MA-402
Queuing Models for Performance Analysis
Assignment # 1

Abhay Kapoor
04010244

Department of Electronics and Communications Engineering, IIT Guwahati

Modelling Web Maintenance Centres through Queue Models

Authors: M. Di Penta, G. Casazza, G. Antonio, E. Merlo

Abstract: The Internet and WEB pervasiveness are changing the landscape of several different areas ranging from information gathering/managing and commerce to software development, maintenance and evolution. Traditionally, phone-centric services, such as ordering of goods, maintenance/repair intervention requests and bug/defect reporting, are moving towards WEB-centric solutions. This paper proposes the adoption of queue theory to support the design, staffing, management and assessment of WEB-centric service centres. Data from a mailing list archiving a mixture of corrective maintenance and information requests were used to mimic a service centel: Queue theory was adopted to model the relation between the number of servants and the performance level. Empirical evidence revealed that by adding an express lane and a dispatcher service time variability is greatly reduced and more complex business rules may be implemented. Moreover; express lane customers experience a reduction of service time, even in the presence of a significant percentage of requests erroneously routed by the dispatcher.

Model:

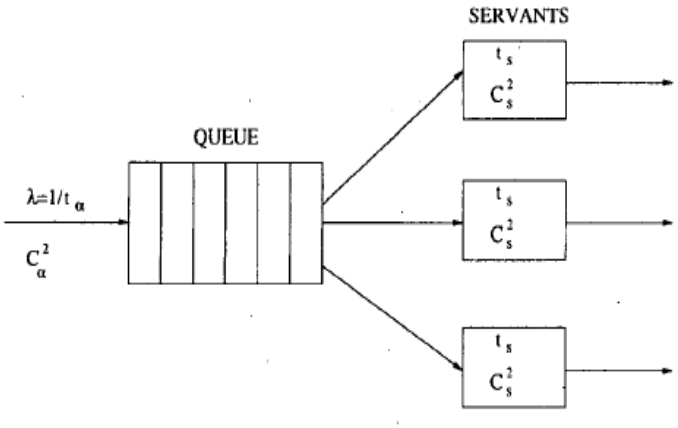

Fig.1: Service Centre Model

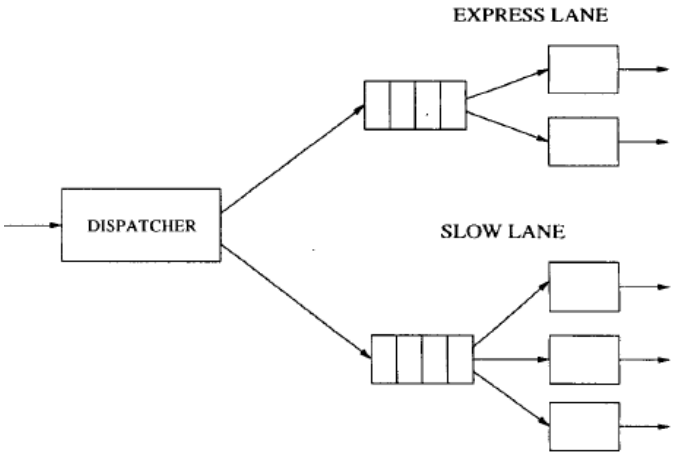

Fig.2: Express Lane Model

Figure 1 shows a simple diagram representing a service centre as a queuing system modelled by a single queue and *m* servants. Requests may be dispatched to any servant: all servants are equivalent. WEB-centric maintenance centres may be described by *M/M/m* models. However, while request inter-arrival times may be described by exponential probability distribution, service times often grossly deviate from *M/M/m* theoretical assumptions.

Figure 2 shows a more sophisticated approach where the incoming requests that can be handled in a short time are dispatched to a dedicated queue. The idea has been borrowed by the supermarket ***express lane.***

Case Study: The case study investigates the relation between the staffing (i.e., the number of servants with and without *express lane)* and the other queuing system parameters for a WEB-centric maintenance service centre. Service requests were extracted from a mailing list of the *STAR (Solenoidal Tracker At RHIC)* experiment. STAR is an experiment at the relativistic Heavy Ion Collider of the Brookhaven National Laboratory. 1165 requests corresponding to the 1999 mailing list traffic have been considered.
The archive was downloaded, all e-mails (stored in HTML format) processed by ***Per1*** scripts. for each request, time stamps were extracted to estimate the average inter-arrival time and service time. The ***theoretical*** number of mailing list servants was estimated as the number of different repliers.

Table 1. Mailing list system parameters (times are expressed in hours).

Requests processed	1165
Number or servants	100
Interval of time considered	from 01-01-1999 to 12-31-1999
t_α	7.44
$\sigma^2_{t_\alpha}$	158.83
C^2_α	2.85
λ	0.13
ν	5.79
SB	0.00
t_s	43.18
$\sigma^2_{t_s}$	49681.83
C^2_s	26.64

Table 2. Queue model parameters varying number of servants (times are expressed in hours).

Servants	ρ	$t_{\omega(exp)}$	t_ω
8	0.72	4.94	68.27
9	0.64	1.83	25.29
10	0.57	0.69	9.54
11	0.53	0.25	3.46
12	0.48	0.10	1.30
13	0.39	0.03	0.05

Table 1 reports the mailing list parameters. Most noticeably, service time variability is extremely high. As a result, the coefficient of variation is well over the expected value for an exponential distribution, suggesting the adoption of *M/G/m* models. Notice that the average time between subsequent customers *(t,)* is about 8 hours with a number of 100 ***repliers*** and an average ***estimated*** service time of 43 hours. Thus a customer of an hypothetical service centre with the same configuration parameters, would always be immediately served.
Table 2 reports the queuing system relevant parameters for the considered mathematical models *(M/G/m* and *M/M/m)* and various servant numbers. Results have been both the models have the same servant use coefficient***;*** they give rise to substantially different waiting time estimates.
Waiting time decreases as servant number increases; moreover, differences between the waiting times estimated by the two models slowly decreases as the servant number increases. A compromise between the number of servants and the time spent in the queue has to be pursued.
The introduction of an ***express lane*** has a pronounced effect: when the threshold varies from 6 to 12 hours the ***express lane*** *t,* is sensibly lower than the correspondent

value for the 11 servants' basic configuration. ***Express lane*** customers experience an average waiting time reduction up to 1550%. In the same time the non ***express lane*** *t*, increases. However, it is worth noting that ***express lane*** serves the majority of the system customers: when the threshold varies from 6 to 12 hours the percentage of ***express lane*** customers ranges from 71% to 77%. In other words, most of the service center customers perceive a clear service improvement in that the service time has substantially been reduced. Such reduction implies that *M/G/m* and *M/M/m* predicted results are closer and the uncertainty on the "in field values" **is** reduced. It is important to highlight that the overall *t*, related to the ***express lane*** configuration, when threshold varies from 6 to 24 hours, is very close to the correspondent obtained with the 11 servants' basic configuration.

Conclusion: The configuration of a service center in terms of both overall number of servants and number of servants in the *express lane* can not be accomplished without taking into account the company business rules. However, results contained in the paper support the adoption of an *express lane* handling a fraction of the incoming requests in that the overall system performance are boosted. Empirical results show that even in the presence of the realistic assumption of imperfect input requests classification; an ***express lane*** based system performs significantly better than the standard model.