Performance Analysis and Evaluation of Database Connection Pool Technology Based on Queueing Model

Huo Zhan-Qiang, Zhang Jin-Cheng and Wang Zhi-Heng
School of Computer Science and Technology, Henan Polytechnic University, Jiaozuo, 454000, Henan, China

Abstract: In order to reveal the dependency relationships between configuration parameters and performance measures of database connection pool technology, the Geom/Geom/c/c discrete-time queuing model with multi-server is built according to its working principles. Using the embedded Markov chain method, the analysis of the model is completed and the mathematical expressions of system performance measures are derived, such as request blocking probability, average number of connections, system utilization, and system throughput etc. Finally, numerical examples are given to discuss the impact of these four performance measures on request arrival rate and maxConn (the maximum number of connections) intuitively. The analysis results of this study provide theoretical bases for the parameter configuration of database connection pool.

Key words: Database connection pool, performance analysis, queueing model

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. Connection reuse can improve the utilization rate of connection and improve the performance of the system about database operations and reduce the system overhead. The core of the connection pool is connection reuse.

Wang et al. (2010) 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 Bingzhang and other people propose the optimal allocation method of database connection pool based on the log file records. Zhu Changsheng et al. give us the configuration method in detail for database connection pool parameter configuration based on XML under J2EE framework. Liang et al. (2006), Merg et al. (2013), Zeng and Fu (2011), Lv et al. (2012) and Liu (2012) et al. 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 three categories: Experimental data analysis, simulation research 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 Tian and Zhang (2006).

WORKING PRINCIPLE

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 free connections in the pool. If there is a free connection, then take out a free connection to assign to the customers to use and at the same time mark the connection as busy connection, as shown in (1) in Fig. 1
- If there is no free connection, check whether the total number of current free connections and the busy connections have reached the maximum number of connections set by the system. If it is not up to the maximum number of connections, create a connection to the request of the customer, as shown in (2) in Fig. 1. If it has reached the maximum number of connections, wait according to the maximum waiting time set by the system, if it exceeds the maximum waiting time, an exception will be thrown to the customer, as shown in (3) in Fig. 1

Corresponding Author: Wang Zhi-Heng, School of Computer Science and Technology, Henan Polytechnic University, Jiaozuo, 454000, Henan, China
Fig. 1: Diagram of management process of database connection pool

- In the end of the use of connection for customers, put the connection back into the connection pool to improve the reuse rate of connection while the connection is marked as free connections, as shown in (4) in Fig. 1

**DESCRIPTION AND ANALYSIS**

**Description of model:** Timeline is divided into fixed interval of small pieces, called time slot. Arrival can only occur at the end of each time slot \( t = n^- (n = 1, 2, \ldots) \), known as late arrival system. In order to clearly define the number of customers in system at a discrete time \( t = n \), all the start and end agreed upon in the service will be at \( t = n \). If customers arrive at \( n^- \) and someone left at \( t = n \), he can immediately begin in the service at \( t = n \). This system is called immediate access. Customer arrival and departure behaviors are shown in Fig. 2.

Each end of time slot shall have an arrival with probability \( p \) and no arrival with probability \( \bar{p} = 1 - p \), the arrival on different time slots are independent. In other words, the arrival interval \( \lambda \) obeys geometric distribution. The customer service time \( \mu \) also obeys the geometric distribution. Arrival interval and service time are independent of each other. There is \( C \) front desks in the system, if the customer arrival is in the situation that \( C \) front desks are occupied and no customers left in his arrival time, the customer will be refused to enter into the system.

**Model analysis:** Assume that \( L_n = L(n') \) indicates the number of customers at the time \( t = n^- \) in the system. According to the arrival and departure behavior of the agreement, the arriving customers at \( t = n^- \) will be included in \( L_n \) and the customers who left at \( t = n \) shall not be included. \( \{L_n, n \geq 1\} \) is a Markov chain with state space \( \Omega = \{0, 1, \ldots, C\} \), the transition probabilities formation is:

![Customer arrival and departure](image)

**Fig. 2**
NUMERICAL EXAMPLES

The relevant parameters in this section are set as follows: connection request arrival probability $p$ in each time slot is set to 0.1-0.9. Serve ratio $\mu$ in each time slot is set to 0.03. The maximum number of connections $C$ is set to 1-32.

Figure 3 shows when the service rate is set to 0.03, the maximum number of connections is set to 4, 8, 16 and 32 separately and the curve of request blocking probability complies with the arrival rate. When the service rate $\mu$ and the maximum number of connections $C$ are under certain conditions, Request blocking probability increases with the increase of request arrival rate $p$ and decreases with the increase of the maximum number of connections $C$.

Figure 4 shows when the service rate is set to 0.03, the request arrival rate is set to 0.2, 0.4, 0.6 and 0.8
separately, the curve of request blocking probability with the maximum number of connections. When the service rate $\mu$ and the request arrival $\theta$ are under certain conditions, the request blocking probability decreases with the maximum number of connections $c$. When the service rate $\mu$ and the maximum number of connections $c$ are under certain conditions, the request blocking probability increases with the increase of request arrival rate $p$.

**Probability comply with the maximum number of connections**: Figure 5 shows when the service rate is set to 0.03, the maximum number of connections is set to 4, 8, 16 and 32 separately, the curve of system average number of connections with the arrival rate.

While the service rate $\mu$ and the maximum number of connections are fixed, the system average number of connections increases with request arrival rate $p$ and closes to the fixed value of $c$. When the service rate $\mu$ and the request arrival rate are under certain condition, the system average number of connections increases with the increase of maximum number of connections $c$.

Figure 6 shows when the service rate is set to 0.03; the request arrival rate is set to 0.2, 0.4, 0.6 and 0.8 separately, the curve of system average number of connections complies with the maximum number of connections.

Figure 7 shows when the service rate is set to 0.03 and the maximum number of connections is set to 4, 8, 16 and 32 separately, the curve of system utilization rate complies with the arrival rate.

Figure 8 shows when the service rate is set to 0.03 and the request arrival rate is set to 0.2, 0.4, 0.6 and 0.8 separately, the curve of system utilization rate complies with the maximum number of connections. When the
Figure 9 shows the changes of system throughput comply with the arrival rate when the service rate $\mu$ and the request arrival rate $p$ are under certain conditions, the system utilization rate decreases with the increase of maximum number of connections $c$. When the service rate and the maximum number of connections $c$ are under certain conditions, the system utilization rate increases with the increase of request arrival rate $p$.

**Comply with the maximum number of connections:**

Figure 9 shows when the service rate is set to 0.03 and the maximum number of connections is set to 4, 8, 16 and 32 separately, the curve of system throughput complies with the arrival rate.

**ACKNOWLEDGMENT**

This study was supported by Natural Science Foundation of Hebei Province (No. F2009000475) and Doctorial Foundation of HPU (No. B2010-26).

**REFERENCES**


