими общую производительность системы являются каналы передачи данных в г. Киев и Николаев и нерациональное размещение данных, которые обуславливают большую длительность выполнения запросов и обновлений на этих узлах. Для устранения этих недостатков предложено: повысить пропускную способность этих каналов с 33,6 до 64 Кбит/с, что позволяет уменьшить суммарное среднее время выполнения запросов и распространения обновлений на 15,1%; перераспределить данные в системе, что позволяет повысить эффективность РБД еще на 18,2%. Таким образом, за счет изменения параметров РБД суммарное среднее время выполнения запросов и обновлений уменьшилось с 111,77 с до 77,16 с, или на 30,56%.

Для модифицированного генетического алгоритма определены значения его параметров: размер популяции — 60, количество поколений — 20, вероятности рекомбинации — 0,5, скрещивания — 0,6, мутации — 0,07, при которых обеспечивается определение близкого к оптимальному значение критерия эффективности РБД. При этом отклонение субоптимального значения от оптимума, полученного путем полного перебора, составляет 3,76%, а суммарное среднее время выполнения запросов и распространения обновлений уменьшается с 111,77 с до 82,19 с, или на 26,47% без дополнительных материальных затрат, за счет оптимизации размещения данных.

Литература

1. Телятников А.О. Разработка объектной модели распределенной базы данных // Наукові праці ДонНТУ. Випуск 74. — Донецьк: ДонНТУ, 2004. — С. 192–200.
2. Лаздынь С.В., Телятников А.О. Оптимизация распределенных баз данных с использованием генетических алгоритмов // Вестник ХГТУ. — Херсон: ХГТУ, 2004. — № 1(19). — С. 236–239.
3. Лаздынь С.В., Телятников А.О. Повышение эффективности распределенных баз данных с использованием объектно-ориентированного моделирования и генетических алгоритмов // Единое информационное пространство: Сб. докл. Междунар. научно-практич. конф. — Днепропетровск: ИПК ИнКомЦентра УГХТУ, 2003. — С. 23–26.
4. Телятников А.О. Моделирование и анализ работы распределенной базы данных с использованием объектно-ориентированного подхода // Наукові праці ДонНТУ. Випуск 90. — Донецьк: ДонНТУ, 2005. — С. 91–98.
5. Курейчик В.М. Генетические алгоритмы: Монография. — Таганрог: Изд-во ТРТУ, 1998. — 242 с.