



Eidgenössische Technische Hochschule Zürich  
Swiss Federal Institute of Technology Zurich

Autumn Term 2009



# Advanced Systems Lab Project

## Milestone 04

Assigned on: **19th Nov 2009**  
Due by: **10th Dec 2009**

### Queuing model of the system

In this milestone, you will model the your system under test using a queuing model and perform some 'Analysis of variation' on different factors.

### Modifications to database servers

You must run the database on all 8 cores of the server machines.

### Modifications to clients

Make the following changes to the clients.

- Queries may no longer be pre-generated. Clients must generate the query when they run them.
- Clients must store all data returned from the database server in response to queries in a file. Use the “/local” storage for this file. All results must be stored in the same file and storage of multiple results cannot be intertwined. (Hint: think about synchronization on writing to file.)

### Think time

The above modifications have introduced some substantial think time on your clients. Before you proceed, make sure you understand what think time is and verify your beliefs with your TA.

Hint:

```
for query in Q1-QN:  
    construct(query)  
    result = execute(query)  
    store(result)
```

Explain what parts of this are think-time and what parts are response-time.

## Run experiments

With this new type of clients run the following experiments. You have three factors, each with two levels:

- Number of server machines {1, 2}
- Number of middleman machines {1, 2}
- Scaling factor of the database {0.1, 0.01}

For each of the 8 setups, measure throughput, response time, think time for 81 clients. You should ensure that the system is in steady state, you have at least 10 steady state measurements for each setup where the time between two successively measurements is greater than the time to run all 22 queries.

## Tasks

After having completed the above experiments, carry out the following three tasks.

### Task 1

For this task you are required to compute the statistical significance of the factors that are involved in the 8 experiments that you were supposed to run for a fixed number of clients (81 clients). The statistical procedure to perform this analysis is called "Analysis of variation" or ANOVA.

$$y_{f_{1..n}r} = \mu + \sum_{i=1..n} t_{i_{f_i}} + \sum_{i=2..n} \sum_{j=1..n} \nu_j + \epsilon_{f_{1..n}r} \quad (1)$$

The equation states that the result value  $y_{f_{1..n}r}$  is a function of the  $n$  variables ( $f_{1..n}$ ) that are independent and influence the result, as well as the number of repetitions of the experiment  $r$ . The result value is a summation of the overall average of all the results obtained in all iterations ( $\mu$ ), the individual effects of all the variables ( $\sum_{i=1..n} t_{i_{f_i}}$ ), the composed effects of all levels of combination of the variables ( $\sum_{i=2..n} \sum_{j=1..n} \nu_j$ ) and of the possible errors ( $\epsilon_{f_{1..n}r}$ ).

The experiments that you conducted can be characterised using this model. You have three factors that influence the response variable (i.e.: throughput, response time, think time, etc.): *number of database servers* —  $S$ , *number of middleman servers* —  $M$  or the *database scaling factor* —  $F$ , having the cardinalities of the levels  $|S| = 2$ ,  $|M| = 2$  and  $|F| = 2$ . You conducted the same experimentation for  $R = 10$  times, by having taken 10 snapshots during the steady state of the experiment.

In this case the model above can be rewritten as equation 2.

$$y_{ijkl} = \mu + s_i + m_j + f_k + \nu_{SM_{ij}} + \nu_{SF_{ik}} + \nu_{MF_{jk}} + \nu_{SMF_{ijk}} + \epsilon_{ijkl} \quad (2)$$

where

$$i = 1, \dots, |S|; \quad j = 1, \dots, |M|; \quad k = 1, \dots, |F|; \quad l = 1, \dots, R \quad (3)$$

We can compute all the terms in the equation by using the experimental results obtained for  $y$ , using the equations in 4.

$$\mu = \bar{y}_{...}; \quad (4)$$

$$s_i = \bar{y}_{i...} - \mu \quad (5)$$

$$m_j = \bar{y}_{.j.} - \mu \quad (6)$$

$$f_k = \bar{y}_{..k} - \mu \quad (7)$$

$$\nu_{SM_{ij}} = \bar{y}_{ij.} - s_i - m_j - \mu \quad (8)$$

$$\nu_{SF_{ik}} = \bar{y}_{i.k.} - s_i - f_k - \mu \quad (9)$$

$$\nu_{MF_{jk}} = \bar{y}_{.jk.} - m_j - f_k - \mu \quad (10)$$

$$\nu_{SMF_{ijk}} = \bar{y}_{ijk.} - s_i - m_j - f_k - \nu_{SM_{ij}} - \nu_{SF_{ik}} - \nu_{MF_{jk}} - \mu \quad (11)$$

$$\epsilon_{ijkl} = y_{ijkl} - \bar{y}_{ijkl}. \quad (12)$$

$$(13)$$

Finally, squaring both sides of the model and reducing the cross products (they are zero), we obtain the sum of square equation 14.

$$\begin{aligned} \sum_{ijkl} y_{ijkl}^2 &= |S||M||F|R\mu^2 + |M||F|R \sum_i l_i^2 + \\ &|S||F|R \sum_j s_j^2 + |S||M|R \sum_k n_k^2 + \\ &|F|R \sum_{ij} \nu_{SM_{ij}}^2 + |S|R \sum_{jk} \nu_{MF_{jk}}^2 + \\ &|M|R \sum_{ik} \nu_{SF_{ik}}^2 + R \sum_{ijk} \nu_{SMF_{ijk}}^2 + \\ &\sum_{ijkl} \epsilon_{ijkl}^2 \end{aligned} \quad (14)$$

$$\begin{aligned} SS_Y &= SS_0 + SS_S + SS_M + SS_F + SS_{SM} + \\ &SS_{MF} + SS_{SF} + SS_{SMF} + SS_E \end{aligned} \quad (15)$$

The influence (percentage) of each variable of compound influence of variables can be computed using

$$100 \frac{SS_X}{SS_Y - SS_0} \quad (16)$$

where  $X$  can be any of the variables or compound variables. Knowing the degrees of freedom ( $DoF$ ) for each of the variables and compound variable as being:

$y$	$ S  M  F R$	$\mu$	1
$y - \mu$	$ S  M  F R - 1$	$s_i$	$ S  - 1$
$m_j$	$ M  - 1$	$f_k$	$ F  - 1$
$SM_{ij}$	$( S  - 1)( M  - 1)$	$MF_{jk}$	$( M  - 1)( F  - 1)$
$SF_{ik}$	$( S  - 1)( F  - 1)$	$SMF_{ijk}$	$( S  - 1)( M  - 1)( F  - 1)$
$\epsilon_{ijkl}$	$ S  M  F (R - 1)$		

The standard deviations can now be computed as follows

$$s_\epsilon = \sqrt{\frac{SSE}{DoF_\epsilon}} \quad (17)$$

$$s_\mu = s_\epsilon \sqrt{\frac{1}{DoF_y}} \quad (18)$$

$$s_X = s_\epsilon \sqrt{\frac{DoF_X}{DoF_y}} \quad (19)$$

where  $X$  in  $S, M, F, SM, SF, MF, SMF$ .

Finally, for each  $(X, V)$  in  $(\mu, \mu), (S, s_i), (M, m_j), (F, f_k), (SM, \nu_{SM_{ij}}), (SF, \nu_{SF_{ik}}), (MF, \nu_{MF_{jk}}), (SMF, \nu_{SMF_{ijk}})$  we can compute the confidence intervals as being:

$$confidence_{90\%}(V) = V \pm 1.645 \times s_X \quad (20)$$

$$confidence_{95\%}(V) = V \pm 1.960 \times s_X \quad (21)$$

As a final result you should present:

- a) all computed Sum of Squares
- b) the percentage of variation explained by different factors and their interactions as well as the percentage of unexplained variation
- c) confidence intervals for the effects

**NOTE:** please refer to the text book for a detailed step-by-step deduction of the equations, their interpretation and how to use them in computing the requested items.

## Task 2

Consider the setup of 1 middleman machine, 3 database servers, and scaling factor of 0.01 for the database. Treating this setup as a blackbox (from the client's perspective), apply the queuing model of M/M/1 for clients in the set  $\{1, 3, 9, 27, 81\}$ .

Accomplish the following objectives for the different number of clients. (Hint: Pay attention to the snapshot time interval as number of clients change.)

- a) Measure the think time (Z), response time (R), jobs arriving (A), jobs completed (C), busy time (B), and average time spent in system ( $E[w]$ ). (Hint: some of these values might be the same.)
- b) Based on the measured values, derive arrival rate ( $\lambda$ ), throughput (X), system utilization (U), service time (S), average active jobs in the system ( $E[n]$ ).
- c) Produce plots of {throughput, response time, think time, utilization, average service time } vs. number of clients and response time vs. throughput.
- d) Plot and compare number of clients to the number of active jobs in the system.
- e) Using the interactive closed system model, and bounding response time, number of clients, and think time to measured values, compute throughput. Compare the computed throughput values to the measured throughput values.

## Task 3

Consider the setup of 1 middleman machine, 3 database servers, and scaling factor of 0.01 for the database. The middleman machine runs only one queue and dispatches queries to the database servers in a round-robin fashion. Apply queuing network models to this setup for clients in the set  $\{1, 3, 9, 27, 81\}$ .

Accomplish the following objectives for the different number of clients. (Hint: Pay attention to the snapshot time interval as number of clients change.)

- a) At the client, measure the think time ( $Z$ ), jobs arriving to the middleman machine ( $A$ ), jobs completed ( $C$ ), busy time ( $B$ ), number of jobs in the system ( $E[n]$ ), and average time spent in system ( $E[w]$ ).
- b) At the middleman machine, measure jobs arriving to each database server ( $A_i$ ), jobs completed at each database server ( $C_i$ ), busy time for each database server ( $B_i$ ), visit ratio for each database server ( $V_i$ ).
- c) For each device (middleman and database servers), derive throughput ( $X_i$ ), mean service time in the device ( $S_i$ ), utilization of the device ( $U_i$ ).
- d) Compare arrival rate at the middleman machine with the arrival rates at all database servers.
- e) Compare number of clients, measured active jobs in all devices, and  $E[n]$  as derived from Little's law.