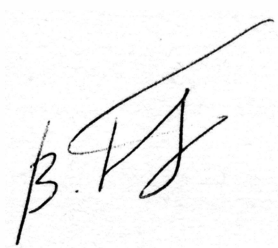
На правах рукописи

Плужников Всеволод Львович

Анализ архитектур параллельных систем баз данных

Специальность 05.13.17 – Теоретические основы информатики (технические науки)

Автореферат диссертации на соискание ученой степени кандидата технических наук



Москва - 2011

Работа выполнена в Московском государственном техническом университете им. Н. Э. Баумана.

Научный руководитель: доктор технических наук,
профессор Григорьев Юрий Александрович

Официальные оппоненты: доктор технических наук,
профессор Юрчик Пётр Францевич

кандидат технических наук,
Бурдаков Алексей Викторович

Ведущая организация: ОАО «Научно-исследовательский центр электронной вычислительной техники»

Защита диссертации состоится «9» февраля 2012 года в 14:30 на заседании диссертационного совета Д.212.141.10. в Московском государственном техническом университете имени Н.Э. Баумана по адресу: 105005, г. Москва, 2-я Бауманская ул., д.5.

С диссертацией можно ознакомиться в библиотеке МГТУ им. Н.Э. Баумана.

Ваши отзывы в 2-х экземплярах, заверенные печатью, просим выслать по указанному адресу.

Автореферат разослан «___» _____ 20__ г.

Учёный секретарь
диссертационного совета,
к.т.н., доцент С.Р. Иванов

Подписано к печати «16» декабря 2011 г. Объем 1 п.л. Тираж 100 экз.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы.

Реализация параллельных систем баз данных с помощью современных технических средств обеспечивает высокую производительность выполнения запросов. В настоящее время существует несколько типов архитектур, позволяющих реализовывать параллельные системы баз данных. Технические средства, используемые для реализации этих архитектур, являются дорогостоящими, что приводит к необходимости учитывать показатель «производительность/стоимость» системы при выборе архитектуры.

Существующие методы анализа и выбора архитектуры систем рассматриваемого класса основаны или на сопоставлении вариантов по качественным критериям (масштабируемости, доступности данных и др.), или на сравнении результатов выполнения конкретных тестов (TPC и др.), не учитывающих особенностей предметной области, для которой разрабатывается система. Выбор архитектуры с помощью этих методов нельзя считать обоснованным, их использование может привести или к чрезмерному завышению стоимости проекта, или к выбору системы с низкой производительностью.

Поэтому разработка математических моделей анализа архитектур параллельных систем баз данных, позволяющих выбирать структуру сложного многопроцессорного аппаратно-программного комплекса с минимальной стоимостью, обеспечивающего выполнение ресурсоёмких запросов к базе данных за допустимое время, является актуальной задачей.

В диссертационной работе указанная задача решается путем разработки моделей оценки индексов производительности параллельных систем баз данных, учитывающих особенности выполнения запросов различных типов к базе данных, механизм распределения таблиц по процессорам системы, параллелизм выполнения запросов в узлах, наличие «узких мест» в многопроцессорных комплексах с различной топологией.

Цель работы. Целью данной работы является разработка метода выбора архитектуры параллельной системы баз данных на основе применения математических моделей оценки характеристик производительности с учетом специфики решаемых ею задач и стоимости.

В работе решаются следующие **задачи**:

1) разработка метода выбора архитектуры параллельной системы баз дынных (ПСБД) на основе показателей стоимости и времени выполнения запросов к системе;
2) разработка аналитических моделей выполнения запросов в ПСБД с различными архитектурами, включая хранилища данных ROLAP;
3) разработка метода оценки стоимости ПСБД для различных архитектурных решений;
4) применение разработанных моделей и методов для выбора архитектуры ПСБД хранилища гидрометеорологических данных.

Объект исследования. Объектом исследования является класс параллельных систем баз данных.

Предмет исследования. Предметом исследования настоящей работы являются процессы обработки запросов в различных структурах параллельных систем баз данных.

Научная новизна. В работе получены следующие новые научные результаты:

1. Разработана модель обработки запросов в параллельной системе баз данных в виде замкнутой и разомкнутой СМО, учитывающая наличие "узкого места" в системе.
2. Выведено преобразование Лапласа-Стилтьеса времени выполнения запроса к одной таблице в параллельной СУБД. Рассмотрены варианты этого преобразования для различных архитектур параллельных систем баз данных (ПСБД).
3. Разработан математический метод оценки времени соединения таблиц в параллельной системе баз данных для различных архитектур (SE, SD, SN) и разных методов реализации соединения (NLJ, HJ).
4. Выведены преобразования Лапласа-Стилтьеса и получены моменты случайного времени выполнения аналитических запросов к хранилищу данных, реализованному на основе ПСБД и использующему специальные планы соединения таблиц измерений и фактов.

Методы исследования. Исследования проводились на основе комплексного использования теории массового обслуживания, теории вероятностей, теории множеств, теории реляционных баз данных.

Практическая ценность полученных результатов.

В диссертации разработан алгоритм выбора архитектуры параллельной системы баз данных, основанный на упорядочивании ПСБД с архитектурами SE, CE, SN, SE-кластер по возрастанию их стоимости.

В работе для практического использования полученных результатов разработано инструментальное средство, позволяющее проводить расчеты временных показателей выполнения запросов к ПСБД. Оно включает в себя модули расчета для различных типов архитектур и позволяет строить зависимости среднего времени выполнения запросов в системе от количества процессоров, параметров запросов и наполнения базы данных.

Внедрение результатов исследований. Разработанные методы и инструментальное средство было использовано в процессе выбора архитектуры хранилища гидрометеорологических данных. Хранилище данных обеспечивает выполнение трех основные задач: накопление данных, их бессрочное хранение и обслуживание потребителей. В соответствии с предъявленными требованиями были определены допустимые архитектуры ПСБД и технические средства для их реализации. Проведены оценки временных показателей выполнения запросов к хранилищу ПСБД, выполнена оценка стоимости систем. На основе этих расчётов решена задача выбора архитектуры ПСБД с минимальной стоимостью.

Публикации по теме. По материалам работы опубликовано 6 печатных работ.

Апробация работы. Материалы работы были изложены автором на НТС кафедры ИУ-5 МГТУ им. Н.Э. Баумана, М., 2009-2011.

Объем работы. Диссертационная работа содержит 154 страниц, 38 рисунков и 19 таблиц, список литературы из 135 наименований.

СОДЕРЖАНИЕ РАБОТЫ

Во **введении** обосновывается актуальность проблемы. Формулируются цели и задачи исследований, приводится перечень основных результатов, выносимых на защиту, и излагается краткое содержание глав диссертации.

В **первой главе** «Анализ существующих методов выбора архитектур параллельных систем баз данных» приведено описание параллельных систем баз данных (ПСБД), особенностей их функционирования и возможных архитектурных решений. Приведено описание форм параллелизма и способов параллельной обработки запросов в ПСБД. Выполнен критический анализ существующих методов выбора архитектуры на этапе проектирования систем. На основе этого анализа предлагается общая методика выбора архитектуры ПСБД.

Процесс выполнения SQL-запроса в параллельной системе баз данных можно представить в виде следующих шагов:

- генерация последовательного плана выполнения запроса,
- тиражирование плана выполнения запроса на все узлы системы,
- обработка запроса над фрагментированными таблицами (распределение фрагментов таблиц БД по узлам системы выполняется заранее и один раз),
- слияние результирующих данных.

Рассмотрим процесс параллельной обработки запроса, где выполняется соединение таблиц R и S базы данных (рис. 1).

$Q = R \triangleright\triangleleft S$ – это логическая операция соединения (join) двух отношений (таблиц) R и S по некоторому общему атрибуту Y. В данном примере таблица R фрагментирована произвольным образом, а таблица S – по атрибуту соединения Y. На рис. 1 показано, что логический план выполнения соединения двух отношений тиражируется на 'n' процессоров в параллельной системе баз данных (на рисунке показаны 2 процессора). Далее выполняется параллельная обработка на каждом процессоре соответствующих фрагментов таблиц R и S. Вследствие того, что таблица R не фрагментирована по атрибуту соединения, при последовательном чтении записей этой таблицы происходит их обработка в операторе exchange, осуществляющем разбор записи и её межпроцессорный обмен. Таблица S фрагментирована по атрибуту соединения и записи, читаемые из фрагментов этой таблицы, обрабатываются на каждом процессоре локально.

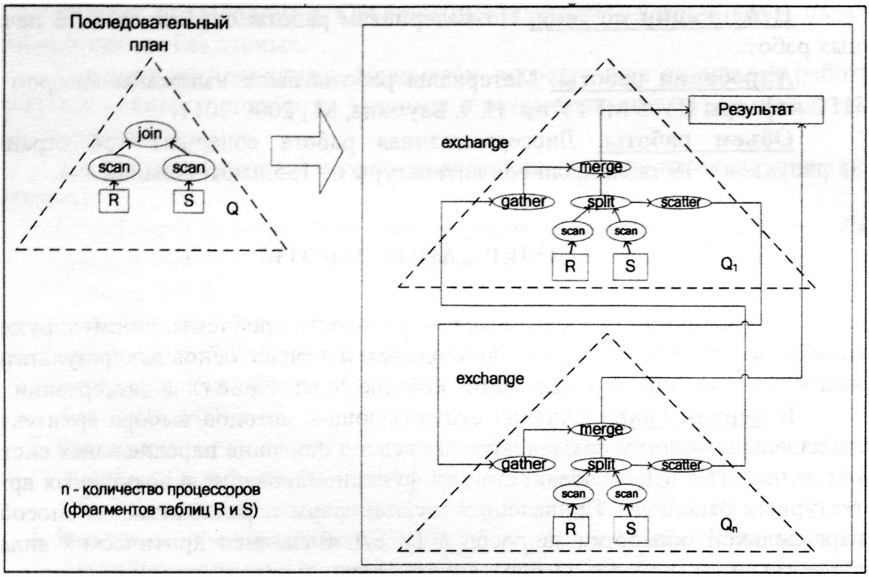


Рис. 1. Обработка запроса $Q = R \triangleright\triangleleft S$ в параллельной системе баз данных

Оператор exchange является составным оператором и включает в себя четыре оператора: gather, scatter, split и merge. Оператор split - это оператор, который осуществляет разбиение кортежей (записей), поступающих из входного потока, на две группы: свои и чужие. Свои кортежи - это кортежи, которые должны быть обработаны на данном процессорном узле. Эти кортежи направляются в выходной буфер оператора split (стрелка вверх). Чужие кортежи, то есть кортежи, которые должны быть обработаны на процессорных узлах, отличных от данного, помещаются оператором split во входной буфер правого дочернего узла, в качестве которого фигурирует оператор scatter. Нульарный оператор scatter извлекает кортежи из своего входного буфера и пересылает их на соответствующие процессорные узлы. Нульарный оператор gather выполняет перманентное чтение кортежей из указанного порта со всех процессорных узлов, отличных от данного. Оператор merge, реализующий логическую операцию join, определяется как бинарный оператор, который забирает кортежи из выходных буферов своих дочерних узлов, соединяет их и помещает результат в собственный выходной буфер. Таким образом, с помощью рассмотренных операций оператор exchange реализует полноценный межпроцессорный обмен записями в параллельной системе баз данных при обработке запроса методом фрагментарного параллелизма.

Анализ существующих решений ПСБД позволил выявить основные используемые на текущий момент типы архитектур:

1. SE (Shared-Everything) - архитектура с разделяемыми памятью и дисками.
2. SD (Shared-Disks) - архитектура с разделяемыми дисками.

3. SN (Shared-Nothing) - архитектура без совместного использования ресурсов.
4. CE (Clustered-Everything) - архитектура с SE-узлами, объединенными по принципу SN.

В настоящее время применяются два основных метода выбора архитектуры: 1) опытное сравнение производительности и стоимости систем на базе компонентных или интегральных тестов, 2) экспертная оценка архитектур ПСБД. Основным недостатком первого метода является то, что здесь используются обобщенные тестовые модели, не учитывающие специфических особенностей предметной области, для которой выбирается архитектура системы. Недостатком экспертных оценок является отсутствие оценок количественных показателей производительности, а также субъективный характер результатов сравнения систем.

В работе предложена концепция выбора архитектуры параллельной системы баз данных на основе оценки временных характеристик выполнения запросов к ПСБД и оценки стоимости системы. На первом этапе определяется множество различных архитектурных решений, которые могут быть использованы для реализации параллельной системы баз данных в рамках заданной предметной области. На втором этапе перечисляются технические и программные средства, с помощью которых данные архитектуры реализуются. На третьем этапе выполняется количественная оценка стоимостных показателей систем и показателей времени выполнения SQL-запросов предметной области. На четвертом этапе происходит непосредственный выбор архитектуры минимальной стоимости при ограничении на допустимое время выполнения запросов.

Во **второй главе** «Разработка математических методов анализа характеристик производительности параллельных систем баз данных» разработаны модели обработки запросов для различных архитектур ПСБД. Предложено аналитическое решение по данным моделям. С помощью преобразования Лапласа-Стилтьеса получены выражения для оценки среднего времени выполнения простого SQL запроса и запроса на соединение таблиц в различных архитектурах ПСБД. Исследованы зависимости времени выполнения запросов от количества процессоров в системе на примере реальной системы.

Разработана модель выполнения запросов к ПСБД с планом $\pi_A(\sigma_F(R))$ в виде замкнутой СМО, узлы которой соответствуют ресурсам системы: диску, ОП, процессорам, соединительной шине. Число заявок в этой СМО равно количеству процессоров в ПСБД. Показано, что при наличии «узкого места» эта модель может быть сведена к модели «ремонтника», а затем - к разомкнутой СМО M/M/1. Определены параметры этой модели.

Но в исходной замкнутой СМО трудно в наглядном виде отразить процесс выполнения запроса к ПСБД: чтение и обработку записей в ресурсах системы, фильтрацию записей, межпроцессорный обмен. В рассмотренной выше модели это учитывается только посредством расчёта переходных вероятностей. Чтобы устранить указанный недостаток, в работе был использован

аппарат производящих функций (ПФ) и преобразований Лапласа-Стилтьеса (ПЛС).

Например, $G(1-P_F(1-\phi_1(s)))$ - это ПЛС времени обработки записей, удовлетворяющих условию поиска с вероятностью P_F, с ПФ числа записей в таблице $G(z)$ и с ПЛС времени обработки одной записи $\phi_1(s)$. Кроме того, этот аппарат позволяет рассчитывать не только математические ожидания случайных величин, но и моменты более высоких порядков.

В работе выведено преобразование Лапласа-Стилтьеса времени выполнения запроса с планом $\pi_A(\sigma_F(R))$ к одной таблице ПСБД:

$$\phi(s) = G(\phi_D(s)\phi_M^2(s)(1-P_F(1-\phi_N(s)))\phi_P(s)), \qquad (1)$$

где $G(z)$ - производящая функция числа записей фрагментной таблицы R, обрабатываемых на одном процессоре; для равномерного распределения записей по процессорам $G = z^{V/n}$; V - общее число записей в таблице R, n – число процессоров в ПСБД; $\phi_D(s)$ - ПЛС времени чтения записи БД фрагментированной таблицы с диска (с учетом общесистемного буфера и очереди к дисковому массиву), $\phi^2{}_M(s)$ - ПЛС времени сохранения и чтения записи фрагментированной таблицы из оперативной памяти (с учетом очереди к шине памяти), $\phi_N(s)$ - ПЛС времени межпроцессорного обмена при передаче результирующей записи по сети N, $\phi_P(s)$ - ПЛС времени обработки записи в процессоре, который является неразделяемым ресурсом; P_F – вероятность, что запись удовлетворяет условию поиска F (эта вероятность рассчитывается по известным формулам).

В работе приведены аргументы, позволяющие считать, что каждое из преобразований Лапласа-Стилтьеса $\phi_D(s)$, $\phi_M(s)$, $\phi_N(s)$, $\phi_P(s)$ соответствует или времени пребывания в СМО M/M/1 (ожидание и обслуживание), если соответствующий ресурс является разделяемым, или просто времени обслуживания, если ресурс является неразделяемым. Время пребывания в СМО M/M/1 распределено по экспоненциальному закону. В диссертации приведены выражения для указанных преобразований для различных архитектурных решений.

В предположении, что в ПСБД с архитектурой SE «узким местом» является диск, из (1) получена оценка среднего времени выполнения запроса (математическое ожидание):

$$M_\xi = -\phi'(0) = \frac{V}{n}\left(\frac{1}{\mu_D - n\lambda_D} + \frac{1}{\mu_{MP}}\right), \quad \frac{1}{\mu_{MP}} = \frac{\mu_M + \mu_P(2+P_F)}{\mu_M\mu_P}, \qquad (2)$$

где λ и μ - это интенсивности поступления и обработки записей в ресурсе (D – диске, M – оперативной памяти, P – процессоре).

График функции $M_\xi(n)$ имеет выраженный минимум в точке $n^* = \frac{\mu_D}{\lambda_D} - \frac{\mu_{MP}}{\lambda_D}\left(\sqrt{\frac{\mu_D}{\mu_{MP}} + 1} - 1\right)$. Это объясняется тем, что сначала с ростом числа процессоров время убывает, благодаря распараллеливанию обработки запроса, затем время возрастает из-за перегрузки подсистемы ввода/вывода.

В работе выполнен анализ процесса соединения таблиц в параллельной системе баз данных (ПСБД). Получены преобразования Лапласа-Стилтьеса времени соединения двух таблиц A и B методами NLJ и HJ:

$$\Psi_i^{NLJ}(s) = H_{Ai}(s, G_{Bi}(\phi_D(s)\phi_M^2(s)\phi_P(s)(1 - P_B\eta_{AB}(1 - \phi_N(s))))), \quad (3)$$

$$\Psi_i^{HJ}(s) = H_{Ai}(s, (\phi_D(s)\phi_M^2(s))^{\omega_A}) \times$$
$$G_{Bi}(\phi_D(s)\phi_M^2(s)\phi_P(s)(1 - P_B(1 - (\phi_D(s)\phi_M^2(s))^{\omega_B}))) \times \quad , \quad (4)$$
$$W_{Ai}(G_{Bi}(1 - P_B + P_B(1 - \frac{1}{r} + \frac{1}{r}\phi_P(s))(1 - \eta_{AB} + \eta_{AB}\phi_N(s))))$$

где i – номер процессора,

$$H_{Ai}(s,z) = V_{Ai}(s, \overset{1}{\phi_N}(s), ..., \overset{i-1}{\phi_N}(s), \overset{i}{z}, \overset{i+1}{\phi_N}(s), ..., \overset{n}{\phi_N}(s)) \times$$
$$\times \prod_{j=1, j\neq i}^{n} V_{Aj}(0, 1_1, ..., 1_{i-1}, z, 1_{i+1}, ..., 1_n) \quad , \quad (5)$$

$$V_{Ai}(s, z_1, ..., z_n) = G_{Ai}(\phi_D(s)\phi_M^2(s)\phi_P(s)(1 - P_A(1 - q_{Ain}(z_1, ..., z_n)))), \quad (6)$$

$G_{Ai}(1 - P_A(1 - z))$ - это производящая функция числа записей, удовлетворяющих условию поиска по таблице A (с вероятностью P_A),

$q_{Ain}(z_1, ..., z_n)$ определяется следующими рекуррентными формулами:

$$q_{Ain}(z_1, ..., z_n) = (1 - p_{Ain})q_{Ain-1}(z_1, ..., z_{n-1}) + p_{Ain}z_n,$$
$$q_{Ain-1}(z_1, ..., z_{n-1}) = (1 - p_{Ain-1})q_{Ain-2}(z_1, ..., z_{n-2}) + p_{Ain-1}z_{n-1}, \quad (7)$$
...
$$q_{Ai1}(z_1) = (1 - p_{Ai1}) + p_{Ai1}z_1, \; p_{Ai1} \equiv 1,$$

p_{Aij} - это вероятность, что запись передаётся из i-го узла в j-й узел при условии, что она не была передана в узлы n...j+1,

$G_{Bi}(1 - P_B\eta_{AB}(1 - z))$ - производящая функция числа записей таблицы B, удовлетворяющих условию поиска по таблице B (вероятность P_B) и условию соединения (вероятность η_{AB}),

$$W_{Ai}(z) = \prod_{j=1}^{n} V_{Aj}(0, 1_1, ..., 1_{i-1}, z, 1_{i+1}, ..., 1_n), \quad (8)$$

r – число разделов (хеш-групп) в хеш-таблице,

показатели степени $\omega_A \leq 2$ и $\omega_B \leq 2$ определяют число чтений/записей на диск хеш-групп.

Полученные формулы учитывают число записей соединяемых таблиц и фрагментацию таблиц по узлам системы ($G_{Ai}(z)=z^{\frac{V_A}{n}}$ и $G_{Bi}(z)=z^{\frac{V_B}{n}}$), фильтрацию записей этих таблиц (вероятности P_A и P_B), вероятность совпадения значений атрибутов соединения (η_{AB}), варианты размещения хэш-таблицы в оперативной памяти (r, ω_A и ω_B), а также параметры межпроцессорного обмена записями между узлами ПСБД (см. (7)).

В работе после дифференцирования выражений (3), (4) в точке s=0 получены формулы для математического ожидания времени соединения двух таблиц в ПСБД с архитектурами SE, SD, SN методами NLJ и HJ. В работе приведён практический пример расчёта среднего времени соединения таблиц большой размерности ($V_A=V_B=10^6$) методами NLJ и HJ в зависимости от числа процессоров в ПСБД, который показал наличие выраженного минимума. Анализ графиков позволил сделать несколько выводов:

1. Графики для архитектур SE и SD практически совпали. Это объясняется высокими значениями интенсивностей μ_M и μ_N.

2. При $n\geq 3$ среднее время для метода соединения HJ на два порядка меньше среднего времени для NLJ (это и следовало ожидать для неиндексированной по атрибуту соединения вложенной таблицы B). Более того, $M_{NLJ}(1)\approx 9400$ с., $M_{HJ}(1)\approx 19$ с.

3. При $n\leq 7$ архитектуры SE и SD не на много хуже SN. Следует также отметить, что при n=7 загрузка диска для SE и SD равна 0,63.

4. Для SE и SD при n=11 перегружается дисковая подсистема, и дальнейший рост числа процессоров не имеет смысла. Для архитектуры SN время продолжает уменьшаться с ростом n (здесь нет разделяемых ресурсов). Однако следует иметь в виду, что для снижения времени выполнения соединения методом HJ с 2 секунд до 1 необходимо увеличить количество процессов с 10 до 20, а это может оказаться экономически не целесообразным.

В **третьей главе** «Разработка математических методов оценки характеристик производительности хранилищ данных на основе параллельных баз данных. Оценка стоимости ПСБД» предложены выражения для определения временных показателей выполнения запроса к хранилищу данных, построенному на основе ПСБД. Приводятся примеры использования этих выражений для расчета среднего времени выполнения запроса к хранилищу. Также приводится описание метода стоимостной оценки ПСБД. Разрабатывается оригинальный алгоритм выбора архитектуры параллельной системы баз данных.

В работе получено преобразование Лапласа-Стилтьеса (ПЛС) времени выполнения запроса к хранилищу данных, которое справедливо для каждого процессорного узла параллельной системы баз данных:

$$\Psi_i(s)=R(s)\cdot H_i(s,\varphi_{PS}(s)\chi(s)\varphi_{PA}(s)), \qquad (9)$$

где i – номер процессора,

$$R(s)=\prod_{j=1}^{K} G_j(1-p_j(1-\chi^2(s))), \qquad (10)$$

$G_j(z)=z^{\frac{V_j}{n}}$ - производящая функция числа записей в таблице j-го измерения в узле, K – число измерений, V_j – общее число записей в таблице j-го измерения, n – число процессоров, p_j - вероятность, что запись таблицы j-го измерения удовлетворяет условию поиска по этому измерению в запросе, $\chi^2(s)$ - ПЛС времени обработки записи блока листового уровня индекса ($\chi(s)$) и записи таблицы измерения ($\chi(s)$),

$$\chi(s)=\varphi_D(s)\varphi_M^2(s)\varphi_{PI}(s), \qquad (11)$$

$$H_i(s,z)=V(s,\varphi_N^1(s),...,\varphi_N^{i-1}(s),z,\varphi_N^{i+1}(s),...,\varphi_N^n(s))\times \\ \times V^{n-1}(0,1_1,...,1_{i-1},\varphi_M^2(s)z,1_{i+1},...,1_n), \qquad (12)$$

$$V(s,Z)=Y_{K-1}(s,G_K(1-p_K(1-\varphi_{PJ}(s)\varphi_M^2(s)q_n(z_1,\ldots,z_n)))), \qquad (13)$$

$$Y_{K-1}(s,z)=Y_{K-2}(s,G_{K-1}(1-p_{K-1}(1-\varphi_{PJ}(s)\varphi_M^2(s)\varphi_N^{n-1}(s)\varphi_M^{2(n-1)}(s)z^n))), \qquad (14)$$

$$Y_2(s,z)=Y_1(s,G_2(1-p_2(1-\varphi_{PJ}(s)\varphi_M^2(s)\varphi_N^{n-1}(s)\varphi_M^{2(n-1)}(s)z^n))),$$

$$Y_1(s,z)=G_1(1-p_1(1-\varphi_N^{n-1}(s)\varphi_M^{2(n-1)}(s)z^n)),$$

$q_n(z_1,...,z_n)$ определяется следующими рекуррентными формулами:

$$q_n(z_1,...,z_n)=(1-p_n)q_{n-1}(z_1,...,z_{n-1})+p_nz_n, \\ q_{n-1}(z_1,...,z_{n-1})=(1-p_{n-1})q_{n-2}(z_1,...,z_{n-2})+p_{n-1}z_{n-1}, \qquad (15)$$
$$\ldots$$
$$q_1(z_1)=(1-p_1)+p_1z_1,\ p_1\equiv 1,$$

p_m - это вероятность, что запись передаётся из данного узла в m-й узел при условии, что она не была передана в узлы n...m+1.

В диссертации приведены выражения для ПЛС $\varphi_D(s)$, $\varphi_M(s)$, $\varphi_N(s)$, $\varphi_{PI}(s)$, $\varphi_{PJ}(s)$, $\varphi_{PS}(s)$, $\varphi_{PA}(s)$ – это ПЛС случайного времени обработки записи (исходной, промежуточной, результирующей) в ресурсе: D - диске, M -ОП, N - шине, процессоре: PI - поиск и чтение из таблицы измерения с помощью индекса, PJ – построение кортежа декартова произведения записей измерений, PS – сравнение записей БД в процессоре при их сортировке, PA – агрегация записей (значений фактов).

Выражения (12)-(15) учитывают создание при выполнении запроса первичного индекса, состоящего из декартова произведения кортежей (записей) таблиц измерений и передаваемого всем процессорным узлам ПСБД.

Это позволяет избежать передачи большого объёма таблицы фактов при межпроцессорном обмене.

В предположении, что запись результирующего декартова произведения измерений остаётся в данном узле или передаётся в другие узлы с одинаковой вероятностью ($p_m = 1/m$ в формулах (15)), из (9) было получено выражение для математического ожидания времени выполнения запроса к ПСБД:

$$
\begin{aligned}
M = -\Psi'(0) = &[2\sum_{i=1}^{K} g_i p_i + n^{K-1}\prod_{i=1}^{K} g_i p_i]\bar{\varphi}_D + \\
&[2(n-1)g_1 p_1 + 2n\sum_{i=2}^{K-1} n^{i-1}\prod_{j=1}^{i} g_j p_j + 4\sum_{i=1}^{K} g_i p_i + \frac{6n-2}{n} n^{K-1}\prod_{i=1}^{K} g_i p_i]\bar{\varphi}_M + \\
&[(n-1)g_1 p_1 + (n-1)\sum_{i=2}^{K-1} n^{i-1}\prod_{j=1}^{i} g_j p_j + \frac{n-1}{n} n^{K-1}\prod_{i=1}^{K} g_i p_i]\bar{\varphi}_N + \\
&[\sum_{i=2}^{K-1} n^{i-1}\prod_{j=1}^{i} g_j p_j + n^{K-1}\prod_{i=1}^{K} g_i p_i]\bar{\varphi}_{PJ} + [2\sum_{i=1}^{K} g_i p_i + n^{K-1}\prod_{i=1}^{K} g_i p_i]\bar{\varphi}_{PI} + \\
&[n^{K-1}\prod_{i=1}^{K} g_i p_i]\bar{\varphi}_{PS} + [n^{K-1}\prod_{i=1}^{K} g_i p_i]\bar{\varphi}_{PA} = \\
&Q_D\bar{\varphi}_D + Q_M\bar{\varphi}_M + Q_N\bar{\varphi}_N + Q_{PJ}\bar{\varphi}_{PJ} + Q_{PI}\bar{\varphi}_{PI} + Q_{PS}\bar{\varphi}_{PS} + Q_{PA}\bar{\varphi}_{PA}
\end{aligned}
\qquad (16)
$$

где $\bar{\varphi}_D$ и т.д. определяются следующим образом: если соответствующий ресурс является "узким местом", то это математическое ожидание времени пребывания в соответствующем разделяемом ресурсе (ожидание и обработка записи БД), в противном случае это математическое ожидание времени только обработки записи в ресурсе (для неразделяемого ресурса и для разделяемого ресурса, который не является "узким местом"),

$g_i = G_i^{'}(1) = V_i / n$ - математическое ожидание числа записей таблицы i-го измерения, которые хранятся в узле,

p_i - вероятность, что запись i-го измерения удовлетворяет условию поиска по этому измерению в запросе.

Помимо архитектур SE, SD и SN на практике применяют смешанные архитектурные решения. Например, узлы SMP (SE) соединяются по схеме "точка-точка" типа SN. В этом случае получается архитектура типа CE. В этом случае формулу (16) можно переписать в следующем виде:

$$
\begin{aligned}
M = &Q_D\bar{\varphi}_D(n_{PR}) + Q_M\bar{\varphi}_M + Q_N(n_{PR}, \Gamma(n))\bar{\varphi}_{NSMP} + \\
&(Q_N - Q_N(n_{PR}, \Gamma(n)))\bar{\varphi}_N + Q_{PJ}\bar{\varphi}_{PJ} + Q_{PI}\bar{\varphi}_{PI} + Q_{PS}\bar{\varphi}_{PS} + Q_{PA}\bar{\varphi}_{PA}
\end{aligned}
\qquad (17)
$$

здесь предполагается, что в SMP-узле "узким местом" является диск,

n_{PR} – число процессоров в одном SMP-узле, n= $n_{SMP} \cdot n_{PR}$ - общее число процессоров в системе, n_{SMP} - число SMP-узлов, Γ(n) – вектор ($g_1(n)$, ..., $g_K(n)$),

$\bar{\varphi}_{NSMP} = 1/\mu_{NSMP}$ - среднее время передачи записи между процессорами одного SMP-узла (чтение из ОП, передача записи (сообщения) по межпроцессорной шине, запись в ОП), остальные обозначения такие же, как и в формуле (16).

Для тех проектов построения информационных систем, для которых важен экономический эффект, должна выбираться архитектура системы с минимальной совокупной стоимостью владения. Совокупная Стоимость Владения (ССВ) - это методика расчета, разработанная зарубежными экономистами, чтобы помочь потребителям и руководителям предприятий определить прямые и косвенные затраты и выгоды, связанные с любым компонентом компьютерной системы.

В работе выполнена оценка ССВ ПСБД, состоящей их нескольких SMP-систем (рис. 2). На рис. 2 введены следующие обозначения: nPR– число процессоров в одной SMP-системе, nSMP– число SMP-систем, NR= N/nSMP – число дисков, закреплённых за одной SMP- системой. Такая конфигурация позволяет исследовать следующие архитектуры:

- SE (одна SMP-система), nSMP=1,
- CE (кластер SMP-систем), nPR > 1 и nSMP>1,
- SN (система с одним процессором в узле), nPR=1, nSMP>1.
- SE-кластер (кластер SMP-систем с общей дисковой памятью), все n процессоров разделяют все N дисков (n= nPR·nSMP).

Формулы для оценки стоимости ПСБД определяются особенностями зависимости стоимости системы от числа процессоров и числа дисков в дисковом массиве. Например, известно, что стоимость одной SMP-системы плавно зависит от числа процессоров nPR до некоторого nPR_{KP}.

Для сравнительной оценки стоимости различных архитектур параллельных систем баз данных предлагается использовать оценку затрат ежемесячного ССВ комплекса на протяжении пяти лет без модернизации комплекса с выделением следующих компонентов ПСБД: SMP-узлов, системы хранения и коммутационной сети.

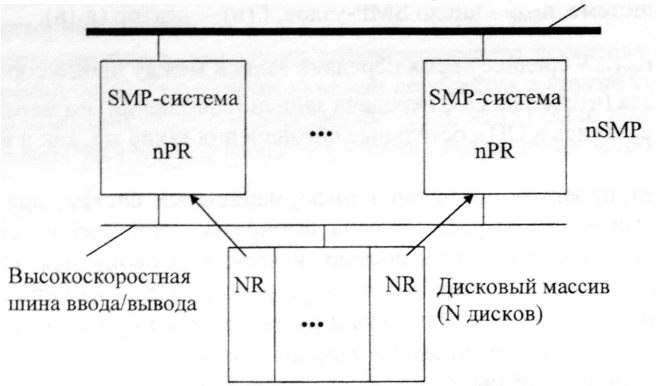


Рис. 2. Общая схема комплекса, состоящего из нескольких SMP- систем

Оценка ССВ определяется по формуле (руб./мес.):

$$C_{ССВ.ПСУБД.М} = \frac{C_{СХД}(N) + nSMP \times C_{SMP}(nPR) + C_{SW}(nSMP) + C_{О.ДР} + N_{CPU} \times C_{ПО}}{60} + \frac{(C_{СХД.Эл}(N) + nSMP \times C_{SMP.Эл}(nPR) + C_{SW.Эл}(nSMP) + C_{Эл.Др} + C_{СХД.Конд}(N) + nSMP \times C_{SMP.Конд}(nPR) + C_{SW.Конд}(nSMP) + C_{Конд.Др} + K_{CPU} \times C_{Сервис.ПО})}{12}, \quad (18)$$

где $C_{СХД}(N)$ – стоимость системы хранения данных, зависящая от числа дисков и дисковых полок в системе хранения, $C_{SMP}(nPR)$ – стоимость SMP-сервера с количеством процессоров nPR, $C_{SW}(nSMP)$ – стоимость коммутатора сети Net на nSMP узлов в системе, $C_{О.ДР}$ – стоимость дополнительного оборудования в комплексе, $\{C_{.Эл}(.)\}$ – составляющие стоимости электроснабжения системы в год, $\{C_{.Конд}(.)\}$ – составляющие стоимости теплоотвода от системы в год,

В работе разработан алгоритм выбора архитектуры ПСБД (алгоритм ВАПСБД), учитывающий специфику сравнения архитектур ПСБД и особенности стоимостной оценки:

Шаг 1. Рассчитать число дисков в RAID-массиве. Расчёт числа дисков проводится по формуле (19).

$$N = \left\lceil \frac{Q}{Q_D \cdot p_D} \times k_{RAID} \right\rceil + 2 \cdot k_{enc}, \quad (19)$$

где Q – общий объём хранимых данных (фактов и измерений), Q_D – объём диска, p_D – доля заполнения диска, k_{RAID} – коэффициент, учитывающий использование технологии RAID для защиты данных от физического отказа дисков, $2 \cdot k_{enc}$ – коэффициент, учитывающий использование технологии горячего резервирования дисков (hot spare).

Шаг 2. Оценить стоимость дискового массива $C_{СХД}(N)$.

Шаг 3. Проанализировать запросы к хранилищу данных. Для каждого i-го запроса

1) определить количество измерений, по которым выполняется поиск (K_i),

2) оценить число записей таблиц измерений в запросе ($VP_{ij} = V_{ij} \cdot p_{ij}$, $j = 1...K_i$),

3) рассчитать среднее значение $VP_i = \sqrt[K_i]{\prod_{j=1}^{K_i} VP_{ij}}$.

Эти данные занести в таблицу и назначить граничные значения для среднего времени выполнения этих запросов.

Шаг 4. Положить nPR=1 и nSMP=1.Это соответствует самой дешёвой конфигурации (одна SMP-система с одним процессором).

Шаг 5. Рассчитать среднее время (M) для всех запросов, используя формулу (17), для SE-кластера $\bar{\varphi}_D$ зависит от общего числа процессоров n= nPR·nSMP. Если для какого-либо запроса время его выполнения превышает граничное значение, то перейти к шагу 6, иначе перейти к шагу 8.

Шаг 6. Проверить nPR: если для текущего значения nPR перегружается диск массива RAID (дальнейшее увеличение nPR не приведёт к уменьшению времени выполнения запросов) или $nPR > nPR_{КР}$, то перейти к шагу 7, иначе увеличить число процессоров в каждой SMP-системе: nPR:=nPR+1, перейти к шагу 5.

Шаг 7. Увеличить число SMP-систем: nPR:=1, nSMP:=nSMP+1. Если $nSMP > nSMP_{ГР}$, то выйти из алгоритма (решение не найдено, заданы слишком жёсткие ограничения на время выполнения запросов), иначе перейти к шагу 5.

Шаг 8. Полученная конфигурация (nPR. nSMP) является оптимальной, оценить ССВ архитектуры ПСБД по формуле (18). Завершить алгоритм (решение найдено).

В алгоритме последовательно наращивается число процессоров (nPR) и SMP-систем (nSMP), и таким образом параллельные системы баз данных с архитектурами SE, CE, SN или SE-кластер упорядочиваются по возрастанию их стоимости. Так как число шагов ограничено, то оптимальное решение или будет найдено, или нет.

В **четвертой главе** «Использование разработанных методов анализа для выбора архитектуры хранилища гидрометеорологических данных» при-

ведены результаты применения предлагаемых моделей при выборе архитектуры ПСБД хранилища гидрометеорологических данных.

Основной задачей построения хранилища гидрометеорологических данных является сбор, обработка и хранение гидрометеорологических данных в едином хранилище с целью их дальнейшей аналитической обработки и использования для нужд геоинформационных систем. Первым этапом построения хранилища была разработка схемы хранилища для проведения оперативного анализа основных показателей климатических данных. Основные климатические показатели определены в постановке задачи и должны быть извлечены из массива поступающих метеорологических данных.

Ежегодное поступление информации в хранилище составляет около 9Тб климатических данных. В дисковом массиве предполагается хранить Q=61,4 Тб данных с последующим архивированием старых сведений. Технические и программные средства системы: Sun Oracle Database Machine (с ячейками Exadata).

Для выбора архитектуры ПСБД хранилища гидрометеорологических данных был применен алгоритм, разработанный в третьей главе (алгоритм ВАПСБД для SE-кластера). Исходя из требований к системе, получено общее число дисков в дисковых массивах N=320. Проанализированы три типа наиболее критичных запросов к хранилищу данных без материализации представлений (табл. 1), $T_{ГР}$ - граничное значение для среднего времени выполнения запроса.

Таблица 1.

Критические запросы к хранилищу

Краткое описание запроса	VP	K	$T_{ГР}$ (с.)
Тип запроса №1. К трем измерениям. Получить усредненные и достоверные значения минимальной, максимальной, средней температуры, температуры точки росы с выводом места сбора информации, хранения и вида наблюдения за определенный промежуток времени	185	3	60
Тип запроса №2. К пяти измерениям. Получить достоверное среднее значение температуры, давления, количества выпавших осадков и продолжительности солнечного сияния с выводом станций наблюдений, мест хранения, видов наблюдений и типов носителей за определенный промежуток времени	24	5	60
Тип запроса №3. К семи измерениям. Предоставить усредненные и достоверные значения минимальной, максимальной, средней температуры, продолжительности солнечного сияния, количества выпавших осадков, скорости ветра, общей облачности, температуры почвы с выводом станций наблюдений, мест хранения, видов наблюдений, типов носителей, видов осадков, типов облаков и направления ветра	9	7	60

По результатам работы алгоритм ВАПСБД была выбрана оптимальная архитектура ПСБД (nSMP=3, nPR=4). Для этой конфигурации была рассчитана оценка ССВ по формуле (18), которая составила 7 431 096,48 руб./мес.

На рис. 3 приведены графики зависимостей среднего времени выполнения запросов от количества SMP-систем (nSMP).

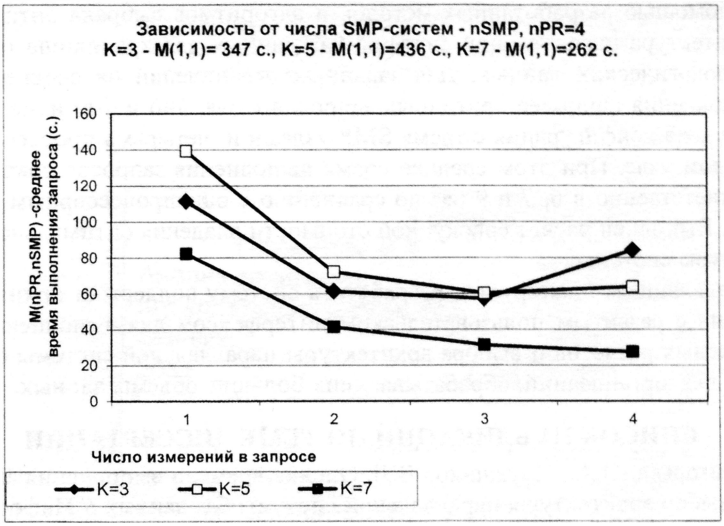


Рис. 3. Зависимость среднего времени выполнения запросов, от количества SMP-систем

Для оптимальной конфигурации были получены следующие значения среднего времени выполнения запросов: запрос №1 – 57 сек., запрос №2 – 60 сек., запрос №3 – 31 сек. Причём для запросов №1, 2 (K=3, K=5) это время является минимальным. Это объясняется большим числом кортежей в декартовом произведении записей измерений (185^3, 24^5) и тем, что при дальнейшем увеличении количества SMP-систем увеличивается общее число процессоров, которые перегружают дисковую систему. По сравнению с однопроцессорной системой среднее время выполнения запросов сократилось почти в 6 (K=3), 7 (K=5) и 9 (K=7) раз.

ВЫВОДЫ

1. Разработаны модели обработки запросов к одной таблице в параллельной системе баз данных в виде замкнутой и разомкнутой СМО, учитывающие основные особенности разных архитектурных решений.
2. Предложен математический метод оценки времени выполнения запросов к нескольким таблицам СУБД для различных архитектур и способов реализации соединений этих таблиц.
3. Разработан математический метод оценки времени выполнения запросов к хранилищу данных, учитывающий особенности реализации плана соеди-

нения таблиц измерений и фактов в параллельной системе баз данных.

4. Предложен способ оценки совокупной стоимости владения для рассматриваемого класса систем в зависимости от конфигурации комплекса.
5. Разработан алгоритм выбора архитектуры параллельной системы баз данных, основанный на упорядочивании вариантов системы по возрастанию их стоимости.
6. С помощью разработанных методов и алгоритмов выбрана оптимальная архитектура параллельной системы баз данных для хранилища гидрометеорологических данных. Для заданных ограничений на среднее время выполнения наиболее критичных запросов с 3-я, 5-ю и 7-ю измерениями получена конфигурация с тремя SMP-узлами и четырьмя процессорами в каждом узле. При этом среднее время выполнения запросов сократилось соответственно в 6, 7 и 9 раз по сравнению с однопроцессорным вариантом. Выполнен расчет совокупной стоимости владения оптимальной архитектуры системы.
7. В дальнейшем планируется разработать систему поддержки принятия решения с развитым пользовательским интерфейсом для выполнения комплексных расчетов и выбора архитектуры параллельной системы баз данных для организаций, обрабатывающих большие объёмы данных.

СПИСОК ПУБЛИКАЦИЙ ПО ТЕМЕ ДИССЕРТАЦИИ

1. Григорьев Ю.А., Плужников В.Л. Оценка времени выполнения запросов и выбор архитектуры параллельной системы баз данных // Информатика и системы управления. – 2009. - № 3. – С. 3-12.
2. Григорьев Ю.А., Плужников В.Л. Модель обработки запросов в параллельной системе баз данных // Вестник МГТУ им. Н.Э. Баумана. Приборостроение. – 2010. - № 4. – С. 78-90.
3. Григорьев Ю.А., Плужников В.Л. Оценка времени соединения таблиц в параллельной системе баз данных // Информатика и системы управления. – 2011. - № 1. – С. 3-16.
4. Григорьев Ю.А., Плужников В.Л. Анализ времени обработки запросов к хранилищу данных в параллельной системе баз данных // Информатика и системы управления. – 2011. - № 2. – С. 94-106.
5. Плужников В.Л. Оценка времени выполнения запроса в параллельной системе баз данных // Наука и образование. - Электрон. журн. – 2011. - №6. – [Электронный ресурс]. URL: http://technomag.edu.ru/doc/188065.html (дата обращения: 25.10.2011).
6. Плужников В. Л. Метод выбора архитектуры параллельной системы баз данных // Проблемы построения и эксплуатации систем обработки информации и управления: Сб. статей / Под ред. В.М. Чёрненького. – 2011. - Вып. 9. – С. 76-83.