электронное научно-техническое издание
НАУКА и ОБРАЗОВАНИЕ
Эл № ФС 77 - 30569. Государственная регистрация №0421100025. ISSN 1994-0408

Анализ процессов выполнения запросов в параллельных системах баз данных с архитектурами SE, SD, SN

77-30569/387243

04, апрель 2012

Плужников В. Л., Гасов В. М.

УДК 004.657

МГТУ им. Н.Э. Баумана
chernen@bmstu.ru

Обоснование разработки модели анализа

В настоящее время сравнительный анализ архитектурных решений выполняется или на основе экспертных оценок качественных критериев (масштабируемости, доступности данных, баланса загрузки, межпроцессорных коммуникаций, когерентности кэшей, организации блокировок и др. [2]), или на основе результатов тестов (TPC и др.) для конкретных платформ. Оба эти способа имеют недостатки.

Экспертные оценки зачастую носят субъективный характер, эксперту трудно оценить количественные показатели будущей системы (индексы производительности, параметры надёжности). Недостатками второго способа является то, что результаты тестирования получаются на основе стендовых испытаний с эталонной моделью нагрузки и при выполнении конкретных программ (тестов). Используемые тесты являются синтетическими и не учитывают специфику предметной области, для которой выбирается архитектура системы. Поэтому результаты тестовых сравнений архитектур довольно проблематично использовать при принятии решения о выборе архитектуры параллельной системы базы данных для конкретной предметной области.

Более того, ни один из перечисленных методов не позволяет получить количественные показатели производительности для заданной системы.

Таким образом, возникает необходимость разработки нового аналитического метода оценки характеристик производительности ПСБД, который должен отвечать следующим требованиям:

1. Учитывать особенности предметной области моделируемой системы, т.е. позволять оценивать временные показатели выполнения запросов к ПСБД (запрос к одной таблице, запрос к нескольким таблицам, запрос к хранилищу данных).

2. Учитывать особенности различных архитектур параллельной системы баз данных.

Метод должен предлагать аналитическое решение, не требующее высоких вычислительных мощностей для его реализации. В статье рассматриваются следующие архитектуры ПСБД:

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.

Модель выполнения запросов в параллельной системе базы данных

На рис. 1 и 2 представлены модели обработки запроса к ПСБД. Архитектуры параллельных систем баз данных подробно описаны в [2]. Ниже изложение ведётся для запроса к одной таблице, хотя эта модель с некоторыми изменениями может быть использована и для анализа запроса к нескольким таблицам. На рисунках приняты следующие обозначения:

- дисциплины обслуживания: PS – Processor Sharing (разделяемый ресурс), IS – Immediately Served (ресурс без очереди),
- L – число записей в блоке БД,
- P_F – вероятность, что запись удовлетворяет условию поиска F,
- 1 – чтение блока БД с L записями с диска RAID-массива в кэш диска,
- 2 – перезапись блока с L записями БД из кэша диска в ОП (интенсивность λL), чтение L записей из ОП в кэш процессора (λL), сохранение записей (маршаллинг для SE), удовлетворяющих условию поиска F, в ОП для слияния их на выделенном процессоре ($\lambda P_F L$),
- 3 – обработка в процессоре L записей БД,
- 4 – передача записей (для SD и SN), удовлетворяющих условию поиска (с вероятностью P_F), выделенному процессору для слияния результатов (межпроцессорный обмен).

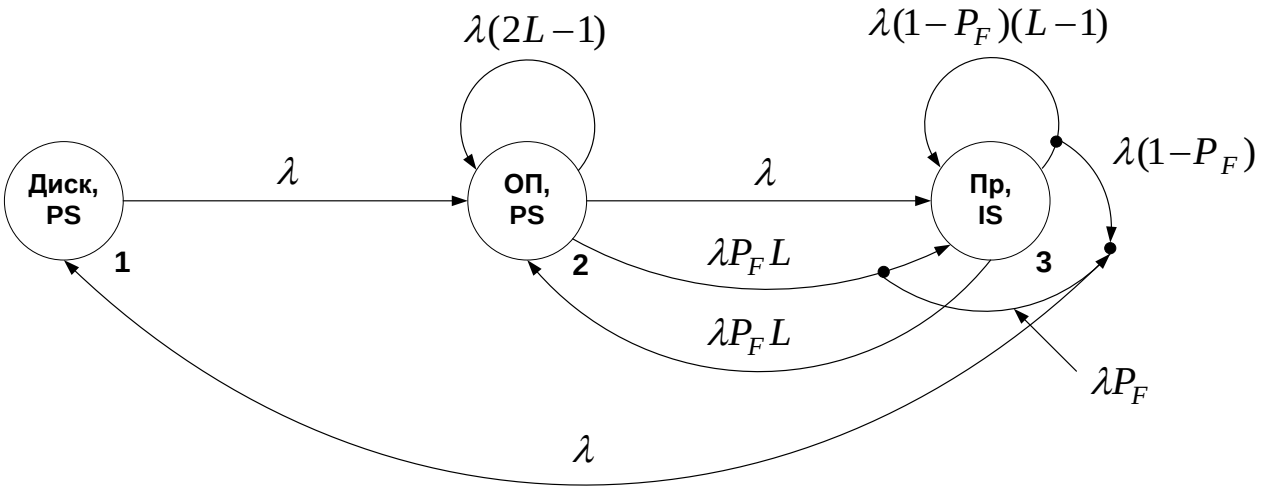

Рис. 1. Модель обработки запроса в параллельной системе баз данных с архитектурой SE.

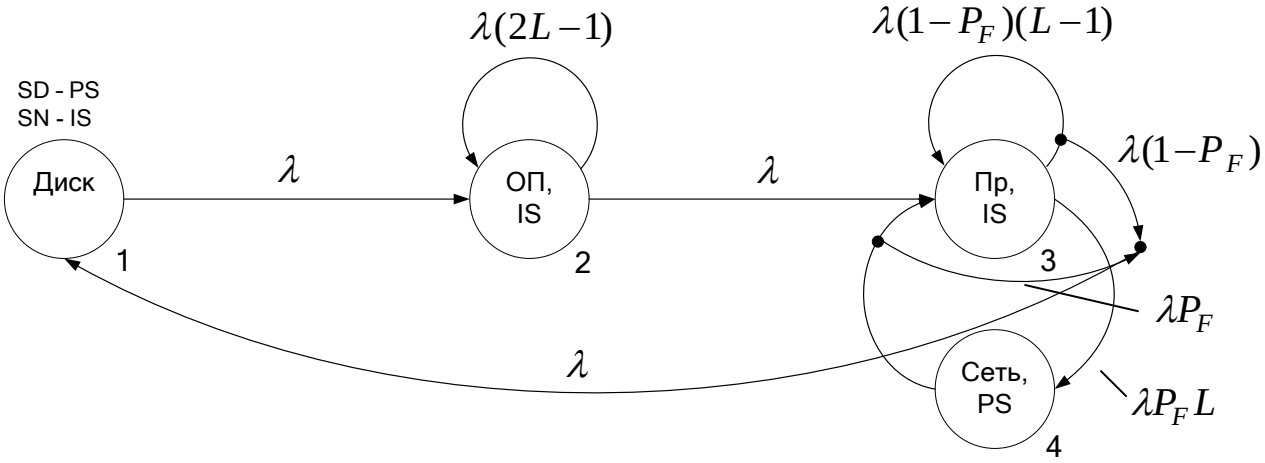

Рис. 2. Модель обработки запроса в параллельной системе баз данных с архитектурами SD и SN.

Как видно из рис. 1 и 2, модель обработки запроса к БД представляет собой замкнутую СМО с различными дисциплинами обслуживания в узлах ("случайная выборка из очереди" – Ethernet на шине, "лифтовый поиск" - в SCSI-диске и т.д.). В модели циркулируют n заявок, которые соответствуют процессорам системы. Точный метод расчёта индексов производительности с помощью этой модели имеет ряд недостатков:

- расчёты по этой модели достаточно сложны для большого числа процессоров n
- результаты анализа нельзя представить в виде простых аналитических формул, с помощью которых можно было бы построить графики зависимостей и сравнить варианты решений.

Ниже предполагается, что разработанные замкнутые СМО являются экспоненциальными (обоснование см. ниже). Чтобы упростить расчёты и сделать их более наглядными, ниже предлагается использовать метод "узкого места". Пусть i-й и j-й узлы – это ресурсы с очередью в замкнутой СМО. Тогда из [11, формула (1.34)] имеем

$$\frac{\lambda_i}{\mu_i} >> \frac{\lambda_j}{\mu_j} \Rightarrow Q_i >> Q_j \quad , \qquad (1)$$

где λ_i, λ_j - интенсивности входных потоков узлов, μ_i, μ_j - интенсивности обслуживания заявок в этих узлах, Q_i, Q_j - среднее число заявок в узлах.

В случае выполнения неравенства (1) для всех $j \neq i$ (т.е. при наличии "узкого места") модели на рис. 1 и 2 можно свести к двухузловой замкнутой СМО (рис. 3) с композиционным центром (1) и i-ым разделяемым ресурсом (2). Здесь "a" и "b" - время обработки в узлах, n - число циркулирующих заявок (процессоров).

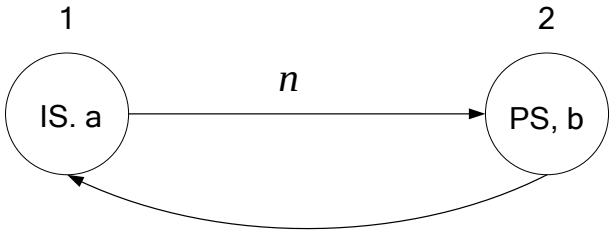

Рис. 3. Двухузловая замкнутая СМО с композиционным центром.

Рассматривая модели на рис. 1 и 2 как дискретные марковские цепи в моменты выхода заявки из узлов, можно рассчитать параметры модели, представленной на рис. 3 (табл. 1).

Таблица 1.

Параметры двухузловой замкнутой СМО с композиционным центром

Архитектура	Условие	"Узкое место"	Параметр "a" модели	Параметр "b" модели
SE	$\frac{1}{\mu_D} >> \frac{2+P_F}{\mu_M}$	Диск	$(\frac{2+P_F}{\mu_M} + \frac{1}{\mu_P})L$	$\frac{1}{\mu_{DB}}$
	$\frac{2+P_F}{\mu_M} >> \frac{1}{\mu_D}$	ОП	$\frac{1}{2+P_F}(\frac{1}{\mu_P} + \frac{1}{\mu_D})$	$\frac{1}{\mu_M}$
SD	$\frac{1}{\mu_D} >> \frac{P_F}{\mu_N}$	Диск	$(\frac{2}{\mu_M} + \frac{1}{\mu_P} + \frac{P_F}{\mu_N})L$	$\frac{1}{\mu_{DB}}$
	$\frac{P_F}{\mu_N} >> \frac{1}{\mu_D}$	Сеть	$\frac{1}{P_F}(\frac{1}{\mu_P} + \frac{1}{\mu_D} + \frac{2}{\mu_M})$	$\frac{1}{\mu_N}$
SN	нет	Сеть	$\frac{1}{P_F}(\frac{1}{\mu_P} + \frac{1}{\mu_D} + \frac{2}{\mu_M})$	$\frac{1}{\mu_N}$

В табл. 1 приняты следующие обозначения:

μ_D – интенсивность чтения записей БД с диска RAID-массива; $\mu_D = \mu_{DB} \cdot L$, где μ_{DB} - интенсивность чтения блоков БД с диска,

μ_M - интенсивность чтения/сохранения записей БД в ОП,

μ_N - интенсивность передачи записей БД по сети (межпроцессорный обмен),

μ_P - интенсивность обработки записей БД в процессоре.

Сведение замкнутой двухузловой СМО к разомкнутой

Модель на рис. 3 проще, чем модели, представленные на рис. 1 и 2. Но она имеет существенный недостаток: результаты анализа нельзя представить в виде простых аналитических формул, с помощью которых можно было бы сравнить варианты решений.

Рассмотрим два случая (см. рис. 3).

1. Загрузка ресурса 2 большая. В этом случае интенсивность выходного потока примерно равна $1/b$. Тогда средняя длина очереди в разделяемом ресурсе 2 равна $k = n - a/b$. Отсюда получим

$$\frac{nb}{a} = \frac{n}{n-k} > 1. \qquad (2)$$

Для этого случая можно получить интересный вывод. Пусть $b > a$ (разделяемый ресурс 2 медленный). Тогда $Q_2 > Q_1$, где Q_2 и Q_1 - среднее число заявок в ресурсах 2 и 1. Это следует из следующего утверждения [11, формула (1.34)]

$$b^i > \frac{a^i}{i!},\ i = 1 \ldots n \Rightarrow Q_2 > Q_1 \quad . \tag{3}$$

Следовательно, $Q_2 = \tau n$, $\frac{1}{2} < \tau < 1$. Отсюда получим оценку для среднего времени обработки записей таблицы базы данных

$$T = (bQ_2 + a)\frac{S}{n} = Sb(\tau + \frac{a}{nb}) \quad , \tag{4}$$

где S - либо число блоков в таблице ("узкое место" – диск), либо число записей в таблице ("узкое место" – ОП или сеть).

Из (4) следует, что T слабо зависит от числа процессоров n. Отсюда можно сделать следующий вывод: если в системе имеется медленный разделяемый ресурс, то распараллеливание выполнения запроса по нескольким процессорам не приведёт к существенному уменьшению времени обработки этого запроса к БД.

2. Загрузка ресурса 2 небольшая. Интенсивность выходного потока равна P/b, где P - вероятность, что разделяемый ресурс 2 занят. Отсюда для достаточно больших n имеем

$$\rho = \frac{nb}{a} = \frac{Pn}{n-k} < 1 \quad . \tag{5}$$

Известно, что для СМО на рис. 3 справедливо следующее распределение вероятностей числа заявок в разделяемом ресурсе 2 [12]:

$$P_i = P_0 (\frac{nb}{a})^i \frac{(n-i+1)\cdot(n-i+2)\ldots n}{n^i} \quad , 1 \le i \le n \, . \tag{6}$$

При малых i и больших n

$$P_i \to P_0 (\frac{nb}{a})^i \quad . \tag{7}$$

Если выполняется условие (5), то $P_0 \to 1 - \rho = 1 - \frac{nb}{a}$. Ошибка вычисления P_i по формуле (7) возрастает при увеличении i, но в этом случае P_i мало.

Но распределение (7) справедливо для разомкнутой СМО M/M/1.

Интенсивность входного потока для M/M/1 определяется по формуле

$$\Lambda = n\lambda, \quad \lambda = \frac{1}{a}, \tag{8}$$

где параметр a рассчитывается по формулам, приведённым в табл. 1. Ниже будем использовать следующие обозначения для интенсивностей входного потока:

λ_{DB} - интенсивность заявок на чтение блоков БД с диска от одного процессора,

$\lambda_D = \lambda_{DB} \cdot L$ - интенсивность заявок на чтение записей БД с диска от одного процессора,

λ_M - интенсивность заявок на чтение/сохранение записей БД в ОП от одного процессора,

λ_N - интенсивность заявок на передачу записей БД по сети межпроцессорного обмена от одного процессора.

Для определения времени пребывания в разделяемом ресурсе целесообразно использовать модель M/M/1. В частности это объясняется несколькими причинами:

1. Для замкнутых СМО даже в предположении их экспоненциальности не существует простых аналитических формул, позволяющих оценивать характеристики системы [11]. Сложность расчёта замкнутых СМО существенно возрастает с увеличением числа процессоров, т.е. 'n'.

2. Как правило, в параллельной системе баз данных присутствует не более одного разделяемого ресурса, который можно охарактеризовать как "узкое место" (чаще всего это внешняя память). В этом случае можно показать, что все остальные ресурсы можно считать неразделяемыми. Если предположить, что время обработки в оставшемся разделяемом ресурсе распределено по экспоненциальному закону, то распределение вероятностей числа требований в обслуживающих аппаратах (ОА) замкнутой СМО зависит от средних значений времени обработки в неразделяемых ресурсах, т.е. не зависит от вида функций распределения (ф. р.) времени обслуживания в этих ресурсах [11]. Поэтому можно считать, что время обработки требований в неразделяемых ресурсах распределено по экспоненциальному закону. Таким образом, замкнутая модель сводится к классической модели "ремонтника", которая при достаточно большом 'n' близка по характеристикам к разомкнутой СМО M/M/1 [12].

3. В работах [12, 13] показано, что для разомкнутой СМО GI/GI/1 без прерывания обработки справедлив закон сохранения работы. В этом случае среднее время ожидания в очереди не зависит от дисциплины обслуживания этой очереди. Т. е. для дисциплин LIFO ("последний пришёл, первый обслужен") , RS ("случайная выборка из очереди" – Ethernet на шине), с пакетной обработкой (LIFO или RS внутри пакета – "лифтовый

поиск" в SCSI-диске), с относительным приоритетом (и др.) можно для расчёта среднего времени пребывания использовать формулы для дисциплины FIFO ("первый пришёл, первый обслужен").

4. Если использовать СМО M/M/1, то для перехода от разделяемого ресурса к неразделяемому достаточно в выражении для ПЛС убрать произведение 'nλ' (см. раздел 2.3, табл. 6).

5. Реально для модели ресурса можно оценить только параметры λ и μ. Поэтому использовать для расчётов другую разомкнутую СМО, отличную от M/M/1, проблематично.

Можно сформулировать стратегию вычислений индексов производительности параллельной системы базы данных, построенной на основе какого-либо архитектурного решения:

1. Выявить "узкое место" системы (см. табл. 1).

2. Если выполняется неравенство (2), то считать архитектуру параллельной системы базы данных неудачной и перейти к другому архитектурному решению.

3. Если выполняется неравенство (5), то для расчётов модели, приведённой на рис. 3, использовать разомкнутую СМО M/M/1 с параметрами, которые указаны в табл. 1.

ЛИТЕРАТУРА

1. М. Тамер Оззу, Патрик Валдуриз. Распределенные и параллельные системы баз данных: [Электронный ресурс]. [http://citforum.ru/database/classics/distr_and_paral_sdb/]. Проверено 26.11.2010.
2. Соколинский Л. Б., Цымблер М. Л. Лекции по курсу «Параллельные системы баз данных”: [Электронный ресурс]. [http://pdbs.susu.ru/CourseManual.html]. Проверено 04.12.2010.
3. Дж. Льюис. Oracle. Основы стоимостной оптимизации. – СПб: Питер, 2007. – 528 с.
4. 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.
5. Производительность СУБД Oracle Database 11g при работе на сервере Sun SPARC Enterprise M9000: [Электронный ресурс]. [http://ru.sun.com/sunnews/press/2010/2010-05-18.jsp]. Проверено 26.11.2010
6. В.А. Варфоломеев, Э.К. Лецкий, М.И. Шамров, В.В. Яковлев. Лекции по курсу “Операционные системы и программное обеспечение на платформе zSeries”: [Электронный ресурс]. [http://www.intuit.ru/department/os/ibmzos/]. Проверено 26.11.2010.
7. Керри Болинджер. Врожденный параллелизм: [Электронный ресурс]. [http://www.osp.ru/os/2006/02/1156526/]. Проверено 04.05.2012.
8. Лев Левин. Teradata совершенствует хранилища данных: [Электронный ресурс]. [http://www.pcweek.ru/themes/detail.php?ID=71626]. Проверено 26.11.2010.

9. Oracle Real Application Clusters Administration and Deployment Guide 11g Release 1 (11.1): [Электронный ресурс]. [http://download.oracle.com/docs/cd/B28359_01/rac.111/ b28254/admcon.htm/]. Проверено 26.11.2010.
10. Григорьев Ю.А., Плутенко А.Д. Теоретические основы анализа процессов доступа к распределённым базам данных. – Новосибирск: Наука, 2002. – 180 с.
11. Жожикашвили В.А, Вишневский В.М. Сети массового обслуживания. Теория и применение к сетям ЭВМ. – М.: Радио и связь, 1988. – 192 с.
12. Клейнрок Л. Теория массового обслуживания. – М.: Машиностроение, 1979. – 432 с.
13. Бронштейн О.И., Духовный И.М. Модели приоритетного обслуживания в информационно-вычислительных системах. – М.: Наука, 1976. – 220 с.
14. Форум/Использование СУБД/Oracle/CPUSPEED на IntelXeon 5500 (Nehalem): [Электронный ресурс]. [http://www.sql.ru]. Проверено 02.12.2010.

electronic scientific and technical periodical
SCIENCE and EDUCATION
EL № FS 77 - 30569. №0421100025. ISSN 1994-0408

Analysis of query execution processes in parallel database systems with SE, SD, SN architectures.

77-30569/387243

04, April 2012

Pluzhnikov V.L., Gasov V.M.

Bauman Moscow State Technical University
chernen@bmstu.ru

The authors propose an approach to estimation of time characteristics of query execution to a parallel database system (PDBS). A model of query execution to PDBS with a p F(R) plan as a closed-loop queue system was developed. It was shown that the model could be reduced to a “repairman” model if there was a bottleneck. The possibility of reducing this model to an open-loop queue system is proved. Parameters of this model are determined. A policy of calculation of performance indexes of PDBS built on any architecture (SE, SD, SN) is formulated.

Publications with keywords: average of distribution time for query execution, parallel system of database, Laplase-Stieltjes transformation, SE, SD, SN architectures
Publications with words: average of distribution time for query execution, parallel system of database, Laplase-Stieltjes transformation, SE, SD, SN architectures

References

1. Tamer Ozsu M., Patrick Valduriez. *Distributed and parallel database systems.* (Russ. version: Tamer Ozzu M., Patrik Valduriz. *Raspredelennye i parallel'nye sistemy baz dannykh*. Available at: http://citforum.ru/database/classics/distr_and_paral_sdb/ , accessed 26.11.2010.).

2. Sokolinskii L. B., Tsymbler M. L. *Lektsii po kursu «Parallel'nye sistemy baz dannykh”* [Lectures on the course “Parallel database systems”]. Available at: http://pdbs.susu.ru/CourseManual.html, accessed 04.12.2010.

3. Lewis J. *Cost-Based Oracle Fundamentals (Expert's Voice in Oracle).* Apress, 2005. 520 p. (Russ. ed.: Dzh. L'iuis. *Oracle. Osnovy stoimostnoi optimizatsii*. SPb., Piter, 2007. 528 p.).

4. Grigor'ev Iu.A., Pluzhnikov V.L. Otsenka vremeni vypolneniia zaprosov i vybor arkhitektury parallel'noi sistemy baz dannykh [Score a run-time of query and selection of architecture of the parallel database systems]. *Informatika i sistemy upravleniia*, 2009, no. 3, pp. 3-12.

5. *Proizvoditel'nost' SUBD Oracle Database 11g pri rabote na servere Sun SPARC Enterprise M9000* [Performance of Oracle Database 11g database while working on the server Sun SPARC Enterprise M9000]. Available at: http://ru.sun.com/sunnews/press/2010/2010-05-18.jsp, accessed 26.11.2010

6. Varfolomeev V.A., Letskii E.K., Shamrov M.I., Iakovlev V.V. *Lektsii po kursu "Operatsionnye sistemy i programmnoe obespechenie na platforme zSeries"* [Lectures on the course "Operating systems and software on zSeries platform"]. Available at: http://www.intuit.ru/department/os/ibmzos/, accessed 26.11.2010.

7. Carrie Ballinger. *Born To Be Parallel* (Russ. version: Kerri Bolindzher. *Vrozhdennyi parallelizm*. Available at: http://www.osp.ru/os/2006/02/1156526/, accessed 04.05.2012.).

8. Levin L. *Teradata sovershenstvuet khranilishcha dannykh* [Teradata improves data warehouse]. Available at: http://www.pcweek.ru/themes/detail.php?ID=71626, accessed 04.05.2012.

9. *Oracle Real Application Clusters Administration and Deployment Guide 11g Release 1 (11.1)*. Available at: http://download.oracle.com/docs/cd/B28359_01/rac.111/b28254/admcon.htm, accessed 26.11.2010.

10. Grigor'ev Iu.A., Plutenko A.D. *Teoreticheskie osnovy analiza protsessov dostupa k raspredelennym bazam dannykh* [Theoretical basis of analysis of the processes of access to distributed databases]. Novosibirsk, Nauka, 2002. 180 p.

11. Zhozhikashvili V.A, Vishnevskii V.M. *Seti massovogo obsluzhivaniia. Teoriia i primenenie k setiam EVM* [Queueing network. Theory and application to computer networks]. Moscow, Radio i sviaz', 1988. 192 p.

12. Kleinrok L. *Queueing Systems. Vol. 1. Theory*. New York, John Wiley & Sons, 1975. (Russ ed.: Kleinrok L. *Teoriia massovogo obsluzhivaniia*. Moscow, Mashinostroenie, 1979. 432 p.).

13. Bronshtein O.I., Dukhovnyi I.M. *Modeli prioritetnogo obsluzhivaniia v informatsionno-vychislitel'nykh sistemakh* [Model of the priority services in the information-computing systems]. Moscow, Nauka, 1976. 220 p.

14. *Forum/Ispol'zovanie SUBD/Oracle/CPUSPEED na IntelXeon 5500 (Nehalem)* [Forum / Using the database/Oracle/CPUSPEED on IntelXeon 5500 (Nehalem)]. Available at: http://www.sql.ru, accessed 02.12.2010.\