Coping with Short-Term Sustained Peak Demands–Several Cost-Effective Strategies.

Authors: Ahmad Hashem
Open University Malaysia
Jalan Tun Ismail
50480 Kuala Lumpur
email: hashem@oum.edu.my

Mansor Fadzil
Open University Malaysia
Jalan Tun Ismail
50480 Kuala Lumpur
email: mansor@oum.edu.my

Abstract

For most open universities, in fact for most universities, there will arise occasions when some online activities will have to be completed by all students in a very short time interval. These occasions usually result in a great rush to get online as quickly as possible; the rush is often made worst by the limited good choices on offer such that usually only the earliest can get the best. Choosing and confirming elective subjects, when only a limited number for each are offered, on a first-come-first-served basis is an example. At the Open University Malaysia the rush occurs when confirming offered subjects and choosing face-to-face meeting timetable. Most students rushed to be earliest to ensure choice slots in the timetable are obtained. Available I.T. resources such as Internet bandwidth and servers cannot cope with these short-term sustained peak demands. Unless these peak demands are met, however, online services will slow down drastically resulting in long queues of users waiting to be served. Providing excess capacity, if at all possible, to ensure these short duration peak demands can be met would not be cost-effective since most of the time these resources are not utilised. This paper will look at some cost-effective approaches to meeting these short-term peak demands.

This abstract was submitted on March 25, 2005 by

Ahmad Hashem
Associate Professor
ICT Services Department
Open University Malaysia
50480 Kuala Lumpur Malaysia

hashem@oum.edu.my
Tel 603 2773 2253
Fax 603 2697 8754

Introduction

The Open University Malaysia practices the blended learning approach to Open and Distance Learning. Students are expected to study on their own from printed modules with .online support available at anytime. Also, the students can avail themselves of face-to-face tutorial sessions every fortnight at each Local Learning Centre. Attendance to these face-to-face sessions is not compulsory but most students make it a point to attend. Even in Open Distance Learning, face-to-face interaction is still important to students. They also insist on getting good time slots, that is, class-class-class but not class-gap-class or class-gap-class-gap-class and some other non-contiguous arrangements of time slots. Bookings of time slots are carried out online during a fixed period, usually lasting about one month.

The time slot selection is part of the whole registration process for old students. This process consists of confirming offered subjects, add and drop subjects and selection of face-to-face time slots, as illustrated in Figure 1.

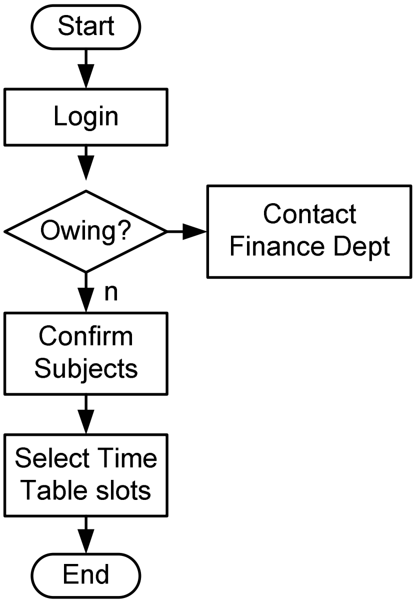

FIGURE 1: Simplified flowchart of old student registration process

Many of the sub-processes require long connection times to the database. The first one is subject confirmation (and add-and-drop subjects) where the students spend time to discuss with friends or looking at various options before finally making up their mind and continuing with the process. Another is when the students select their time table slots. Here, the students typically linger over their selection by consulting friends or taking time to account for personal selection criteria such as whether members of a car pool can start and finish at the same time. These relatively long connection times mean the probability of having many concurrent users connected to the database at any instant is very high. The large number of concurrent users blocks off access to later arrivals leading to long queues. Thus, many students will experience very long delays before getting access to the online registration system. Also, the many concurrent users will use up all available internal resources such as memory and processors which in turn will slow down each transaction drastically. A similar scenario occurs at many conventional universities when students must choose elective subjects with limited seats.

During the first few days from the start of the exercise there will be periods of sustained very high demands on resources, such as computing power and Internet bandwidth, that if not catered for can lead to very serious slowing down of all the Internet services. The simplest way of accommodating this relatively short duration peak demand is to scale up the entire I.T. infrastructure involved so that there is always available sufficient Internet bandwidth as well as computing environment no matter the demand. This is not the most cost effective solution since very heavy investments would be required but for most of the time all the extra capacity is not being used. Also, catering to peak demands will not always work no matter the capacity (unless limitless) because peak demands tend to use up all available resources, no matter how big.

Empirical Study

At the OUM some studies were carried out to try to define the peak demands. These studies were mostly empirical but the results do help to indicate approaches to catering to the peak requirements. To understand these peak demands the number of students that have successfully completed the registration and time-table selection process were recorded daily, see Figure 2. The number of students involved was almost similar: about 20,000 for semester 1 '05 and about 22,000 for semester 2 '05.

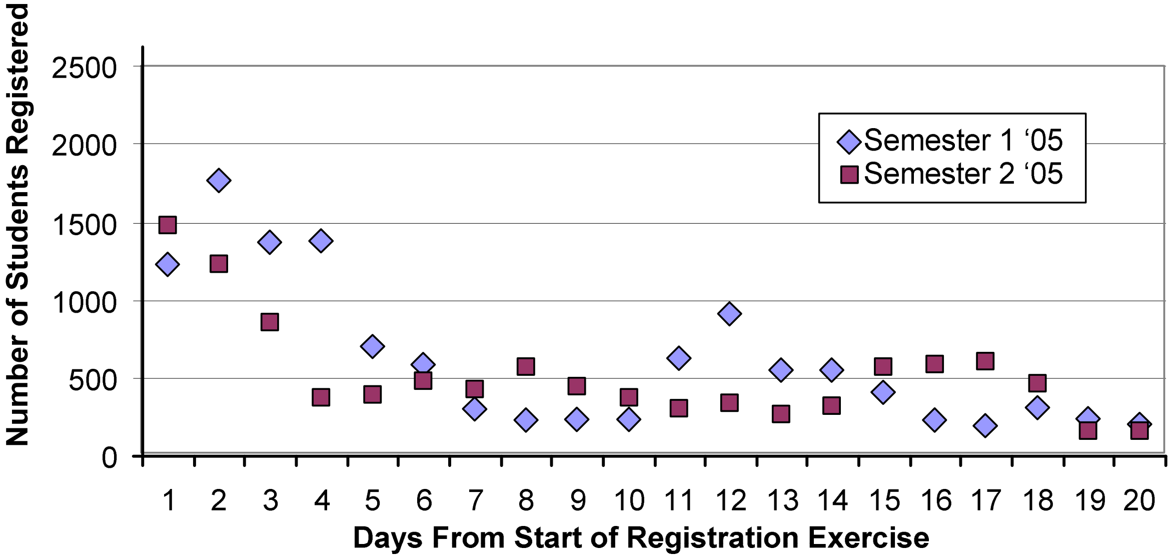

FIGURE 2: Number of old students that have successfully completed the registration process daily.

An interesting feature observed in Figure 2 is the similarity between peak values for both semester 1 '05 and semester 2 '05. For both semesters the maximum number of students that managed to complete the registration process was never more than 2000 per day. This maximum limit was also observed for all previous semesters. Also the graphs show high activities in the first few days of the exercise but rapidly tapering off to a few hundred a day.

The existence of the maximum seems to indicate that the present infrastructure can only allow not more than 2000 students to complete the registration process. There are many aspects of the infrastructure that can contribute to bottlenecks and hence delays in processing. Figure 3 shows a schematic of the server arrangement and the connection to the Internet. The readings for the graphs in Figure 3 were obtained using a simple, twotier server arrangement that consists of an application cum web server and a database server. The data is stored in a Storage Area Network (SAN) with connection from database server to SAN storage via fibre channel.

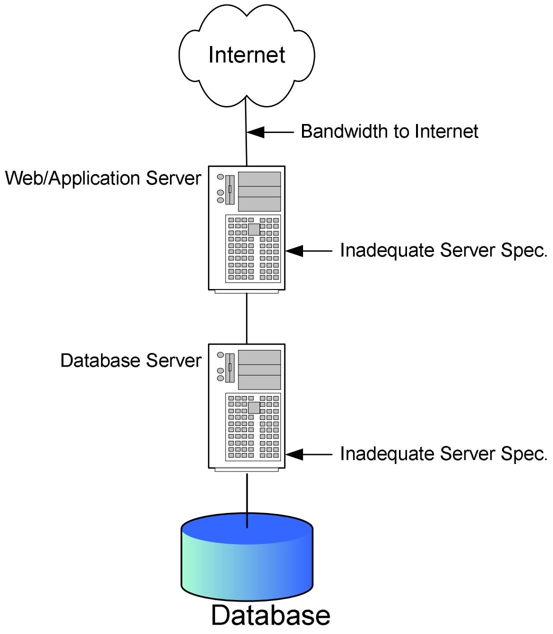

FIGURE 3: Simple server arrangement

Looking at Figure 3, there are three possible areas where bottlenecks can occur: inadequate bandwidth and inadequate server specifications for both servers. The servers may not have sufficient resources such as central processing units (central processing units) and random access memory (RAM) to cope with peak demand.

From monitoring of Internet activities it was found that the available bandwidth is sufficient to cope with demand. In fact, only about 40% of total available bandwidth was used up during this period. The bottleneck was narrowed down to the application server. This was a high specification server running on 4 central processing units with 12 GB of Random Access Memory.

Minimizing server bottlenecks

An obvious solution to inadequate server resources is to get a bigger server or to increase resources (more central processing units and more RAM). The first option can be very expensive. The second option may not be feasible. All the slots for central processing units and RAM have been populated. Another option is to employ a cluster of many small servers (usually normal personal computers) to replace the large server. This is the server clustering concept. Figure 4 show a simple cluster of four servers.

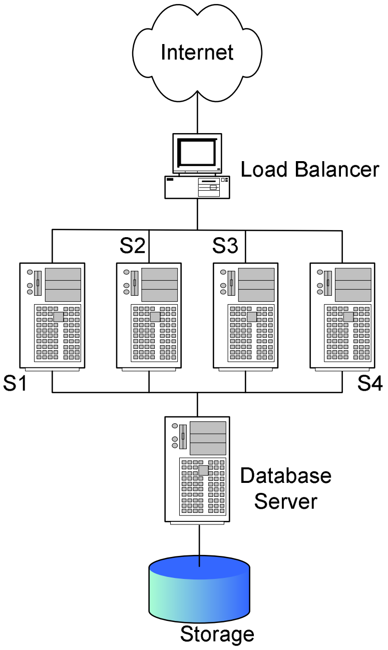

FIGURE 4: A cluster of four application servers

The servers in the cluster do not have to have high specifications. In most cases normal personal computers with added RAM (one to two Gigabytes) would suffice. The load balancer is a low specifications personal computer running an open source DNS load balancing program (http://www.zytrax.com/books/dns/ch9/rr.html), that redirects traffic to the servers in the cluster according to certain rules. For the present a simple round robin rule was used. The DNS balancing is not the best technique but it is simple to implement and from performance monitoring it was found that it manages to redistribute traffic quite well. Figure 5 shows the improvement in peak throughput that was observed after the simple clustering compared to peak throughputs without clustering.

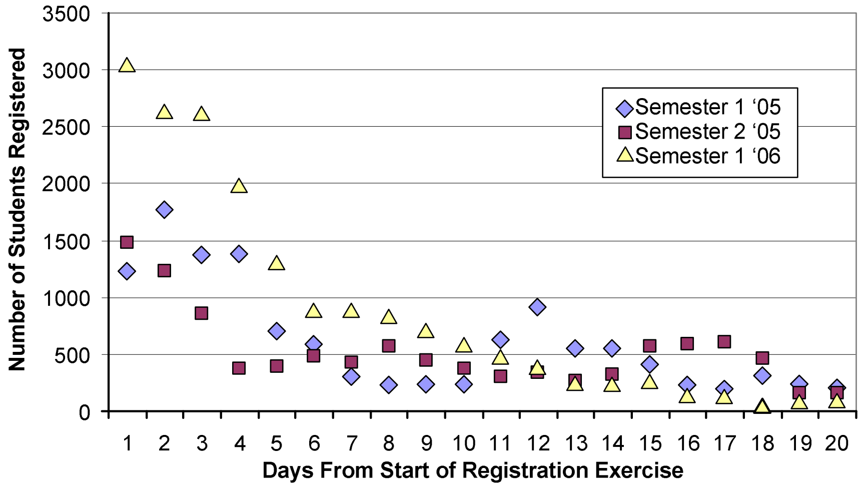

FIGURE 5: Improvement in peak throughput (Semester 1, '06) after simple clustering.

Many more students can now complete the registration and time-table selection process. The maximum at around 3000 is not because of bottlenecks but because that is the number of students that enters and successfully complete the process.

Spreading the load

The initial sharp peak and the gradual tapering off to some very low value that was observed with and without clustering does not constitute efficient resource usage. It is better to spread out the load so that many more peak loadings occur throughout the registration period. One way to spread the load is to schedule students into "zones". It does not really matter what the criteria for grouping students into a particular zone are. The important consideration is that the total number of students in each zone should be roughly similar. At the OUM students were divided into three groups of about 10,000 students each based on Learning Centre groupings. The first group were allowed to register in the first three days, followed by the second group for the next three days and the third group, also for three days. Then it is back to free-for-all for the rest of the registration period.

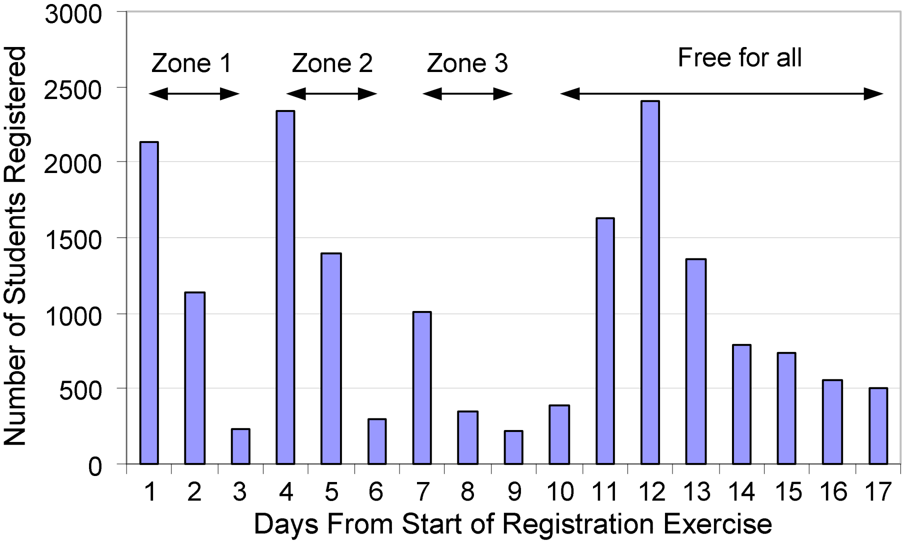

FIGURE 5: More occurrence of peaks when students grouped into "zones"

Observe in Figure 5 the trend of the very sharp rise in the first two days followed by the very quick drop off subsequently each time a group registers. This scheduling can be seen to spread the load better and is therefore more efficient.

Conclusions

The simple clustering techniques and DNS load-balancing appears to reduce bottlenecks and enable many more transactions to be completed. More effective utilization of resources was achieved by dividing the student population into groups and scheduling their online activities to minimize concurrencies.