Peer-reviewed and Open access journal
e-ISSN: 1804-4999 | www.academicpublishingplatforms.com
The primary version of the journal is the on-line version

ATI - Applied Technologies and Innovations
Volume 11 | Issue 2 | 2015 | pp.63-70
DOI: http://dx.doi.org/10.15208/ati.2015.07

Maximum load of database service with recovery

Victor Boicov
Department of Electronics and Telecommunication
Riga Technical University, Latvia

e-mail: viktor.boicov@inbox.lv
postal address: Lomonosov str.1, Riga, LV-1019, Latvia

The paper analyzes the database services load with recovery. A mathematical model of the interaction with the database services is developed. On the basis of mathematical modelling procedures for the exchange services, the possible delays and the presence of large queues to services are defined. Suggestions to improve the exchange of services are developed.

Keywords: Heterogeneous services, client and server requests, network queuing systems, heterogeneity

Citation: *Boicov V., 2015. "Maximum load of database service with recovery", Applied Technologies and Innovations, Vol.11(2), pp.63-70, http://dx.doi.org/10.15208/ati.2015.07*

Introduction

The main purpose of any database is to provide efficient services to users. One of the fundamental works on relational databases (Codd, 1970) identifies four main types of database services: "Projection", "Intersection", "Selection" and "Union" and four types of special services: "Subtraction", "Descartes product", "Connection" and "Division". All these services are independent from each other, although special services can be expressed in terms of the basic relational operators. Each of the services in the database is processed during the period of time determined by the type of query. This indicates the presence of heterogeneity in database network queries processing.

For the use of services to be efficient, it is necessary to build models that allow for an unlimited number of services and take into account the presence of heterogeneity. In a number of works of the author, models of network queuing systems with heterogeneous services were constructed (Boicov, 2009; 2011a). However, these studies do not take into account the possibility of failures and recoveries. On the other hand, there are many works that take into account the possibility of recovery and failures in databases (Kumar, 2001; Gelenbe and Derochette, 1979), but still do not take into account the heterogeneity. The following work encompasses both accounting for the heterogeneity of services and the possibility of restoring the database.

The mathematical model of the database services interaction without recovery

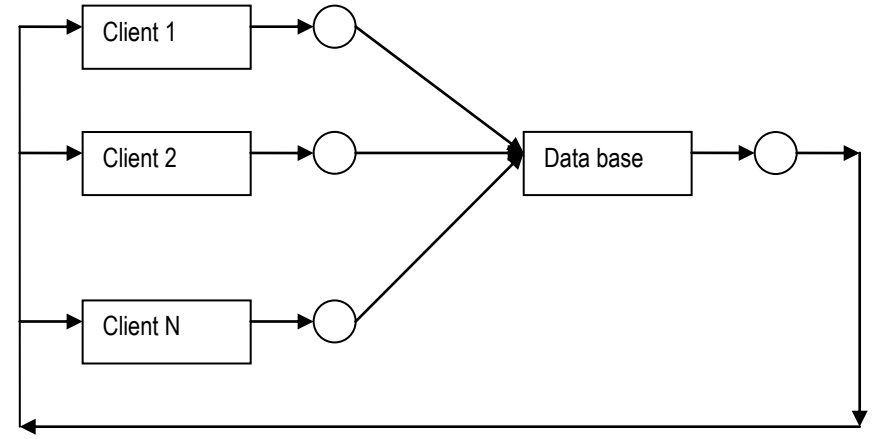
In (Boicov, 2012) a mathematical model of the interaction between the workstation and the server is presented, one that takes into account the presence of two types of services. However, it was noted above that for the exchange of data in a database the number of services may be of more than eight types. To account for these factors, a mathematical model of the interaction of the services with the network database is presented below. The

Applied Innovations and Technologies
ISSN: 1804-4999

general scheme of interaction of clients and the database server can be represented as a closed network queuing system (NQS).

The scheme of such a network is shown in Figure 1.

FIGURE 1. SCHEMATIC PORTRAYAL OF THE INTERACTION OF THE CLIENT AND SERVER QUEUES



Each client program sends a request to the database. This request may be of any length. In general, the length of this request is random. Each client's request lengths depend on many random factors. This randomness is manifested in servicing these requests by the database. The database is required to respond to each client request. Furthermore the database itself can generate queries. The times of occurrence of these requests are random, and the length of these requests is random as well. In each case, the client must respond to the query of the database server. Thus, a closed loop of interaction between the database programs and the client emerges. Client work can be interpreted with the help of QS models. In Figure 1, each QS of the client is marked with its own rectangle with the name 'Client' and the corresponding number of the client. Transitions of requests from the client to the server and from the server to the client are marked with arrows. These transitions are probabilistic in nature. The times of servicing requests by both the server and the client also random. Client service time is homogenous because this time is determined by the time of the implementation of the programs of one client. The time of servicing requests by the server is heterogeneous. A study of QS networks with heterogeneous service laws is presented in (Boicov, 2012). However, in our case, the network QS is mixed, since it contains both homogeneous and heterogeneous queries. In (Boicov, 2009) an expression was obtained that estimates of mean number of customers in one particular QS network node with heterogeneity. This expression has the following form:

$$N_{(R)} = \sum_{i=1}^{R} \lambda_i/\mu_i + (\sum_{i=1}^{R} \lambda_i) [(\sum_{i=1}^{R} \lambda_i/\mu_i^2)/ (1 - \sum_{i=1}^{R} \lambda_i/\mu_i)] \qquad (1)$$

Here λi - the average intensity of entry into the system of the i - th type queries, μ_i - average service rate of these requests, R - the number of service types, $N_{(R)}$ - the number of requests that are in the queue and serviced by the heterodyne QS.

In the particular case when R = 1, we turn to the homogeneous service. Average number of demands in a homogeneous QS for this case is determined by the following expression:

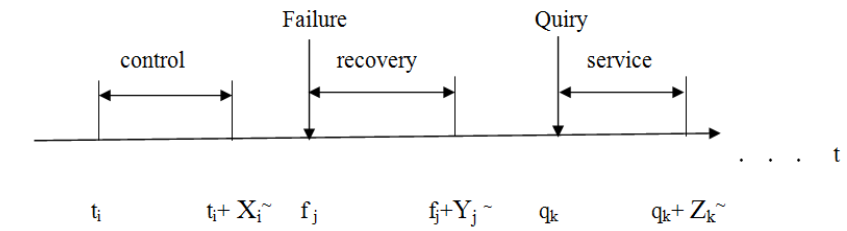
$$N = \lambda/\mu + \lambda^2/\mu^2 / [(1-\lambda/\mu)] \quad (2)$$

Equation (1) yields the average queue length for a database server, and using the relation (2) we can find the average queue length on the client side. These relationships make it possible to assess the effectiveness of the database without taking into account the emergence of its failure and recovery.

Mathematical model of interaction of database services with recovery

All questions related to data recovery are defined by restoration procedures and failure rate. Various algorithms are used to ensure recovery; they are implemented in the program as special points of rollback, control and data recovery. For the implementation of the recovery model, we assume that the occurrence of failure is instantaneous. Creating control points, the time of data recovery and service requests are distributed along the time axis in the form of appropriate actions that are shown in Figure 2.

FIGURE 2. DATA RECOVERY MODEL



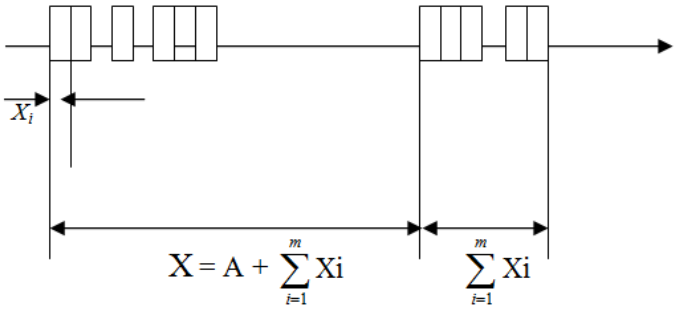
Here, $X_i^{\sim}$ indicates the time of the i- th control request , $Y_j^{\sim}$ - recovery time of the j- th failure and $Z_k^{\sim}$ - service time for the k- th order.

That said, the instants of occurrence of queries and the moments of occurrence of errors may take place during the processes of control, repair and service. We can assume that two or more operations cannot be executed simultaneously by the database server. Because of this assumption, and in accordance with the analogous assumptions that have been made in a number of works on the restoration of databases (Kumar, 2001; Gelenbe and Derochette, 1979) that arise in the course of requests' servicing, we can assume that failures and moments of queries in the database take place in accordance with the laws of Poisson. The average service time of requests can be accepted as exponentially distributed (Boicov, 2012). I must say that this assumption may well be justified, more so because the estimations of the load in the databases which are exploring the performance in reality almost always take the critical value. Let us now turn to the assumptions regarding the

Applied Innovations and Technologies
ISSN: 1804-4999

timing control. Stationary probability distribution for the database monitoring system can be obtained using the generating function, which has many generations (Boicov, 2011b). Figure 3 shows the distribution of the steady-state control over time, which consists of at least two generations of requests - service requests and requests for control.

FIGURE 3. DISTRIBUTION OF THE STEADY-STATE CONTROL OVER TIME



In Figure 3, X denotes the time between the control moments in a single monitoring session, A - the time interval between the moment of closure of the previous control session and the start of the next, Xi - the time of treatment for commands executed in a single control point, m - the number of control points in the session.

The generating function of the first generation is the function with respect to the duration of the monitoring process:

$$Q^{(1)}(s) = \exp(-X + X * s) \tag{3}$$

The generating function of the second generation is a function that is determined by the probability parameter p - the probability of controlling with a single control point:

$$Q^{(2)}(s) = \sum_{j=1}^{\infty} p * [(1-p) * Q^{(1)}(s)]^{j-1} \tag{4}$$

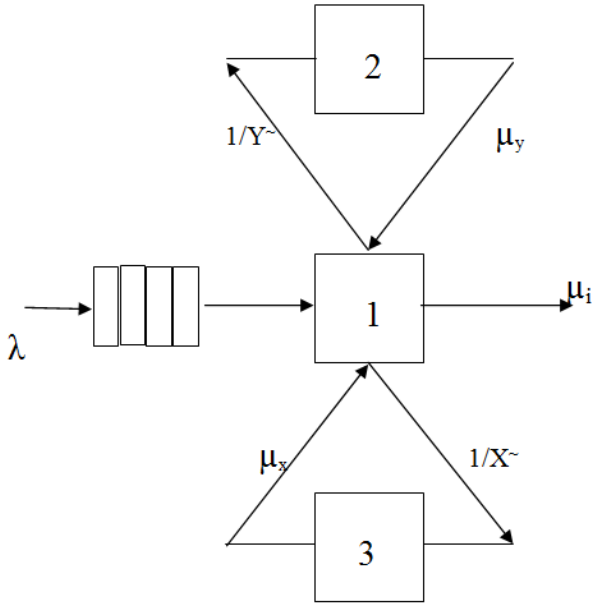
Substituting 3 in 4, we find that:

$$Q^{(2)}(s) = p / [1 - \exp(-X + X * s) + p * \exp(-X + X * s)] \tag{5}$$

The obtained expressions (4) and (5) make it possible to determine the average number of queries to the database for a single control session and monitoring of requests by the server. For overall assessment, a general model of requests service is considered below.

The general model for incoming requests and their service by the database server is shown in Figure 4.

FIGURE 4. GENERAL MODEL FOR INCOMING REQUESTS AND THEIR SERVICE BY THE DATABASE SERVER



Here Y~ is the average time of occurrence of failures, X~ average time of control.

The system of differential equations describing the behaviour of the process (Figure 4) is as follows:

$$d\left(p(0,1,t)\right)/dt = \mu_i p(1,1,t) - (\lambda + Y^{\sim} + X^{\sim}) p(0,1,t);$$

$$d\left(p(n,1,t)\right)/dt = \mu_i p(n+1,1,t) + \mu_x p(1,3,t) + \mu_y p(1,2,t) + \lambda p(n-1,1,t) - (\lambda + Y^{\sim} + X^{\sim}) p(n,1,t), n \geq 1;$$

$$d\left(p(0,2,t)\right)/dt = \mu_y p(1,2,t) - Y^{\sim} p(0,2,t);$$

$$d\left(p(n,2,t)\right)/dt = \mu_y p(n+1,2,t) + Y^{\sim} p(n+1,1,t) - (Y^{\sim} + \mu_y) p(n,2,t) n \geq 1;$$

$$d\left(p(0,3,t)\right)/dt = \mu_x p(1,3,t) - X^{\sim} p(0,3,t);$$

$$d\left(p(n,3,t)\right)/dt = \mu_i p(n+1,3,t) + X^{\sim} p(n+1,1,t) - (\mu_i + X^{\sim}) p(n,3,t), \; n \geq 1.$$

From the resulting system of equations, one can find the distribution of stationary probabilities of states *P(n,k) (n=0,∞, k=1,2,3)*. Here, n is the number of requests in the

Applied Innovations and Technologies
ISSN: 1804-4999

system. The solution of this system can be found by using the method of substitution or by using the following generating function:

$$G_k(s) = \sum_{n=0}^{\infty} P(n,k)s^n \quad k=1,2,3 \tag{6}$$

The values of the stationary probabilities $P(n,k)(n=0,\infty,k=1,2,3)$ determine the length of the queue to the server. By solving the above system of differential equations, and simplifying and taking into account the heterogeneity of services in accordance with the expression (1), we establish the following expression for the mean queue length on the server:

$$N^{\sim}_{servser} = [\ \lambda / \sum_{i=1}^{R} (\mu_i + 1/X^{\sim} + 1/Y^{\sim})\ ] / [1-(\lambda / \sum_{i=1}^{R} (\mu_i + 1/X^{\sim} + 1/Y^{\sim}))\ ] \tag{7}$$

Evaluation of database service downloads

Expression (7) makes it possible to estimate the average utilization of services in a client - server database in the presence of failures and recoveries. From the point of view of the efficiency of use of these devices, the important features are the dependencies of the queue lengths on the amount of client requests, that is, their work intensity. To account for these possibilities, in evaluating the performance of queuing systems, one typically introduces the concept of system load. Queuing system load is the ratio of the intensity of the incoming requests to the system to the intensity of their service. For our case - when the servicing system of a database server is used - this ratio takes the following form:

$$\varrho = \lambda / \sum_{i=1}^{R} (\mu_i + 1/X^{\sim} + 1/Y^{\sim})\ ] \tag{8}$$

Using the obtained expression (7) and the designation (8), we can construct a graph of how the length of the queue depends on the system load for the server and client parts of the database. These dependences - for different mixtures of requests in the server database - are shown in Figures 5 and 6.

On the graphs, the mixtures of requests to the server are related to the different types of queries. According to database theory, proposed by (Codd, 1970), the main types of queries are "Projection", "Intersection", "Selection" and "Union" queries. These are the most time-consuming ones from the point of view of their treatment by database management systems. On the presented graphs, the types of mixtures of requests are scaled in relation to the "Selection" type of query and are defined as follows:

The first type —··—► $X^{\sim} = 0.3;\ Y^{\sim} = 0.4;\ \sum_{i=1}^{R} \mu_i = 0.5,$

The second type ·····► $X^{\sim} = 0.2;\ Y^{\sim} = 0.1;\ \sum_{i=1}^{R} \mu_i = 0.8,$

The third type - - - - -► $X^{\sim} = 0{,}1;\ Y^{\sim} = 0.08;\ \sum_{i=1}^{R} \mu_i = 1.,$

The fourth type $\longrightarrow$ $X^{\sim} = 0.08, Y^{\sim} = 0.05 \sum_{i=1}^{R} \mu_i = 1.8.$

FIGURE 5. THE DEPENDENCE OF THE LENGTH OF CLIENT SERVICE QUEUES ON THE LOAD

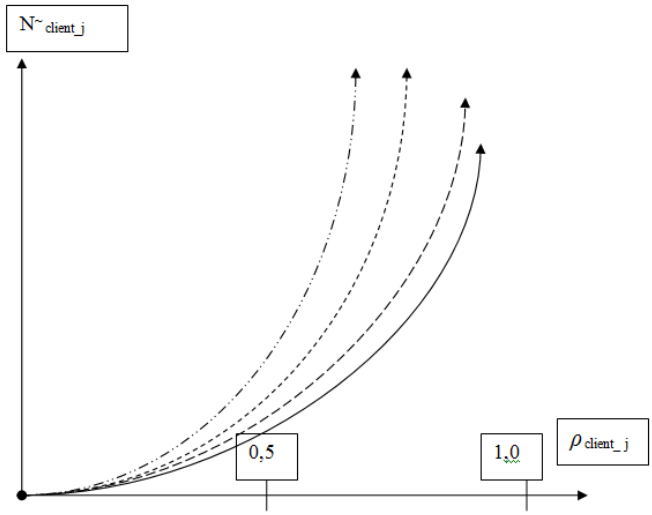
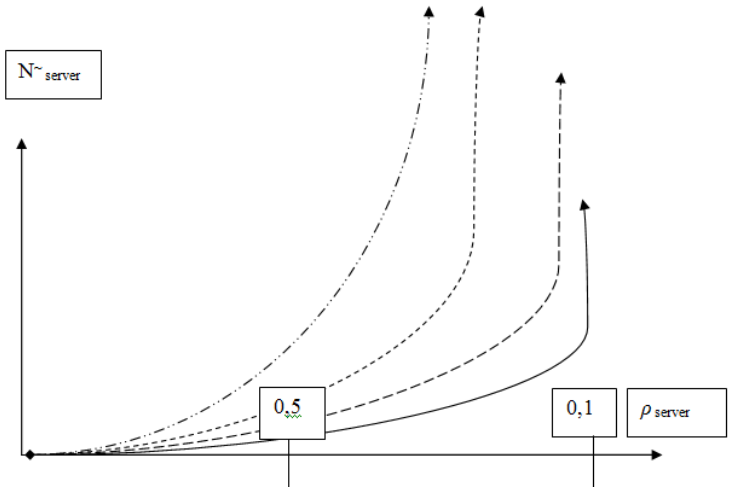


FIGURE 6. DEPENDENCES OF THE LENGTH OF SERVER SERVICE QUEUES ON THE LOAD



Applied Innovations and Technologies
ISSN: 1804-4999

The first type of mixture is calculated based on the fact that the request pertains only to the "Projection". The second type includes "Intersections" requests. The third type includes "Selection" and the fourth type pertains to "Union".

The graphs show that the greatest load on the server comes from "Projection" type requests. When client - server systems are at a load around 0.8 - 0.9, variations in service delay time increase tenfold compared to a load of 0.1.

According to queuing theory (Kleinrock, 1979), the length of the queue in the QS can be significantly reduced if the service discipline is changed. In our case, a rational change in the disciplinē of servicing is to apply heterogeneous services. The alignment of services in the queue can be changed, for example, by using the ORACLE database management system's utility QUEUE_SIZE. In (Boicov, 2012), it is shown that, given the prioritising of query service, it is most advantageous to assign a higher priority to the shortest queries. To implement such a service discipline in the database server , one must insert a service type estimation block and a request service length estimation block. Then, using the utility QUEUE_SIZE, the service priority is assigned according to its minimum assign data length. Thus, it is possible to reduce the average queue length in a database server to 20 %.

Conclusion

This paper presents a study of database load. A service interactions model is developed. Expressions for estimating the load of the database server are derived both for working conditions without and with failure recovery. It is shown that services have properties of varying degrees of load of the nodes in the network, i.e. the properties of heterogeneity. Constructing mathematical models allowed us to make the calculations of average services' queue length to the database. These calculations showed that the database server can function catastrophically badly if run with a load of close to 0.9. Improving the efficiency of services can be achieved using heterogeneous service disciplines. The performed calculations have shown that the expected time gain in using services by network servers through the selecting the optimal load of network nodes and application of heterogeneous services can exceed 20%.

Reference

Codd E.F., 1970. "Relation model of data for large shared data banks", Comm. ACM, Vol.13(6), pp.377-383

Boicov V.N., 2009. "Heterogeneity factors in stochastic mass service systems", Automatic Control and Computer Sciences, Vol.43(3), pp.123-128

Boicov V.N., 2011a. "The investigation of possible errors in equivalence models of queuing systems with heterogeneous requests", Automatic Control and Computer Sciences, Vol.43(6), pp.303-308

Boicov V.N., 2011b. "Probability distribution functions for servicing two types of requests", Automatic Control and Computer Sciences, Vol.45(6), pp.201-205

Boicov V.N., 2012. "Customer flows in a network with two types of service", Automatic Control and Computer Sciences, Vol.46(3), pp.112-118

Gelenbe E. and Derochette D., 1978. "Performance of rollback recovery systems under intermittent failures", Communications of the ACM, Vol.21(6), pp.493-499

Kleinrock, L., 1979. Queuing systems. Volume 2, Computer application. University of California, Los Angeles

Kumar V., 2001. Introduction to database systems. Database recovery. Department of Computer Networking, University of Missouri-Kansas City