УДК 004.652.5

БУЗОВСКИЙ О.В.,
ПОДРУБАЙЛО А.А.

МЕТОДЫ И АЛГОРИТМЫ ОБЪЕДИНЕНИЯ ТАБЛИЦ ДЛЯ РАСПРЕДЕЛЕННЫХ ХРАНИЛИЩ ДАННЫХ В ОПЕРАТИВНОЙ ПАМЯТИ

В данной статье приведен обзор существующих методов распределенного объединения таблиц. Предложен метод и алгоритм модифицированного путевого объединения, оптимизированный для работы с распределенными хранилищами данных в оперативной памяти. Данный метод отличается от известных аналогов тем, что минимизирует объем пересылаемых по сети данных.

This paper provides an overview of existing distributed join methods and proposes modified track join method and algorithm, optimized for usage with in-memory data grids. This method differs from known analogs minimizing data size sent across the network.

Введение

В последнее десятилетие, в связи с постоянно повышающимися требованиями к скорости доступа к данным, широкое распространение получили распределенные хранилища данных в оперативной памяти (in-memory data grid, IMDG). Первоначальная идеология таких хранилищ предусматривала выборку данных из IMDG исключительно методом доступа по ключу (т.н. NoSQL подход). Также была предусмотрена возможность выполнения пользовательских функций непосредственно на серверах хранилища. Однако опыт применения таких хранилищ показал необходимость выполнения SQL-подобных запросов. На текущий момент в наиболее распространенных реализациях IMDG (VmWare Gemfire, Oracle Coherence, Hazelcast) частично поддерживается объектный язык запросов (object query language, OQL), но ни одна из реализаций распределенных хранилищ в оперативной памяти не предполагает возможности выполнения слияний распределенных таблиц. Операция join (объединение) поддерживается только для таблиц, копии которых хранятся на узлах хранилища целиком (реплицированные таблицы)[1]. Таким образом, в случае необходимости выполнения объединения распределенных таблиц, пользователь вынужден выполнять эту операцию самостоятельно.

Современные технологии передачи данных могут быть медленными по сравнению с обработкой данных в оперативной памяти. Сеть, построенная по технологии InfiniBand с номинальной пропускной способностью 40 Гбит/сек, показала 3 Гбит/сек на узле во время выполнения секционирования данных при помощи хэш-функции. При выполнении в оперативной памяти, секционирование на несколько тысяч частей выполняется со скоростью, приблизительно равной пропускной способности оперативной памяти [2,3]. Следовательно, алгоритм объединения таблиц для хранилищ в оперативной памяти должен минимизировать пересылки данных по сети.

Анализ существующих решений

В большинстве распределенных баз данных на долговременных запоминающих устройствах для объединения таблиц используется Grace-метод объединения хешированием[4,5]. Особенностью этого метода является разбиение входных таблиц на части (партиции) с последующей записью их на диск, и последовательным чтением. Очевидно, что данный метод неэффективен для хранилищ в оперативной памяти. Применительно к IMDG, за основу можно взять метод гибридного объединения хешированием, являющийся модификацией Grace-метода. Гибридное объединение предусматривает хранение партиций в оперативной памяти, причем, во избежание переполнения памяти предусмотрена запись части партиций на диск. Такой подход позволяет, при наличии достаточного количества свободной памяти, выполнять объединение таблиц без использования медленных устройств. С другой стороны, при высокой интенсивности операций объединения или при дефиците памяти будет использовано устройство долговременного хранения, что позволит обработать запрос даже при недостатке ресурсов. Однако объединение хешированием далеко от оптимального использования сетевых ресурсов, так как при этом способе обе

таблицы-операнда передаются по сети практически полностью [6]. Использование предопределенных хэш-функций гарантирует балансировку нагрузки, но ограничивает вероятность того, что хэшированный кортеж не будет передан по сети до 1/N, где N – количество узлов распределенного хранилища.

Метод путевого объединения (track join), представленный в [6], минимизирует количество пересылок кортежей по сети. Основная идея этого метода состоит в определении цели отправки каждого конкретного кортежа. Метод путевого объединения существенно снижает использование сетевых ресурсов по сравнению с другими известными методами выполнения объединений распределенных таблиц. Однако он ориентирован на использование в традиционных базах данных, где хранимые кортежи строго типизированы и фактический объем хранимых данных в разных строках одной таблицы отличается несущественно. В отличие от них, IMDG хранят объекты, способные, в свою очередь, содержать другие объекты. Таким образом, объем хранимых данных в разных кортежах одной таблицы может отличаться на несколько порядков (а, теоретически, эта разница может быть и больше). Поэтому простой подсчет количества пересылаемых кортежей, используемый методом путевого объединения для оценки использования сетевых ресурсов, применительно к распределенным хранилищам в оперативной памяти не дает объективной количественной оценки объема пересылаемых данных.

Цель

Целью данной работы является разработка метода объединения распределенных таблиц, позволяющий количественно оценить и минимизировать фактический объем пересылаемых по сети данных при выполнении операции объединения в хранилищах данных в оперативной памяти.

Изучению подвергается внутреннее симметричное объединение по признаку равенства ключа таблиц распределенного хранилища данных в оперативной памяти.

Исходными данными являются две таблицы, R и S, произвольно распределенные по N узлам хранилища. Каждый узел может взаимодействовать со всеми остальными и все каналы передачи данных между узлами обладают одинаковой пропускной способностью.

Описание метода

Рассмотрим три версии модифицированного метода путевого объединения для хранилищ в оперативной памяти, по мере возрастания их сложности. Для всех трех случаев примем, что три процесса (R и S, работающие с локальными для каждого узла частями таблиц R и S соответственно, а также управляющий процесс T) работают одновременно на каждом узле, но только один раз на каждом узле.

Двухфазное путевое объединение – это простейшая версия метода. Она определяет узлы хранилища, содержащие как минимум один подходящий кортеж для каждого из ключей объединения (значения поля, по равенству которого выполняется объединение кортежей). Затем одна из таблиц кортеж за кортежем передается на все узлы, содержащие соответствующий ключ объединения. Такую процедуру передачи будем называть селективным широковещанием.

Необходимо отметить, что в данной статье описан оригинальный алгоритм двухфазного путевого объединения, представленный в [6]. Описание этого алгоритма приведено для облегчения понимания 3х- и 4х-фазного алгоритмов модифицированного путевого объединения.

Метод двухфазного путевого объединения состоит из двух этапов. Первый этап – определение путей передачи, второй - селективное широковещание.

Во время первой фазы обе таблицы R и S проецируются по ключу объединения, затем проекции пересылаются по сети. Узел назначения при этом определяется при помощи хеш-функции, так же, как пересылаются кортежи при обычном объединении хешированием. Каждый узел принимает уникальные ключи и хранит их вместе с идентификатором отправившего их узла. Эта операция выполняется отдельным процессом T, который также генерирует расписания передачи. Расположение процесса T определяется хэш-функцией ключа объединения, таким образом, распределяя процесс управления по узлам хранилища.

Во второй фазе выполняется передача кортежей таблицы R на узлы, хранящие соответ-

ствующие им по ключу объединения кортежи таблицы S. При этом процесс T отправляет сообщения на каждый узел, содержащий хотя бы один кортеж таблицы R, для которого было найдено совпадение в таблице S. Каждое сообщение содержит ключ объединения и соответствующее ему множество идентификаторов узлов, содержащих части таблицы S.

Выбор передаваемой таблицы (R или S) в этом случае должен быть выполнен оптимизатором до начала выполнения запроса, подобно выбору внутреннего и внешнего отношений в традиционных алгоритмах объединения хешированием.

Метод двухфазного путевого объединения (или однонаправленного селективного широковещания) передает по сети данные только одной из таблиц. В случае, когда входные таблицы имеют в основном уникальные значения ключей объединения, а селективность высока, объем пересылок данного алгоритма равен сумме стоимости путевого поиска и минимума кардинальных чисел множеств R и S.

Блок-схемы процессов R, S и T представлены на рис. 1 – 3.

Трехфазное путевое объединение

Метод модифицированного трехфазного путевого объединения (или двунаправленного селективного широковещания) позволяет во время выполнения определить, кортежи которой из таблиц (R или S) следует пересылать. Это решение принимается индивидуально для каждого из значений ключа объединения. Для этого каждый узел высылает координирующему процессу информацию не только о наличии определенного ключа объединения, но и о связанном с ним объеме хранимых данных.

Двунаправленное селективное широковещание позволяет определить случаи, когда пересылка кортежей таблицы S означает передачу меньшего объема информации, чем в случае пересылки кортежей таблицы R. Решение о направлении пересылки (R к S или наоборот) принимается для каждого уникального значения ключа объединения независимо, таким образом, оптимизатор запросов освобождается от необходимости при каждом выполнении оценивать общую выгодность и выбирать единственное направление пересылки для всей операции объединения.

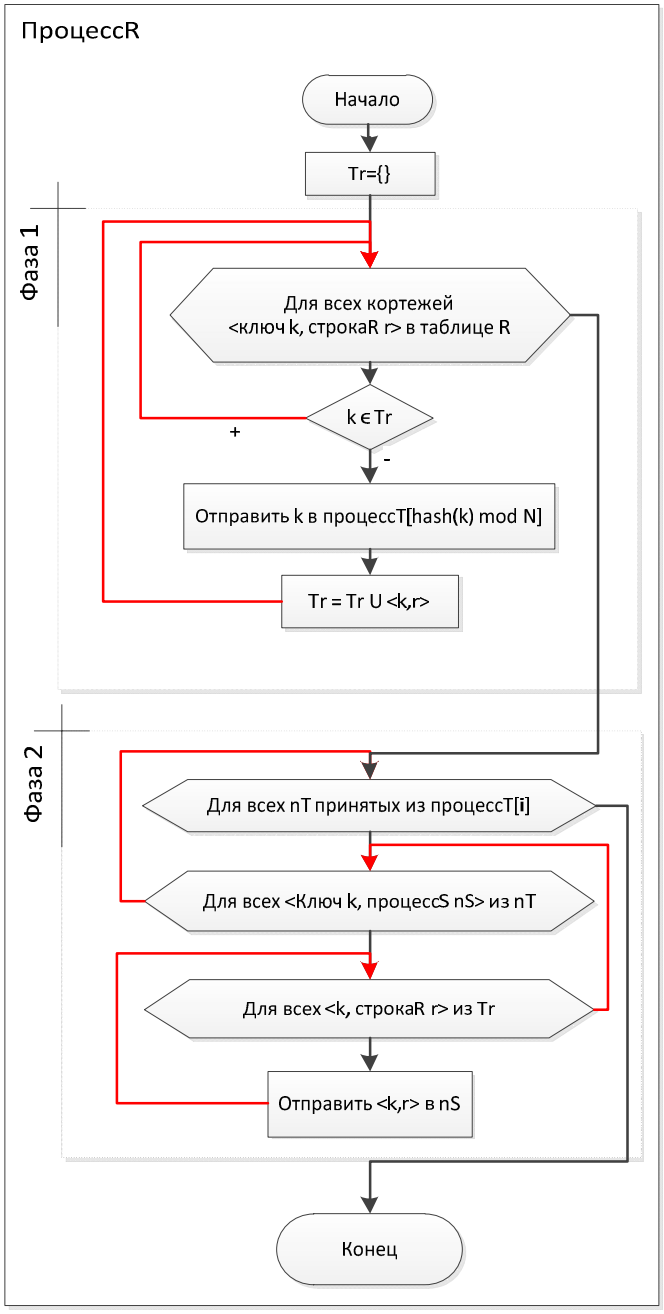

Рис 1. Алгоритм работы процесса R для двухфазного путевого объединения

В трехфазном алгоритме процессы, оперирующие таблицами R и S, симметричны. Блок-схема алгоритма такого процесса (на примере процесса R) приведена на рис. 4. Для того, чтобы получить блок-схему алгоритма процесса S, необходимо на рис.4 заменить все буквы r на s, и наоборот.

Алгоритм управляющего процесса представлен на рис. 5. Как видно, он использует процедуру определения стоимости однонаправленного селективного широковещания.

Метод выполнения оценки симметричен по направлению пересылки. Приведем оценку стоимости пересылки кортежей R к кортежам S.

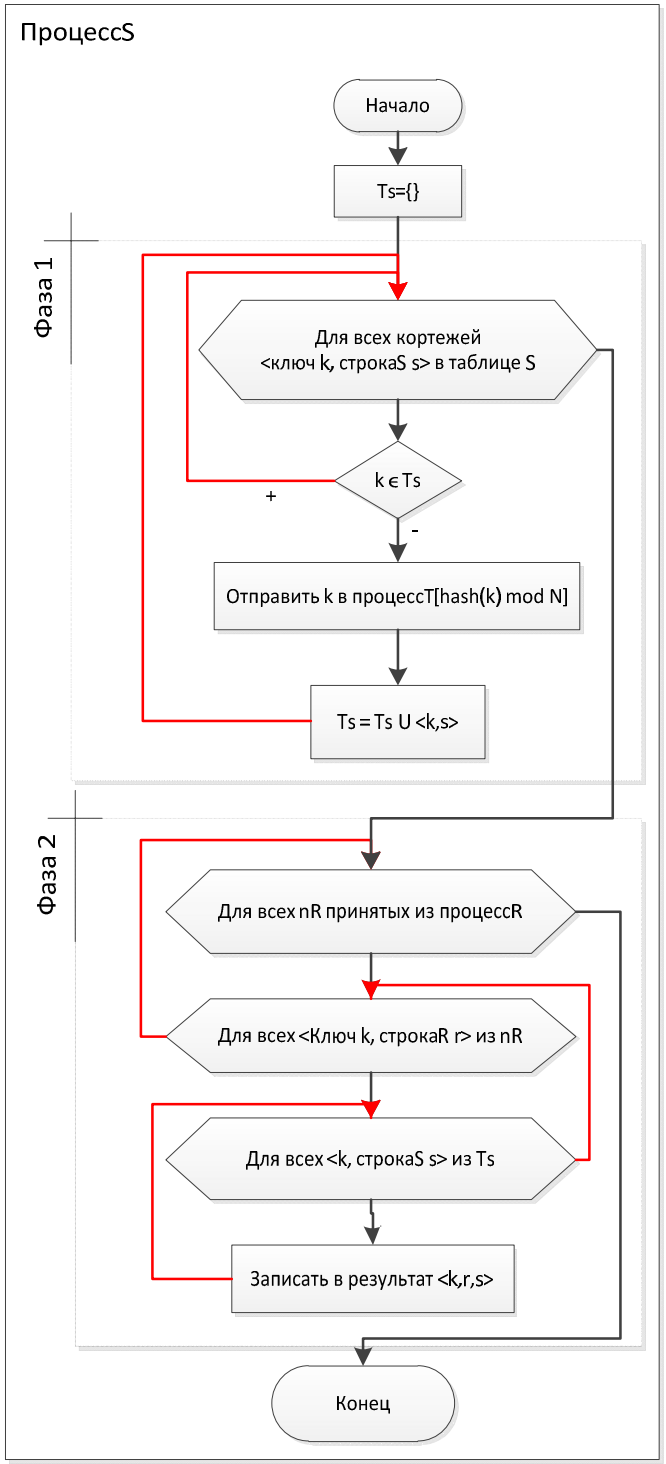

Рис 2. Алгоритм работы процесса S для двухфазного путевого объединения

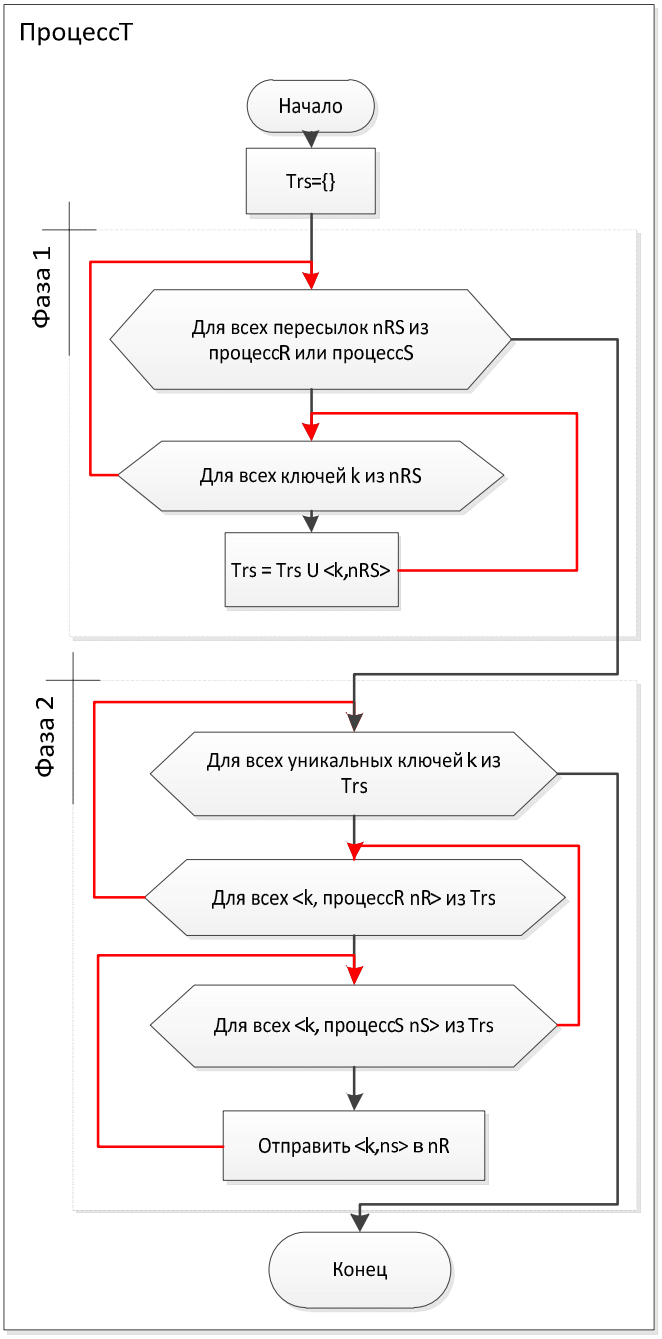

Рис 3. Алгоритм работы процесса T для двухфазного путевого объединения

В качестве исходных данных примем массивы R = {<k, id_r, c_r>} и S = {<k, id_s, c_s>}, каждый кортеж которых содержит значение ключа объединения, идентификатор узла, хранящего этот ключ и суммарный объем хранимых узлом кортежей, которые содержат ключ k. Оценка стоимости пересылки выполняется для каждого значения ключа k независимо, т.е. в каждом конкретном случае множества R и S содержат кортежи с единственным значением k.

Общий объем хранимых в таблице R данных, содержащих ключ k:

$$R_{общ} = \sum_{i=1}^{N} c_{r_i} . \qquad (1)$$

Здесь и далее N – количество узлов в распределенном хранилище, содержащих ключ объединения k.

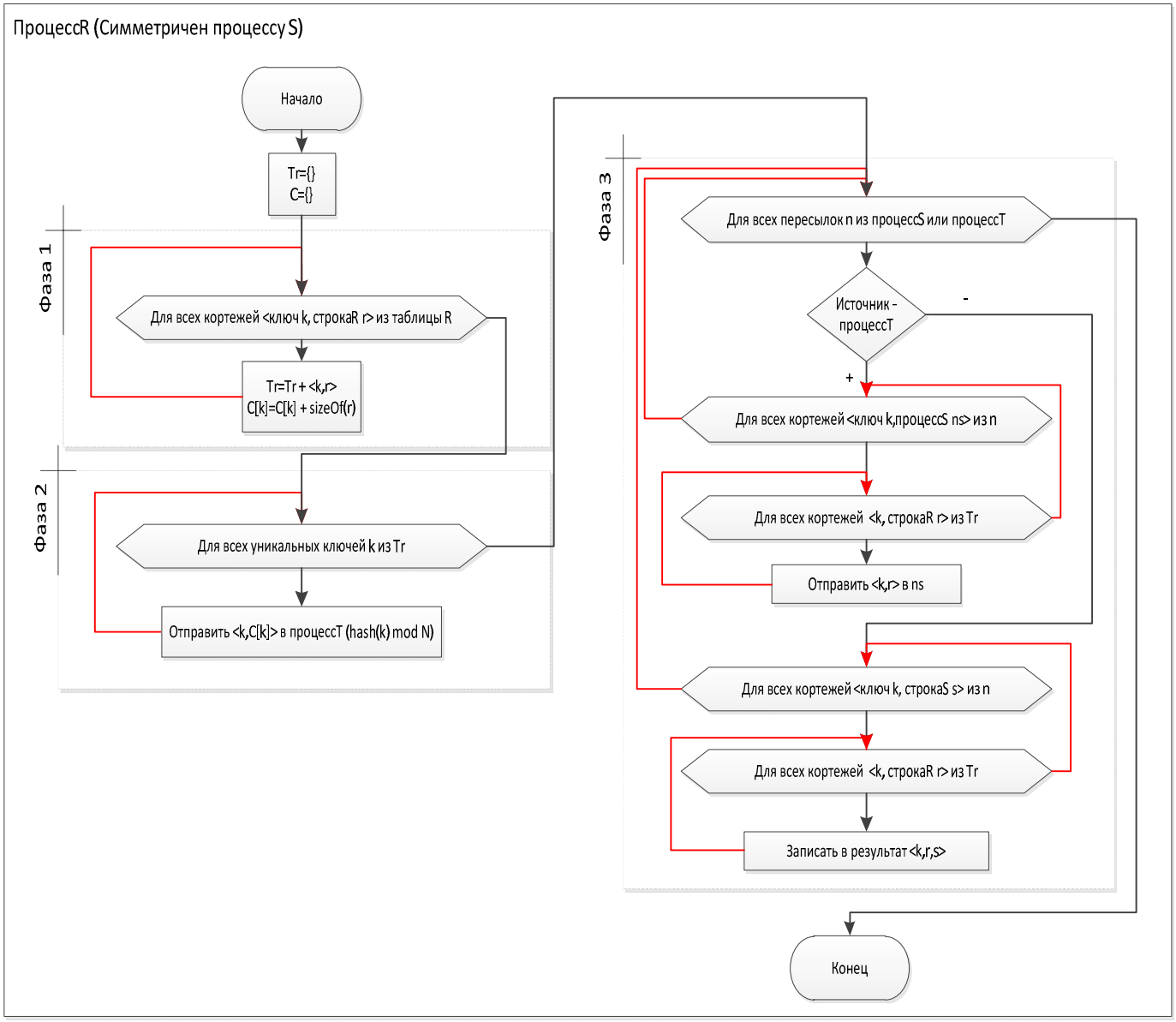

Рис 4. Блок-схема алгоритма процесса R в модифицированном трехфазном путевом объединении

Объем кортежей R, хранимых на одном узле с соответствующими им по ключу объединения кортежами таблицы S:

$$R_{лок} = \sum_{i=1}^{N} q_i \quad (2) \qquad где$$

$$q_i = \begin{cases} c_{r_i}, c_{s_i} > 0 \\ 0, c_{s_i} = 0 \end{cases} \quad (3)$$

Количество узлов, которые необходимо оповестить управляющему процессу:

$$N_r = \sum_{i=1}^{N} k_{r_i}, \quad (4) \qquad где$$

$$k_{r_i} = \begin{cases} 1, c_{r_i} > 0 \vee процессT \notin узел_i \\ 0, c_{r_i} = 0 \vee процессT \in узел_i \end{cases} \quad (5)$$

Количество узлов, содержащих кортежи таблицы S:

$$N_s = \sum_{i=1}^{N} k_{s_i} \quad (6)$$

$$Здесь \quad k_{s_i} = \begin{cases} 1, c_{s_i} > 0 \\ 0, c_{s_i} = 0 \end{cases} \quad (7)$$

Стоимость селективного широковещания от R к S равна сумме объемов пересылаемых кортежей первой таблицы, кроме локально хранимых с кортежами второй таблицы, и объема генерируемых управляющим процессом уведомлений об отправке:

$$RS_{cost} = R_{общ} \cdot N_s - R_{лок} + N_r \cdot N_s \cdot M, \quad (8)$$

Где M – размер уведомления об отправке, содержащего идентификатор узла-получателя и ключ объединения.

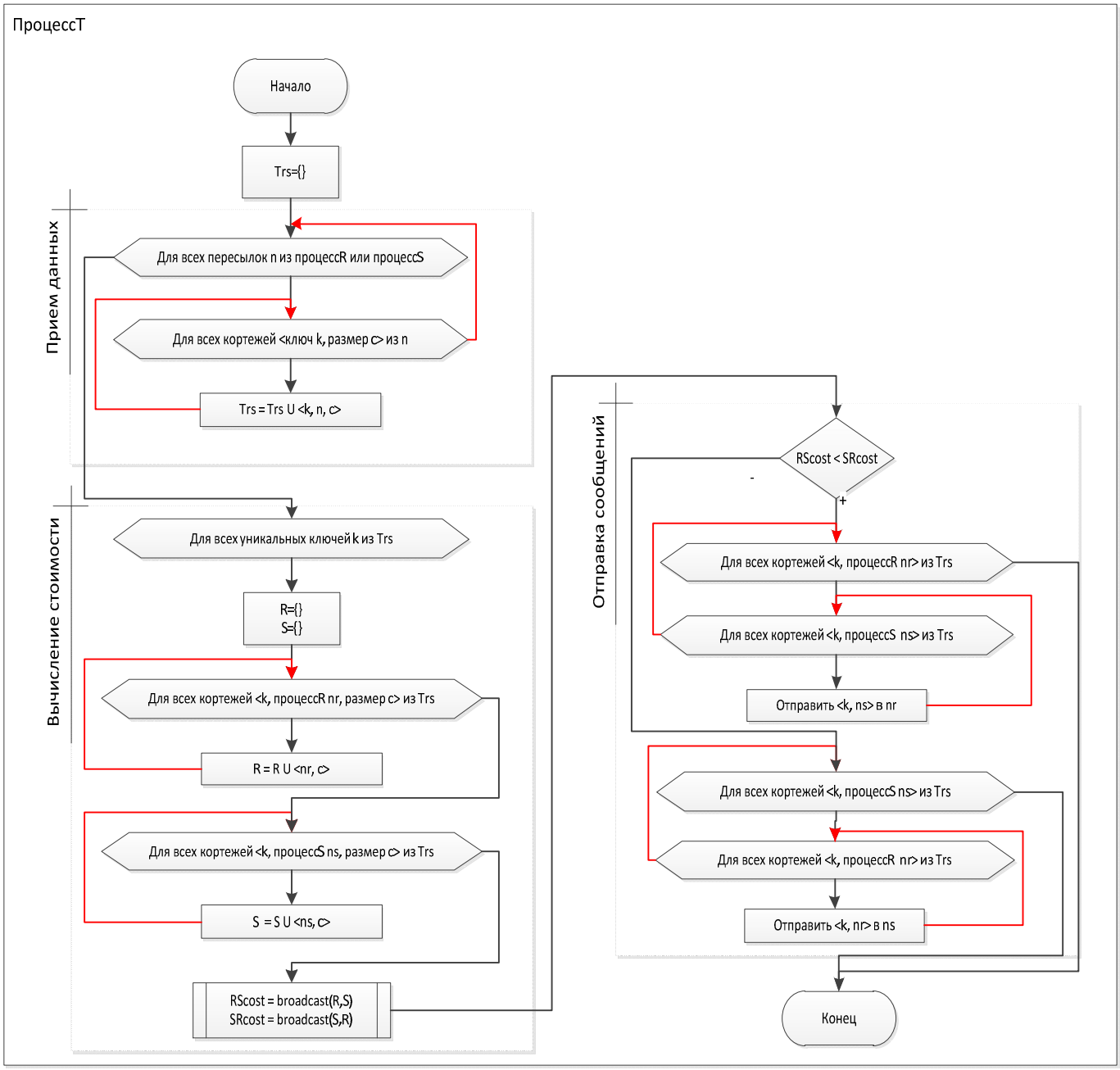

Рис. 5. Блок-схема управляющего процесса алгоритма модифицированного трехфазного путевого объединения

Количество выполняемых операций меньше количества анализируемых кортежей, следовательно, алгоритм является линейным по классу вычислительной сложности (O(N)).

Четырехфазное путевое объединение

В полной версии метода путевого объединения объем пересылок по сети оптимизируется за счет учета возможностей предварительных группировок кортежей одной из таблиц, содержащих определенный ключ объединения. Это достигается путем введения более сложного управляющего алгоритма, содержащего фазу миграции. Задачей этой фазы является определение подмножества узлов, содержащих кортежи одной из таблиц-операндов объединения, для которого временная сложность предварительного переноса кортежей первой таблицы и последующего селективного широковещания со второй будет наименьшей.

Блок-схема симметричных процессов R и S представлена на рис.6, управляющего процесса T – на рис.7.

Управление пересылками в четырехфазном методе иногда демонстрирует поведение, сходное с методом объединения хешированием, где все кортежи пересылаются на один узел хранилища. Даже в этом случае алгоритм путевого объединения, чтобы минимизировать стоимость

пересылок, выберет узел с наибольшим количеством уже хранимых кортежей.

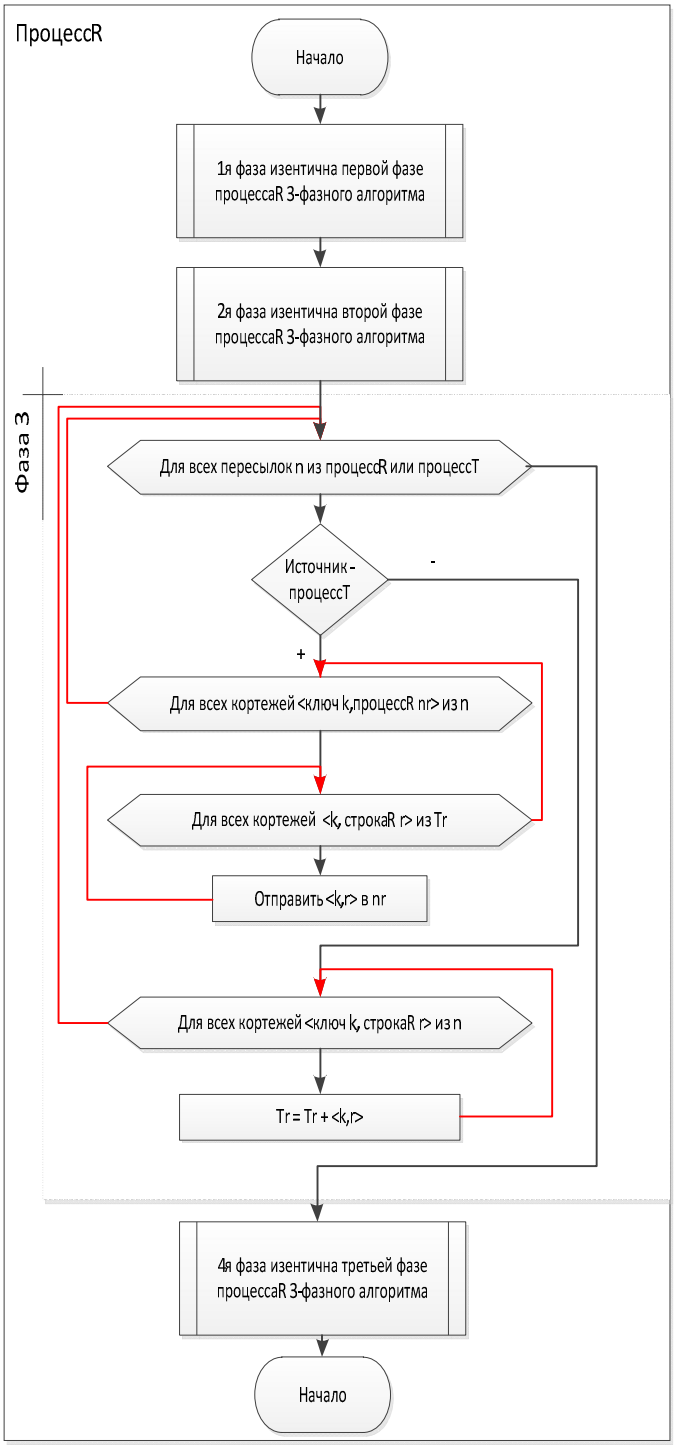

Рис 6. Блок-схема алгоритма процесса R в модифицированном четырехфазном путевом объединении

Введем функцию оценки временной сложности объединения двух таблиц по одному значению ключа объединения. Обозначим через x_{ij} булево решение о пересылке кортежей таблицы R с узла i на узел j, а y_{ij} – решение о пересылке кортежей S с i на j. Примем c_{r_i} как общий объем хранимых на узле i кортежей таблицы R, содержащих данный ключ объединения. Аналогично c_{s_j} обозначает объем кортежей таблицы S на узле j. Тогда объем пересылок для объединения кортежей по одному значению ключа будет

$$Q = \sum_i \sum_{j \neq i} x_{ij} \cdot c_{r_i} + y_{ij} \cdot c_{s_j} \qquad (9)$$

Где $\forall i, j \sum_k x_{ik} \cdot y_{kj} \geq 1$

Таким образом, определение способа объединения таблиц, оптимально использующего сетевые ресурсы, сводится к нахождению минимума функции Q.

Четырехфазный метод путевого объединения решает задачу минимизации сетевого трафика путем миграции, т.е. перегруппировки кортежей таблицы, выступающей получателем кортежей второй таблицы. Так, перед селективным широковещанием R→S, происходит миграция кортежей S таким образом, чтобы уменьшить объем последующей пересылки кортежей R.

Алгоритм определения стоимости пересылки $RS_{\cos t}$ и узлов, кортежи с которых должны мигрировать S_{migr} (в случае направления пересылки R→S), представлен на рис. 8. Алгоритм симметричен по направлению пересылки.

Рисунок 9 иллюстрирует схему сетевых пересылок для объединения таблиц хешированием, а также для разных версий путевого объединения. Для хеш-объединения примем, что хэш-функцией был выбран первый узел. Алгоритм двухфазного объединения пересылает кортежи таблицы R к кортежам таблицы S. Трехфазный алгоритм выполняет пересылку в обратном направлении, т.к. это снижает объем пересылок. Четырехфазный алгоритм выполняет предварительную миграцию кортежей R, также выполняя затем селективное широковещание кортежей S.

Оценка объема пересылок

Для упрощения примем равномерное распределение кортежей по узлам распределенного хранилища. Это наихудший случай для метода путевого соединения, так как он исключает локализацию данных[6].

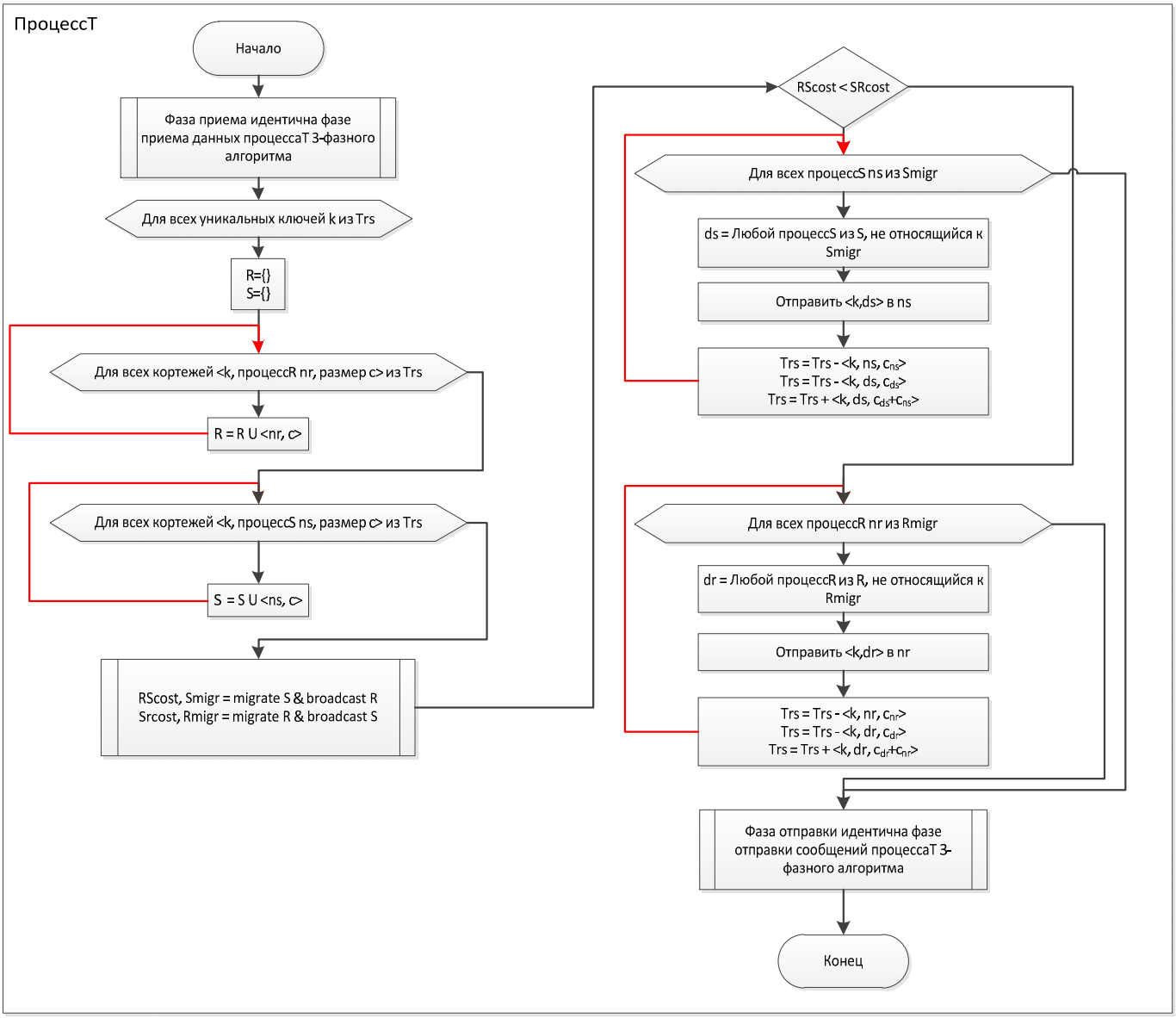

Рис. 7. Блок-схема управляющего процесса алгоритма модифицированного четырехфазного путевого объединения

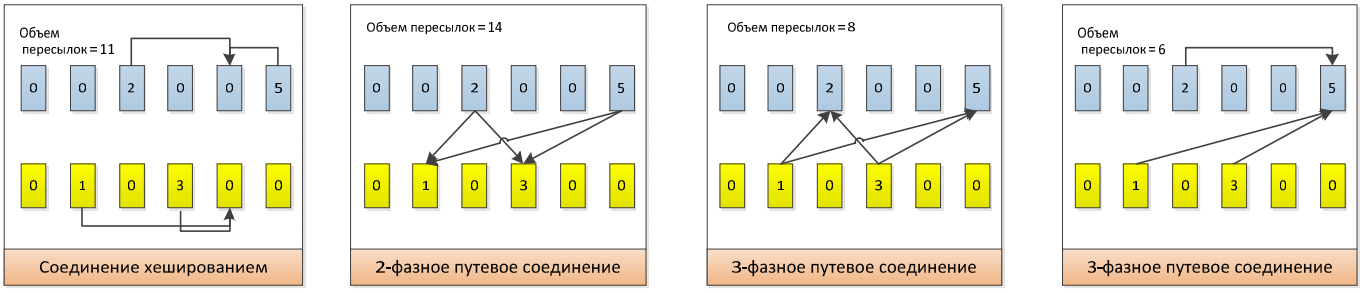

Рис. 9. Сравнение выполнения объединения путевыми алгоритмами и алгоритмом хеширования

Оценка объема пересылаемых по сети кортежей приводится применительно к распределенным хранилищам данных в оперативной памяти, использующим объектную модель.

Объем пересылок в случае использования метода соединения хешированием:

$$w_k \cdot (t_r + t_s) + \sum (c_{r_i} + c_{s_i})$$

Где w_k - объем памяти, занимаемый ключом объединения

t_r и t_s – кардинальные числа множеств R и S соответственно

c_{r_i} и c_{s_i} – совокупные объемы хранимых на i-м узле подлежащих объединению данных из таблиц R и S соответственно.

Для оценки объема пересылок в двухфазном путевом соединении, прежде всего, необходимо определить объем пересылок при определении пути. Количество узлов хранилища, содержащих конкретный ключ объединения, ограничено сверху количеством узлов N, и, в худшем случае, т.е. при равномерном распределении равных ключей по всем узлам, составляет t/d, где t – количество кортежей и d – количество уникальных ключей объединения. Мы будем обозначать количество узлов, содержащих конкретный ключ соединения как $n_R \equiv \min(N, t_R/d_R)$ и $n_S \equiv \min(N, t_S/d_S)$.

Введем понятие входной селективности (s_R и s_S) как процент кортежей одной таблицы, для которых существуют совпадения в другой таблице. Исходя из того, что совпадения существуют для всех значений ключа объединения, определим количество узлов, содержащих совпадения: $m_R \equiv \min(N, t_R \cdot s_R/d_R)$ и $m_S \equiv \min(N, t_S \cdot s_S/d_S)$

Исходя из этого, объем пересылок при использовании двухфазного метода путевого объединения будет

$(d_R \cdot n_R + d_S \cdot n_S) \cdot w_k$ (поиск путей)

$+ d_R \cdot m_S \cdot w_k$ (передача идентификаторов узлов с кортежами S)

$+ t_R \cdot s_R \cdot m_S \cdot (w_k + \sum c_{R_i}/d_R)$ (передача кортежей R к кортежам S)

Трехфазное путевое объединение предусматривает передачу общего объема хранимых данных, содержащих данный ключ объединения. Отношение $t/(d \cdot s)$ показывает среднее количество повторений ключей на каждом узле при равномерном распределении кортежей. Тогда максимальные размеры счетчиков объема пересылаемых по ключу данных будут $q_R = \log_2(\max(c_{R_i}) \cdot t_R/(d_R \cdot s_R))$ и $q_S = \log_2(\max(c_{S_i}) \cdot t_S/(d_S \cdot s_S))$.

Таким образом, объем пересылок при трехфазном путевом объединении составляет

$d_R \cdot n_R \cdot (w_k + q_R) + d_S \cdot n_S (w_k \cdot q_S)$ (поиск путей)

$+ d_{R_1} \cdot m_{S_1} \cdot w_k + t_{R_1} \cdot s_{R_1} \cdot m_{S_1} \cdot (w_k + \sum c_{R_{1i}}/d_{R_1})$

$(R_1 \to S_1)$

$+ d_{R_2} \cdot m_{S_2} \cdot w_k + t_{R_2} \cdot s_{R_2} \cdot m_{S_2} \cdot (w_k + \sum c_{R_{2i}}/d_{R_2})$

$(S_2 \to R_2)$

Здесь и далее R_1, S_1, R_2, S_2 – подмножества кортежей соответствующих таблиц, разделенные по направлению передачи.

Оценка объема пересылок для четырехфазного путевого объединения имеет вид

$d_R \cdot n_R \cdot (w_k + q_R) + d_S \cdot n_S (w_k \cdot q_S)$ (поиск путей)

$+ d_{R_3} \cdot m_{R_1} \cdot w_k + t_{R_3} \cdot s_{R_3} \cdot m_{R_1} \cdot (w_k + \sum c_{R_{3i}}/d_{R_3})$

(миграция $R_3 \to R_1$)

$+ d_{S_3} \cdot m_{S_2} \cdot w_k + t_{S_3} \cdot s_{S_3} \cdot m_{S_2} \cdot (w_k + \sum c_{S_{3i}}/d_{S_3})$

(миграция $S_3 \to S_2$)

$+ d_{R_1} \cdot m_{S_1} \cdot w_k + t_{R_1} \cdot s_{R_1} \cdot m_{S_1} \cdot (w_k + \sum c_{R_{1i}}/d_{R_1})$

$(R_1 \to S_1)$

$+ d_{R_2} \cdot m_{S_2} \cdot w_k + t_{R_2} \cdot s_{R_2} \cdot m_{S_2} \cdot (w_k + \sum c_{R_{2i}}/d_{R_2})$

$(S_2 \to R_2)$

Здесь R_3, S_3 обозначают множества кортежей, мигрирующие на другие узлы с целью минимизации объема пересылок.

Результаты экспериментов

Для сравнительной оценки использования методов модифицированного путевого объединения была создана система имитационного моделирования.

В системе были реализованы алгоритмы объединения хешированием, оригинальный 2х-, 3х- и 4х-фазный алгоритм путевого соединения и модифицированный алгоритм путевого соединения, представленный в данной статье.

Моделируемая система хранения данных содержит 16 узлов, на которых размещается в общей сложности 10^9 кортежей. В рамках модели было принято, что таблицы R и S содержат одинаковое количество кортежей. Изменяемым параметром в модели является допустимое отклонение размеров кортежей относительно минимального значения. Рассматривались значения 1 (т.е. все кортежи равны, и каждый кортеж содержит только примитивные типы данных), 2 и 10. Допустимое отклонение в 10 раз означает, что размер кортежа может быть от x до 10x, где x – заданное минимальное значение. В качестве минимального размера кортежа для таблицы R было принято 60 байт, для таблицы S – 40 байт.

Результаты имитационного моделирования приведены на рис. 10.

Проведенные исследования показали, что использование модифицированных методов путевого объединения целесообразно только при наличии вариативности размера пересылаемого кортежа. Таким образом, использовать его следует в случаях, когда пересылаемый кортеж содержит в себе объектные структуры данных. В случае же, когда кортеж содержит только примитивные типы, целесообразнее использовать оригинальный метод путевого объединения.

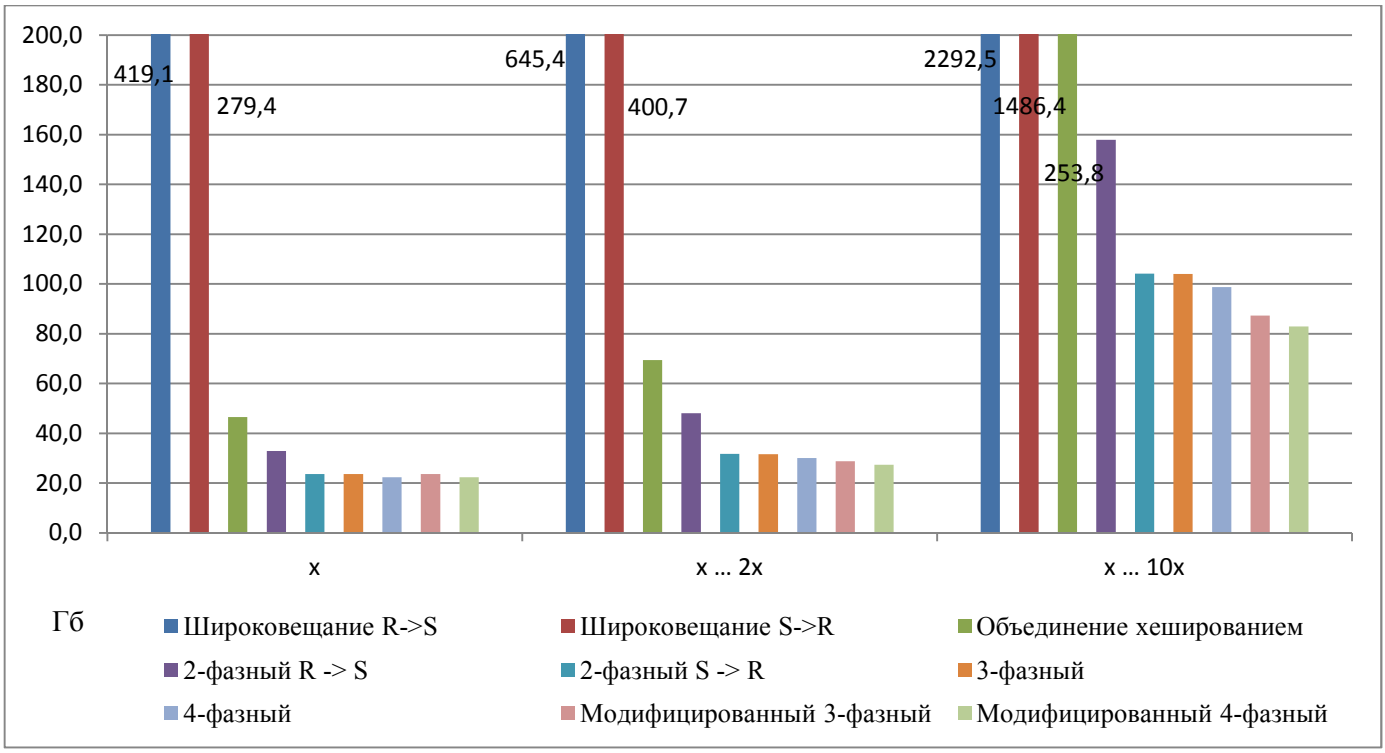

Рис. 10. Объем пересылок при использовании различных методов объединения

Вывод

Предложенные в данной статье модифицированные методы и алгоритмы путевого объединения распределенных таблиц позволяют уменьшить объем пересылаемых данных в хранилищах, использующих объектную модель данных, прежде всего, в хранилищах данных в оперативной памяти. Их применение позволяет значительно ускорить время выполнения операции объединения в распределенных хранилищах в оперативной памяти, а также увеличить количество одновременно выполняемых в хранилище операций.

Список литературы

1. А. Александров. Ударные СУБД в оперативной памяти // Открытые системы. СУБД – 2008 – №7 – С.36-44.
2. M.-C. Albutiu, A. Kemper, and T. Neumann. Massively parallel sort-merge joins in main memory multi-core database systems. // PVLDB, 5(10) –2012 – p.1064-1075.
3. C. Balkesen et al. Multicore, main-memory joins: Sort vs hash revisited. // PVLDB, 7(1) – Sept. 2013 – p.85-96.
4. D. J. DeWitt et al. The Gamma database machine project. // IEEE Trans. Knowl. Data Engin., 2(1) – 1990 – p.44-62.
5. M. Kitsuregawa, H. Tanaka, and T. Moto-Oka. Application of hash to data base machine and its architecture. //New Generation Computing – 1983 – p.63-74.
6. O. Polychroniou, R. Sen and K. Ross. Track join: distributed joins with minimal network traffic. // SIGMOD Conference – 2014 – p. 1483-1494.