Journal of
**Software
Engineering**

Research Article
Performance Analysis of Database Connection Pool Technology Based on Queueing Model with Mixed Waiting and Loss

Jun Wu and Zhanqiang Huo

School of Computer Science and Technology, Henan Polytechnic University, 454000 Jiaozuo, China

Abstract

Background and Objective: To reveal the dependency relationships between configuration parameters and performance measures of database connection pool technology, Geom/Geom/c/c+m discrete-time queuing model with multi-server and mixed waiting and loss is built according to its studying principles. **Materials and Methods:** Using the embedded Markov chain method, the one-step transition probability matrix and the recursive formula of this model are derived. The mathematical expressions of system performance measures are calculated by the theoretical analysis results of the model such as request blocking probability, system average waiting time, system utilization and system throughput etc. **Conclusion:** Finally, numerical examples are intuitively given to discuss the dependence relationships between system performance measures and configuration parameters. The study results can provide theoretical bases for the parameter configuration of database connection pool technology.

Key words: Database connection pool, performance analysis, discrete-time queuing theory, Geom/Geom/c/c+m model, mixed waiting and loss

Corresponding Author: Zhanqiang Huo, School of Computer Science and Technology, No. 2001, Shiji Road, 454000 Jiaozuo, China
Tel: +8603913987742

Competing Interest: The authors have declared that no competing interest exists.

Data Availability: All relevant data are within the paper and its supporting information files.

INTRODUCTION

Database connection as a key resource to interact with the database is particularly important in the applications especially, in web application of multi-user. Database connection pool technology is proposed to solve the problem of frequently creating and releasing database connection that reduces the system performance in software system. An important problem in database connection pool technology is how to configure the parameters of the connection pool in order to make the software system at the optimal performance.

Wang *et al.*[1] gives a quantitative data about the improvement of the performance of the database access based on the connection pool through performance contrast tests under three different kinds of database access modes. Li *et al.*[2] propose the optimal allocation method of database connection pool based on the log file records. Zhu and Shane[3] give us the configuration method in detail for database connection pool parameter configuration based on XML under J2EE framestudy. Liang *et al.*[4], Meng *et al.*[5], Zeng and Fu[6], Lv *et al.*[7] and Liu[8] make a study and have obtained some significant results from database connection pool technology.

The performance analysis methods of database connection pool technology can be divided into 3 categories: Experimental data analysis, simulation study method and mathematical model analysis methods. In this study, the mathematical model analysis method is used for the database connection pool technology to establish a discrete-time multi-server queuing model. The analysis of classical discrete-time queuing model is shown by Takagi, Tian and Zhang[9].

According to the study principles of database connection pool technology and considering to database connection cache mechanism, a Geom/Geom/c/c+m discrete-time queuing model with multi-server and mixed waiting and loss was established. Using an embedded Markov chain method,

the system performance measures of request blocking probability, system average waiting time, system utilization and system throughput were derived. The dependency relationships between the performance measures and the system parameters were shown intuitively through some numerical results. The study results provide a theoretical basis for the parameter configuration of database connection pool technology.

Management process of database connection pool: Database connection pool takes the database connection as the management entity. It well solved too much overhead and low response speed of the system and so on caused by frequent database access by the system. The process of the connection pool management is shown in Fig. 1:

- When a client requests a database connection, the first is to see if there are any more free connections in the pool. If there is a free connection then take out a free connection assigned to customers for using and mark the connection as busy connections at the same time as shown in Fig. 1
- If there is no free connection check whether the total number of busy connection has reached the maximum number of connections system settings. If it is not then create a connection to the request of the client as shown in Fig. 1. If it is then to set the maximum waiting time for waiting, if exceed the maximum waiting time throw an exception to the customer as shown in Fig. 1
- When at the end of the connection that the customer uses, the connection is returned to the connection pool in order to improve the reuse rate of connection and will mark the connection as free connections at the same time as shown in Fig. 1

In the management process of database connection pool, connecting a database as a service desk, regarding a
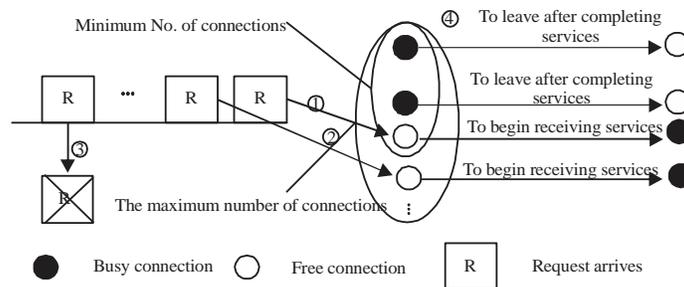

Fig. 1: Database connection pool management process

connection request as a customer arrive. Assuming that customer request arrival process is Bernolli arrival process, the connection using time obey the geometric distribution, considering the buffer in the connection pool, there is no free connections and cache area has reached maximum capacity, connection request is rejected, adopt the first come first service strategy was adopt, the management process of database connection pool can be modeled for multi service desk mixed the Geom/Geom/c/c+m discrete time queuing model.

Descriptions and analysis of queuing model

Model description: Time line is segmented into small pieces with fixed interval which is called slot. Arrival only happens at the terminal of each slot $t = n^-(n = 1,2\cdots)$ which is called late arrival system (Fig. 2).

For Geom/Geom/c/c+m discrete time queuing model with multiple server and mixed waiting and loss, the description is as follow:

- Late arrival system: customer arrival can only happen at (n^-, n) $n \geq 0$. The beginning and ending of service happen at (n, n^+). Customer arriving at (n^-, n) cannot depart at (n, n^+) any customer service time must across one length of slot
- At slot (n^-, n) $n \geq 0$, a customer arrives at probability p $(0<p<1)$ does not arrive at probability $\bar{p} = 1-p$, the arrival process is Bernoulli process with parameter p
- The customer service time obey the geometric distribution and independent identically distributed which is $p\{s = k\} = \bar{\mu}^{k-1}\mu, 0 <\mu< 1, k = 1,2,3...$
- The system has $c(c \geq 1)$ parallel server, with capacity is $m(1 \leq m \leq c)$ the queuing rule of queuing place FIFO
- Arrival interval and service time are independent of each other

If all the seats c+m were occupied when customers arrive and in the instantaneous there was no customers leave, customer's service requirements were refused by the system and leave directly. When number of customers in the system is less than c, the customers are accepting the service directly, when the number of customers in the system is between numerical c and numerical c+m and the customers are waiting in line accessing the system.

Model analysis: If $L_n = L(n^+)$ represents the number of customer in the system at time $t = n^+$. According to the provisions of the arrival and departure behavior, customer

arrived at $t = n^-$ will store in L_n, if $L_n = L(n^+)$ represents the number of customer in the system at time $t = n^+$. For the system with late arrival delay entrance, customer arrived at (n^-, n) will store in L_n and customer will depart at (n, n^+) will also store in L_n $\{L_n, n \geq 0\}$ is the finite state Markov chain which has the state space $\Omega = \{0,1,\cdots c, c+1,\cdots, c+m\}$ where the number of customers in the service is a transformation from L_n to L_{n+1} transition probability is as follow:

$$p_{ij} = P\{L_{n+1} = j \,|\, L_n = i\} \; n\geq1, i, j \in \Omega$$

If set:

$$\binom{c}{-1} = \binom{c}{c+1} = 0$$

the following equation can be derived:

$$p_{c+i,j} = \bar{p}\binom{c}{j-i}\mu^{c+i-j}\bar{\mu}^{j-i} + p\binom{c}{j-i-1}\mu^{c+i+1-j}\bar{\mu}^{j-i-1}$$
$$0 \leq i \leq m, i \leq j \leq c+i+1$$

Similarly, the transition probability matrix is expressed as:

$$P = \begin{pmatrix}
P_{0,0} & P_{0,1} \\
P_{1,0} & P_{1,1} & P_{1,2} \\
\vdots & \vdots & \vdots & \ddots \\
P_{c-1,0} & P_{c-1,1} & P_{c-1,2} & \cdots & P_{c-1,c} \\
P_{c,0} & P_{c,1} & P_{c,2} & \cdots & P_{c,c} & P_{c,c+1} \\
P_{c+1,0} & P_{c+1,1} & P_{c+1,2} & \cdots & P_{c+1,c} & P_{c+1,c+1} & P_{c+1,c+2} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
P_{c+m-1,0} & P_{c+m-1,1} & P_{c+m-1,2} & \cdots & P_{c+m-1,c} & P_{c+m-1,c+1} & P_{c+m-1,c+2} & \cdots & \cdots & P_{c+m-1,c+m-1} & P_{c+m-1,c+m} \\
P_{c+m,0} & P_{c+m,1} & P_{c+m,2} & \cdots & P_{c+m,c} & P_{c+m,c+1} & P_{c+m,c+2} & \cdots & \cdots & P_{c+m,c+m-1} & P_{c+m,c+m}
\end{pmatrix}$$

Steady state probability distribution algorithm: When $0<p<1$, $0<\mu<1$ and $1 \leq m \leq c$ because of the structure of transition probability matrix P $\{L_n, n \geq 0\}$ is finite state irreducible, non-periodic so, it is back normally using $\Pi = \{\pi_0, \pi_1, \pi_2, \cdots, \pi_c, \pi_{c+1}, \cdots \pi_{c+m}\}$ to represent the limitation distribution of $\{L_n, n \geq 0\}$ and Π satisfy:

$$\Pi P = \Pi \tag{1}$$

Set $\pi_{c+m} = 1$ from Eq. 1:

$$\pi_{c+m} = \pi_{c+m-1}P_{c+m-1, c+m} + \pi_{c+m}P_{c+m, c+m} \tag{2}$$

From Eq. 2 the following equation can obtain:

$$\pi_{c+m-1} = \frac{\pi_{c+m} - \pi_{c+m}P_{c+m, c+m}}{P_{c+m-1, c+m}} \tag{3}$$

According to Eq. 1, the following equation can also get:

$$\pi_1 = \sum_{i=0}^{c+m} p_{i,1}\pi_i, \ \pi_2 = \sum_{i=0}^{c+m} p_{i,2}\pi_i, \ ..., \ \pi_k = \sum_{i=k-1}^{c+m} p_{i,k}\pi_k$$

summarize π_0, π_1, π_2, $\cdots$, π_k the following equation can be derived:

$$\pi_k = \sum_{i=k}^{c+m} p_{i,k}\pi_k + p_{k-1,k}\pi_{k-1}, \ k = 1,2,...,c+m \qquad (4)$$

From Eq. 4 and 5 can be derived as:

$$\pi_{k-1} = \frac{\pi_k - \sum_{i=k}^{c+m} p_{i,k}\pi_k}{p_{k-1,k}}, \ k = 1,2,..., \ c+m \qquad (5)$$

From Eq. 2, π_{c+m-1} can be expressed as the expression of π_{c+m}, π_{c+m-2} can be expressed as the expression of π_{c+m-1} and π_{c+m}. Using the formula (5) and so on π_0 can be expressed as the expression consist of $\{\pi_1, \pi_2, \cdots, \pi_c, \pi_{c+1}, \cdots \pi_{c+m}\}$. Since, by the recursive algorithm and regularized conditions the steady state probability distribution $\pi_0, \pi_1, \pi_2, \cdots, \pi_c, \pi_{c+1}, \cdots \pi_{c+m}$ can be calculated.

Performance analysis of the database connection pool: Request blocking probability, namely a database connection request is rejected because of no connection available in the system marked as p_l. Request blocking probability is an important measure of service quality measuring a database connection pool system. If request blocking probability is too high, the higher maximum number of connections shall be configured by the system. The mathematical expression for request blocking probability is given by following:

$$p_l = \pi_{c+m}\bar{\mu}^c$$

where, π_{c+m} is steady state probability distribution, μ is service rate, accordingly $\bar{\mu} = 1 - \mu$ is not served probability and c is maximum number of connections.

System average waiting time that is the mean of average waiting time under the steady state database connection pool connection requests marked as w. System average waiting time reflects the waiting time of connection requests in the system after the system runs stably, it provide an important theoretical basis for configuration parameters of database connection pool technology, its mathematical expression is:

$$w = \frac{\sum_{i=0}^{m} i\pi_{c+i}}{cu}$$

where, i is the number of connection in the system, π_{c+i} is the steady state probability when connection number is i, c is maximum number of connections and μ is rate of service.

System utilization that is ratio of connected users numbers $\bar{k}$ and the maximum number c of connections can be established in the connection pool marked as η. System utilization has a close relationship with the system resource usage, analyzing the effect parameters of system utilization has an important reference significance for parameter configuration of database connection pool technology and its mathematical expression is given by following:

$$\eta = \frac{\sum_{i=0}^{c} i\pi_i + c\sum_{i=1}^{m} \pi_{c+i}}{u}$$

where, i is the number of connection in the system, π_i and π_{c+i} is the steady state probability when connection number is i, c is maximum number of connections and μ is rate of service.

System throughput, the number of customers accomplished the service per unit of time, mark as θ. System throughput directly reflects the whole system capability to handle a connection request. As an important measure to evaluate system performance and its mathematical expression is given by following:

$$\theta = \sum_{i=0}^{c} i\mu\pi_i + c\mu\sum_{i=1}^{m} \pi_{c+i}$$

where, i is the number of connection in the system, π_i and π_{c+i} is the steady state probability when connection number is i, c is maximum number of connections and μ is rate of service.

NUMERICAL EXAMPLES

In order to visually display the dependency between the database connection pool performance measures and system configuration parameters, Fig. 3-14 show request blocking probability, the average waiting time, the system utilization and system throughput as the request arrival rate, the maximum number of connections and queuing place capacity change trend graph in different conditions. The tests were set the parameters for the following two cases:
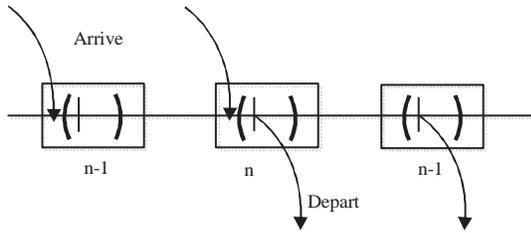
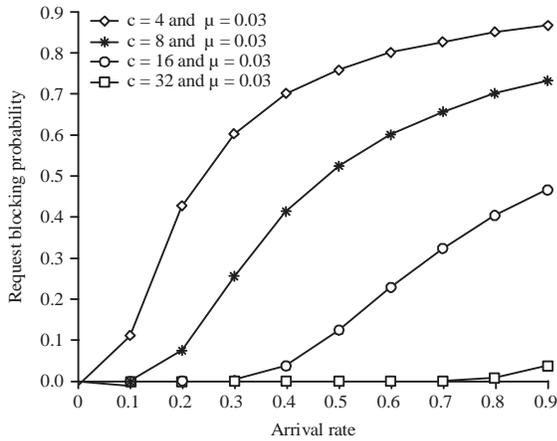Fig. 2: Customer arrival and departure

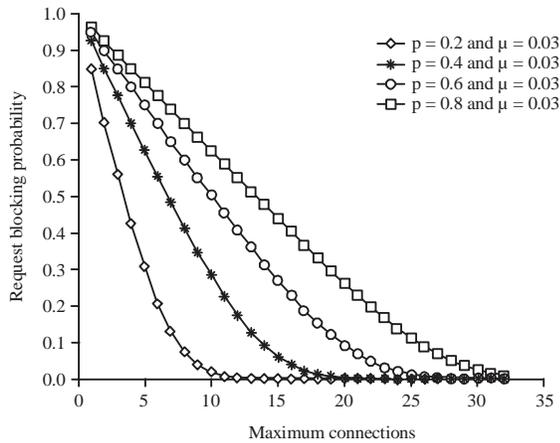Fig. 3: Request blocking probability changes as the arrival rate

Fig. 4: Change of request blocking probability with the maximum connection No.

First one: The analysis of the system performance measures with the arrival rate and the maximum number of connections will be related to changes in parameters: Each time slot connection request arrival probability p is set 0.1-0.9, probability of each time μ slot connection service complete set 0.03 and the maximum number c of connections is set 1-32 and the current queuing capacity m is set 2.
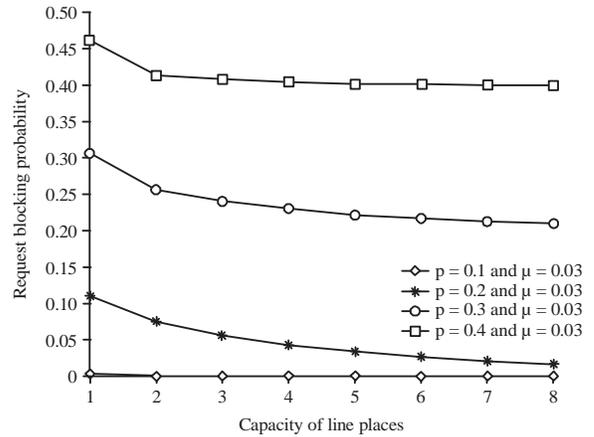
Fig. 5: Request blocking probability with queuing place capacity change
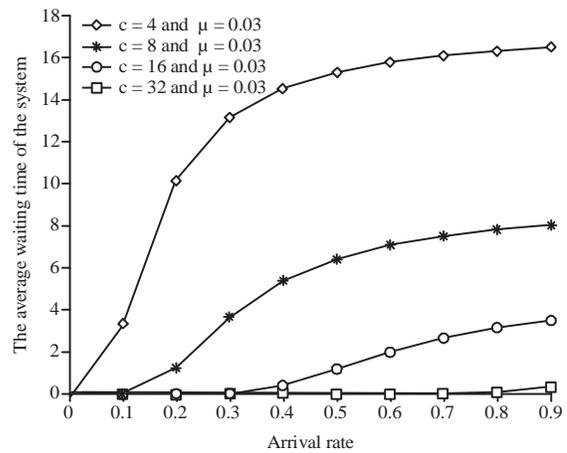
Fig. 6: Average waiting time varies with the arrival rate
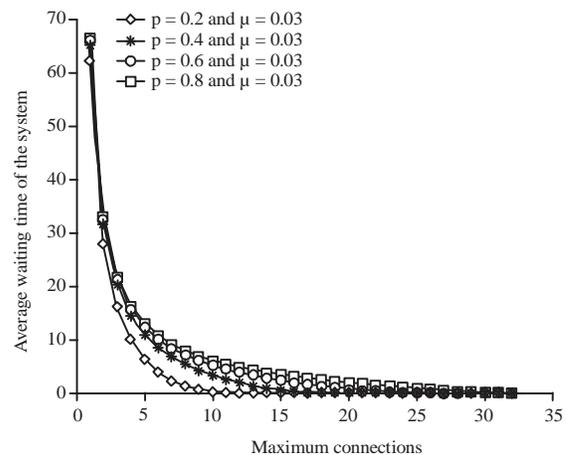
Fig. 7: Average waiting time changes with the maximum number of connections

Second one: The analysis of the system performance measures with the queue capacity change places, the related

parameters: Each time slot connection request arrival probability p is set 0.1-0.4, probability of each time μ slot connection service complete set 0.03, the maximum number of connections c is set 8 and the current queuing capacity m is set 1-8.

Figure 3 shows request blocking probability curve with arrival rate when the service rate is set 0.03, queuing place capacity is set 2 and the maximum number of connections are set 4, 8, 16 and 32. When the service rate, the maximum number of connections, queuing capacity were fixed and the request blocking probability increases with the request arrival rate. This is due to the increase of connection requests, the service station to the timely processing of connection requests is reduced and the queuing place capacity is fixed, the queue waiting for service customers there are fixed so, failing to deal with the number of connections is increased then the request blocking probability increases; when the connection completion of service rate and the request arrival rate of a certain and the queuing place fixed capacity conditions and request blocking probability decreases with the increase of the maximum number of connections. This is due to when the maximum number of connections increases, the number of connections per unit time can increase. The number of connections to be completed, the corresponding processing is reduced and the corresponding request blocking probability decreases.

Figure 4 is the request change curve of the blocking probability with the maximum number of connections when the service rate is set to 0.03 and the number of queuing place capacity is set to 2. The request arrival rate is set to 0.2, 0.4, 0.6 and 0.8. When the service rate, the request arrival rate and queue places certain capacity are fixed, the request blocking probability increases with the maximum number of connections decreases and approaches zero when the service rate, the maximum number of connections to certain places queuing capacity are fixed and request blocking probability increases with the request arrival rate increases.

Figure 5 is request blocking probability curve with queuing place capacity when the service rate is set 0.03, the maximum number of connections is set 8, the request arrival rate 0.1, 0.2, 0.3 and 0.4, respectively. When the service rate, the maximum number of connections and the request arrival rate under certain conditions, request blocking probability increases with the queuing places capacity decreases and approaches to a constant value, this is due to the connection request in can not be timely serviced and will enter into the wait queue places with increases with queuing place capacity then able to queue wait for the request to be many and the blocking probability corresponding will reduce.
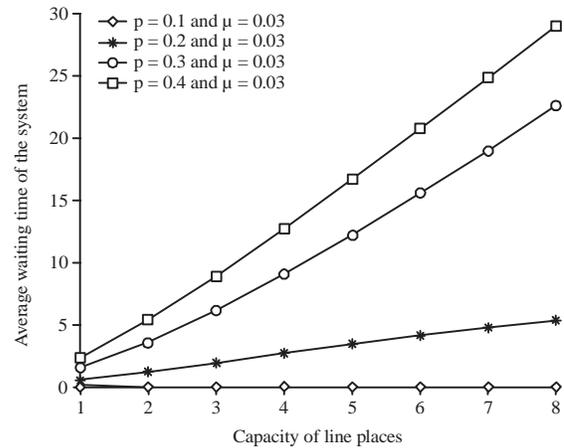

Fig. 8: Average waiting time change with queue capacity

Figure 6 is the average waiting time curve with arrival rate when the service rate is set 0.03, queuing place capacity is set 2 and the maximum number of connections is set 4, 8, 16 and 32. When the service rate and the maximum number of connections to certain places and queuing capacity of fixed conditions, the average waiting time as the request arrival rate increases. This is due to with increase of connection requests, the corresponding can not be timely service connection request will increase, this connection request will enter into the queue of waiting places so, that the average waiting time will increase when more than queuing place capacity, the average waiting time approaches the set value; when the service rate and the request arrival rate some places and queuing capacity of fixed conditions and increasing the average waiting time decreases with the maximum number of connections.

Figure 7 is the average waiting time change curve with the maximum number of connections when the service rate is set 0.03, queuing place capacity is 2, the customer arrival rate is set 0.2, 0.4, 0.6 and 0.8. When the service rate, the request arrival rate and queue places certain capacity are fixed, the average waiting time increases with the maximum number of connections and increase and approaches zero, when the condition of service rate, the maximum number of connections and the queuing system are fixed and the average waiting time increases with the request arrival rate increase.

Figure 8 is the average waiting queue capacity curve with the place when the service rate is set 0.03, the maximum number of connections is set 8, the request arrival rate 0.1, 0.2, 0.3 and 0.4, respectively. When the service rate, the maximum number of connections and the request arrival rate are under certain conditions, the average waiting time increases with the queuing place capacity, this is due to the
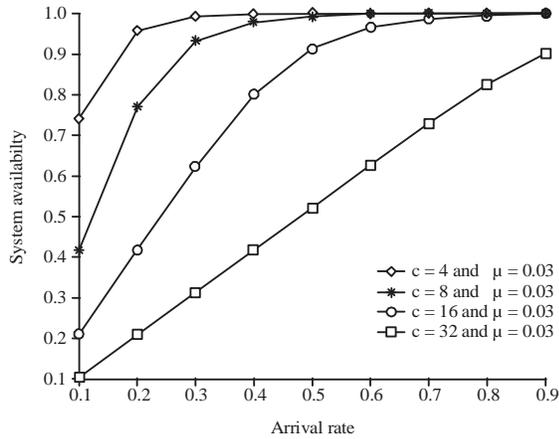
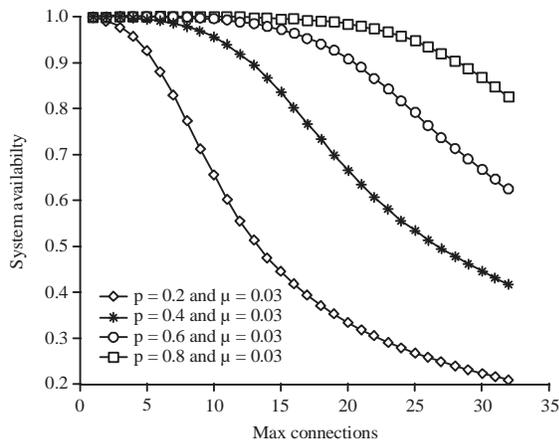Fig. 9: System utilization rate changes with the arrival rate

Fig. 10: Utilization ratio of the system changes with the maximum No. of connections

Fig. 11: Utilization ratio of the system with the queue capacity changes place

queue capacity changes with increasing of place and connection requests cannot be timely serviced will enter into the wait queue places, greater capacity and able to queue the connection request. The average waiting time of the system is longer, if the arrival rate is low enough, the connection request is directly into the service, the waiting time is shorter or even zero.

Figure 9 is the system utilization rate curve with arrival rate when the request arrival rate is set 0.03, queuing place capacity is set 2 and the maximum number of connections is set 4, 8, 16 and 32. When the service rate, the maximum number of connections to certain places and queuing capacity conditions are fixed and the system utilization rate increases and approaches 1 as the request arrival. This is due to with increase of connection requests, the unit time connection establishment services will increase, the corresponding system utilization increased until the system full load operation; when the service rate and the request arrival rate and queue

places certain capacity under the condition of system utilization increases with the number decreased. This is due to with increasing the maximum number of connections, the number of connections per unit time to be completed and the corresponding increase and the idle connection will increase which reduce the utilization rate of system.

Figure 10 is the system utilization rate curve with the maximum number of connections when the service rate is set 0.03, queuing place capacity is set 2 and the request arrival rate was set 0.2, 0.4, 0.6 and 0.8. When the service rate, the request arrival rate and queue capacity of certain are fixed, the system utilization increases with the maximum number of connections decreases; when the service rate and the maximum number of connections to certain places and queuing capacity are fixed, system utilization rate increases with the increase of request arrived.

Figure 11 is the system utilization rate curve with queuing place capacity when the service rate is set 0.03 and the maximum number of connections is set 8. The request arrival rate is, respectively 0.1, 0.2, 0.3 and 0.4. When the service rate, the maximum number of connections and the request arrival rate are under certain conditions, the system utilization rate increases with the queuing places capacity increase and tends to a fixed value; when the service rate, the maximum number of connections and queuing places capacity are under certain conditions and system utilization rate increases with the increase of request arrived. When you reach a certain degree, the system will run at full capacity and the utilization rate is close to 1.

Figure 12 is the system throughput curve with arrival rate when the request arrival rate is set 0.03, queuing place capacity is set 2 and the maximum numbers of connections are set 4, 8, 16 and 32. When the service rate and the maximum number of connections to certain places and
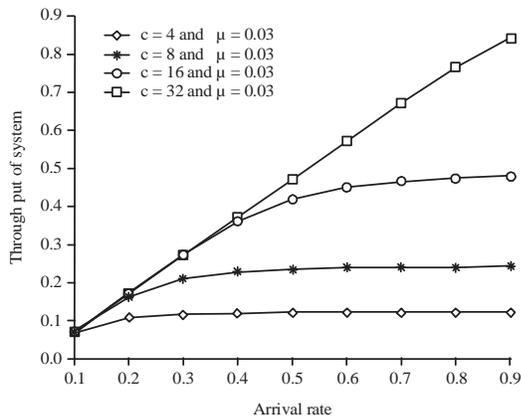
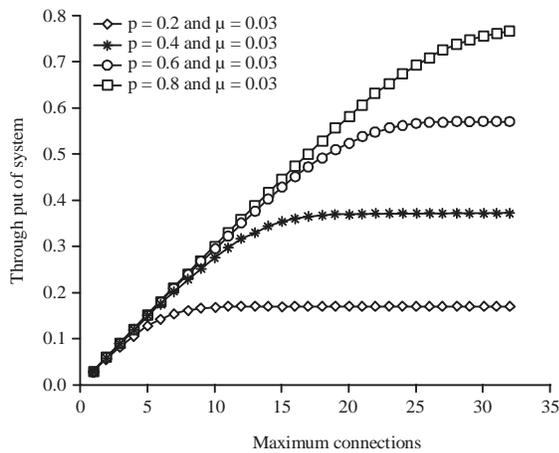Fig. 12: System throughput changes with the arrival rate

Fig. 13: System throughput changes with the maximum No. of connections

Fig. 14: Throughput of the system changes with the queue place capacity

queuing capacity are fixed, the system throughput ratio increase with the request and approaches to the setting value. This is due to the increase of connection requests, the unit time connection establishment services have increased then the corresponding throughput of the system increased until, the system study in full load when the service rate, the request arrival rate and queue places certain capacity are fixed, the throughput of the system increases with the maximum number of connections. This is due to increase of the maximum number of connections, the establishment number of connection service can increase and the corresponding system throughput increases.

Figure 13 is the variation curve of the throughput of the system with the maximum number of connections when the service rate is set 0.03 and the request arrival rate is set respectively 0.2, 0.4, 0.6 and 0.8. When the service rate, the request arrival rate and queue places certain capacity are
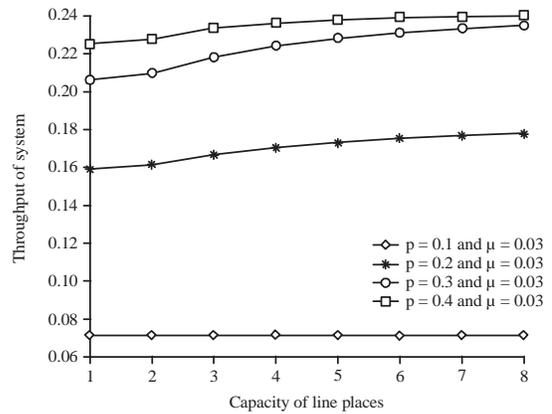
fixed, system throughput increase with the maximum number of the connections increases; when the service rate, the maximum number of connections and conditions queuing field capacity are fixed and the system throughput rate increases with the request arrival.

Figure 14 is the system throughput curve with queuing place capacity when the service rate is set 0.03, the maximum number of connections is set 8 and the request arrival rate is 0.1, 0.2, 0.3 and 0.4, respectively. When the service rate, the maximum number of connections and the request arrival rate are under certain conditions, the throughput of the system increases with the queuing places capacity increase and tend to a constant value. It is because that with increase of the queue place capacity, connection requests are under not be timely serviced and then will enter into the waiting so that queue place capacity increases. The more queue connections are able to request, the service station can promptly to the connection request is processed and the more the system throughput.

DISCUSSION

The main innovation of this study lies in the application of discrete time queuing theory to the performance analysis of database connection pool. Qualitative analysis was performed for database connection pool technology from the angle of mathematical model. Through three kinds of performance comparison tests under different database access patterns, Wang *et al.*[1] provides data support for database connection pool technology to improve database access performance. The design and implementation of database connection pool technology is studied in the Liang *et al.*[2] based on a specific system which proves the validity of the technology of

database connection pool. The optimization strategy of database connection pool technology was studied by Zhu and Shane[3]. The above study belongs to the study of the same object from different angles.

Section 4 of this study reveals the inner dependency relations between system performance measures for example request blocking probability, system average waiting time, system utilization and system throughput etc. System configuration parameters for example request arrival rate, the maximum number of connections and queuing places capacity. From the 5th section in this study can conclude: (1) If the other conditions are fixed, the request blocking probability increased with the increase of request arrival with the increase of the maximum number of connections decreases and tends to zero and decreases with the places capacity becomes larger, (2) If other conditions fixed, the average waiting time increased with the increase of request arrival rate with the increase of the maximum number of connections decreases and tends to zero and increases with larger capacity queuing places, (3) If the other conditions are fixed, the system utilization rate increases and tends to be constant with the increase of request arrival, decreases with larger the maximum number of connections and with increases with the places capacity becomes larger and (4) If other conditions fixed, the system throughput increased and reached close to a fixed value with the larger request arrival rate with the increase of the maximum number of connections increases and closer to the fixed value with the larger capacity of the queue becomes larger.

In this study, the model is based on the discrete time queuing theory which makes the analysis result more close to the actual situation. The model uses the Bernoulli arrival process to describe the arrival process of the request and the randomness of the arrival process is embodied. Model also both considers the loss of request in the database connection pool technology and the waiting for the situation, making the model more effective. Application of the model to complete the database connection pool technology performance analysis is an effective supplement for the database connection pool technology study.

CONCLUSION

This study analyzed the database connection pool management process. According to the management process a Geom/Geom/c/c+m discrete-time queuing model with multi-server and mixed waiting and loss was established then completed the theoretical analysis of the model and derived the request blocking probability, the average waiting time, the system utilization and system throughput formula of the system performance measures. Using the numerical examples showed the dependency relationships between the performance measures of the database connection pool and the system configuration parameters. The next step is to simulate the model then complete performance analysis and evaluation of the database connection pool technology in more accurate level.

ACKNOWLEDGMENT

REFERENCES

1. Wang, Y.J., F. Wang and W.B. Yang, 2010. Performance comparison test of database access based on database connection pool. Ind. Control Comput., 6: 92-94.
2. Li, B.Z., F.S. Wu, Z.J. Jing and H.J. Huang, 2008. Self-adaptive database connection pool technology in NVMS. Comput. Eng., 23: 41-43.
3. Zhu C.S. and Y.F. Shane, 2003. Study on self-adaptive database connection pool. Comput. Eng. Applic., 36: 187-209.
4. Liang, Q.H., Z.F. Shen, J.C. Luo, H. Fan, D.P. Ming and J.L. Li, 2006. Study of database connection pool in LBS platform. Comput. Eng., 32: 39-41.
5. Meng, P.C., S.B. Hu, H. Shu and F.Y. Yan, 2013. Optimization strategy of connection pool based on ADO. Comput. Eng. Design, 34: 1706-1710.
6. Zeng, G.L. and X.F. Fu, 2011. A new model of database connection pool. Comput. Digital Eng., 39: 163-166.
7. Lv, J.B., G.Z. Dai and D.J. Mu, 2012. Implementation of absolute delay guarantee in database connection pool in Web application servers. Applic. Res. Comput., 29: 1838-1841.
8. Liu, F., 2012. A method of design and optimization of database connection pool. Proceedings of the 4th International Conference on Intelligent Human-Machine Systems and Cybernetics, Volume 2, August 26-27, 2012, Nanchang, Jiangxi, pp: 272-274.
9. Tian, N. and Z.G. Zhang, 2006. Vacation Queueing Models: Theory and Applications. Springer, New York.