Teaching About the Impact of Transaction Volume on System Performance and Capacity Planning

Dr. Michel Mitri
mitrimx@jmu.edu

CIS/OM Program
James Madison University
Harrisonburg, VA 22801 USA

ABSTRACT

Courses in information resource management often include discussion and practice in capacity planning. This requires students to understand a variety of topics related to business transaction processing, workload characteristics, system demand, forecasting methods, and system performance measurement. This article presents a student project that combines these topics in a spreadsheet application. The spreadsheet is designed to take transaction history as input, use this history to make forecasts of future workload demand, and then predict future system performance based on these forecasts. The paper discusses forecasting methods, system performance metrics, and presents a comprehensive description of the spreadsheet assignment.

Keywords: system performance metrics, workload forecasting, spreadsheets, capacity planning, service levels

1. INTRODUCTION

Information resource management is a broad area that includes strategy development, tactical planning, and operational implementation. Teaching such a course requires the instructor to present material that ties together disparate topics into a cohesive whole. For example, students should be aware of the effect of workload demand on system performance, and prepared to use this knowledge to manage user expectations of the services, performance, and availability provided by their computer system. Such expectations are formally described in a service level agreement (SLA), a contract between the information technology (IT) group and the user community that states system availability and performance commitments. A well-executed SLA reduces departmental conflicts, improves cost containment, and places IT services on a more business-like basis (Frenzel 1999; Vijayan 1997; Vincent 1988).

For an IT manager to intelligently negotiate an SLA, he or she must understand business- related concepts such as customer and employee population, inventory stocks, and business transaction rates. This kind of volume data is expressed in *business units*, the language of the user community. IT managers must be able to communicate with the users in these terms. However, IT managers also must be able to translate business units into measurements related to intensity of service demand (e.g. system arrival rates, transaction-processing time, etc.). Furthermore, these must be translated to system performance metrics such as utilization, queuing levels, and response time. In addition, IT managers realize that business trends are changing over time and that business growth rates will cause changes to system requirements. Therefore, estimating the computing needs of an organization requires application of forecasting techniques and relating these to the system performance metrics mentioned above. The combination of these skills form a discipline called *capacity planning*, the determination of predicted future system needs (Menasce et al 1994), which impacts system acquisition decisions and service level expectations.

With this in mind, the author created a spreadsheet assignment for an information resource management course that allows students to apply these concepts in a practical setting. The following sections describe the

issues of workload forecasting and system performance modeling that are discussed in the class, and present a description of the spreadsheet application. The topics discussed in the article contribute to IS educators and students in the following ways:

1) They integrate concepts of workload forecasting, system performance and capacity planning into a coherent course offering.
2) They present a simple description of queuing theory, geared toward an MIS student population, with implications for IT management.
3) They illustrate a computational exercise for putting the concepts into practice via a challenging but not overwhelming spreadsheet application for students to complete.

2. ISSUES OF WORKLOAD FORECASTING

In order to facilitate skill-acquisition for capacity planning, students of IT management should learn about the standard mathematical models for forecasting, and use these models when analyzing how business trends (such as transaction rates) affect IT resource demand. Three common forecasting models are particularly useful for facilitating such learning: moving average, exponential smoothing, and linear regression (Anderson et al. 1994; Menasce et al. 1994)

The *moving average* takes, as the projected value of a period under investigation, the average value of the actual rates over a fixed number of periods (the number of periods are determined by the decision-maker).

The *exponential smoothing* forecasting technique bases its projection for a given time period (e.g. month) by taking into account both the previous month's forecasted value and the present month's actual value. It uses a "*smoothing constant*" to determine how much weight to assign to the actual values. The exponential smoothing formula is described later in this article.

The final forecasting method described is *linear regression*. This is a statistical technique that attempts to fit a straight line through a series of data points. The X-axis of the linear regression will be the numbers related to time periods (such as months). The Y-axis will be the actual values for Months 1-8.

In teaching these forecasting techniques within the context of IT management, an instructor can apply them to predictions of business trends in an organization. It is useful to discuss various types of business transactions (e.g. customer orders, bills, work-orders, etc.) and discuss how the forecasting methods can be used to predict future months' transaction rates based on the changes to business rates in historical data of past time periods.

3. ISSUES OF SYSTEM PERFORMANCE

In addition to teaching forecasting techniques, instruction related to capacity planning should also expose students to system performance metrics. The instructor can start by discussing performance measurements that are directly visible to the end users, namely response time and throughput. *Response time* refers to the amount of time it takes, in an online system, for the computer system to return a screen to the user after the user has submitted a request. *Throughput* refers to the number of transactions, typically in a batch system, that can be processes in a given time period.

After discussing these user-based metrics, the instructor should discuss their underlying system causes. Specifically, one should discuss metrics such as *system utilization* and *queue lengths*, and relate response time and throughput. In a general information resource management course (as opposed to a rigorous computer science class), it is best to keep the discussion and any underlying mathematical models quite simple. For example, although the instructor may briefly discuss queuing theory and Markov state-space diagrams (Sevcik and Mitrani 1981), in practice, it is best allow students to rely on simplifying assumptions underlying Little's Law (Little 1961).

Little's Law is a simple, generalizable, and intuitively appealing theorem regarding queuing behavior in single-server systems. Little's Law states that for any system in which users are waiting in a queue to be serviced by a single server, the average length of users in the queue is equal to the arrival rate of users times the average amount of time a user spends being serviced. This can be applied to any type of single-server system (e.g. customers in a grocery line, transactions in a computer system, etc.).

Little's Law can be used to measure system response times and throughput, which impacts user satisfaction, and therefore has implications for information resource management, particularly as it relates to capacity planning problems. From Little's Law, the following measures are relevant:

Let λ = arrival rate (rate at which requests for service come to the system…e.g. number of transactions arriving per second). Let μ = service rate (rate at which the system satisfies the request…e.g. number of transactions that can be processed per second). Based on these two variables, one can estimate the following values

system utilization = λ / μ
throughput = λ
average queue length = $\lambda / (\mu - \lambda)$
average response time = $1 / (\mu - \lambda)$

The queue length and response time formulae assume a balanced flow…i.e. $\lambda < \mu$. Otherwise. The queue just keeps building.

Using these measures, students are able to make reasonable, if simplistic, estimates of important system performance parameters based on the workload demand forecasts described in the previous section.

After classroom discussion of these topics, it is important to give students exposure to practical applications of the material. The following section presents a spreadsheet assignment in which students combine the concepts of business transactions, workload forecasting, and system performance measurement into a set of linked worksheets that predict future system performance based on historical transaction data. The assignment is described in its entirety, and includes images of spreadsheet components.

4. THE SPREADSHEET ASSIGNMENT

For this assignment, students create a spreadsheet that uses three different methods for forecasting transaction activity for the months of September through December based on known transaction demand of January through August. The spreadsheet will then use the performance analysis formulas discussed in class to estimate average system utilization, queue lengths, and response times for the months of September through December based on each of the three forecasting results. After students build the spreadsheet, they use it to help answer questions related to system performance and capacity planning issues.

The spreadsheet is composed of three separate worksheets:

1) The Transaction Input Data, consisting of 8 months of transaction data, in the form of transaction arrival rates during peak periods.
2) The Workload Forecast for months 9-12 of the transaction demand (anticipated arrival rates) for peak periods, based on three forecasting methods: *moving average, exponential smoothing, and linear regression*.
3) The anticipated system performance (including system utilization, average queue lengths, and average response time) based on the projections using each of the three forecasting methods.

4.1 The Transaction Data Sheet

Figure 1 is a display of the first worksheet of the spreadsheet that students create in their assignment. It simply includes the transaction data for the first eight months of the year. This transaction data is in terms of average transactions per second for the peak periods. The user of the spreadsheet should be able to enter and change this at will. All other items on this worksheet should be locked.

Widgets 'R Us	Transaction		History					
Category	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug
Peak Period Transaction Rate (transactions/sec)	34	36	41	40	40	42	43	44

Figure 1: Transaction Data Sheet

4.2 The Workload Forecast Sheet

This worksheet, shown in figure 2, consists of data that results for the Workload Forecast. Given the input data shown above, students should obtain the results below when implementing the forecasting methods (which are described below). Note that the *projections* are predicting the peak period transaction arrival rates for months 9 through 12, based on the actual values for transaction arrival rates in months 1 through 8.

In this sheet all cells should be locked (in order to prevent users from accidentally modifying cells containing formulas), except for the one to the right of the label "Constant" in the exponential smoothing portion. That constant value (displayed as 0.6 in this example) should be the only one that the user can modify. However, by changing values in the Transaction Data sheet, the values in the Workload Forecast sheet will be automatically altered accordingly. To facilitate this, students should *link* the cells containing actual values (Transaction Data in the above example) to the corresponding cells in the Transaction Data Sheet. This linking process is accomplished using the Paste Special option of the Edit menu. First, students Select and Copy the cells they want. Then, they go to the position where they want the linked values to be stored. Then they choose the Paste Special option of the Edit menu, making sure to click the Paste Link button. This way, any changes made in the Income Statement will automatically be copied to the forecast sheets.

Figure 2 shows the projected (forecasted) values based on three forecasting methods (Moving Average, Exponential Smoothing, and Linear Regression). In

addition, the sheet shows the error (squared difference)

Widgets 'R Us					Workload Forecasts (expected transactions/second)									
		Months												
Category		1	2	3	4	5	6	7	8	9	10	11	12	
Trans Rate		34.0	36.0	41.0	40.0	40.0	42.0	43.0	44.0					
Moving Avg					37.0	39.0	40.3	40.7	41.7	43.0				
Change	1.2					2.0	1.3	0.3	1.0					
Projection		**34.0**	**36.0**	**41.0**	**37.0**	**39.0**	**40.3**	**40.7**	**41.7**	**43.0**	**44.2**	**45.3**	**46.5**	
Difference					3.0	1.0	1.7	2.3	2.3					
Squared Diff.	4.7				9.0	1.0	2.8	5.4	5.4					
Exp. Smooth		34.0	35.2	38.7	39.5	39.8	41.1	42.2	43.3					
Constant	0.6													
Change	1.3		1.2	3.5	0.8	0.3	1.3	1.1	1.1					
Projection		**34.0**	**35.2**	**38.7**	**39.5**	**39.8**	**41.1**	**42.2**	**43.3**	**44.6**	**46.0**	**47.3**	**48.6**	
Difference				2.3	0.5	0.2	0.9	0.8	0.7					
Squared Diff.	1.3			5.4	0.3	0.0	0.8	0.6	0.5					
Lin Regress	1.3	34.2												
Projection		**35.5**	**36.8**	**38.1**	**39.4**	**40.6**	**41.9**	**43.2**	**44.5**	**45.8**	**47.1**	**48.4**	**49.6**	
Difference				2.9	0.6	-0.6	0.1	-0.2	-0.5					
Squared Diff.	1.6			8.6	0.4	0.4	0.0	0.0	0.3					

Figure 2: Workload Forecast Sheet

for each of the methods. The forecasting techniques and the calculation of squared error are listed and described below.

Moving Average: The moving average takes, as the projected value of a period under investigation, the average value of the actuals over a fixed number of periods (the number of periods are determined by the decision-maker). In the example above, the line labeled "Moving Average" shows, for each month, the average of the three preceding months. For example, the moving average for Month4 is the average of the actual values for Month1, Month2, and Month3. Likewise, Month5's moving average is based on Month2, Month3, and Month4.

Once each moving average has been established, the spreadsheet should calculate the actual change in moving average values (see the line labeled "Change"), and then calculate the average change. This average change is what will be used to calculate projections for the *future* (unknown) months. Note that the projections for those months where moving average calculation is possible (i.e. those months where actual data exists for the three months prior to them) is equal to the moving average values themselves. For the assignment, these are months 4-9, since there is actual data for months 1-8. For the future months, the projections are equal to the previous month's projection value plus the average change. For example, Month10's projection is equal to Month9's projection plus the average change value

Exponential Smoothing: This forecasting technique bases its projection for a given month by taking into account both the previous month's forecasted value and the present month's actual value. It uses a "*smoothing constant*" to determine how much weight to assign to the actual values.

> *PresentMonthForecast =*
> *PreviousMonthForecast +*
> *SmoothingConstant * (PresentMonthActual - PreviousMonthForecast).*

Thus, in the line labeled "Exp. Smooth" above, Month2's value is set to Month1's exponential smoothing value + 0.6 * (Month2's actual value - Month1's exponential smoothing value). The 0.6 value is the value seen on the line labeled "Constant". This value should be available for manipulation by the user.

From the exponential smoothing values, the spreadsheet should calculate the change from month to month (see the line marked "Change"). After the changes are calculated, the average of the changes can be obtained. This average is what will be used to give the projected values for the *future* months (Months 9-12 in the example above). The projected values for Months 1-8 are simply the values obtained from exponential smoothing.

Note how the smoothing constant affects the exponential smoothing values. This number should be a value between 0 and 1. If it is 1, then the exponential smoothing value is simply equal to the actual value. If it is zero, then the exponential smoothing value is equal to the projection from the prior month. It is usually recommended to keep the smoothing constant at a number between 0.3 and 0.6 for forecasting purposes.

Linear Regression: The final forecasting method used is linear regression. This statistical technique attempts to fit a straight line through a series of data points. In this assignment, the X-axis of the linear regression (the independent variables) has the month numbers (i.e. 1-8). The Y-axis will be the actual values for Months 1-8.

To use Linear Regression in Excel, students are taught to use the statistical function called LINEST. This returns an array of values. The first value returned by LINEST is the *b* value (which indicates the slope of the line). The second value is the *a* value, which indicates the y-intercept point. The *a* value in the above spreadsheet (based on the input data) should be 34.2, and the *b* value should be 1.29. Based on these values (*a* and *b*), the formula used for each month in the

Linear Regression line is*: a+b*Month#*. The linear regression line thus contains the projections for this method.

To compare the accuracy of each of these forecasting methods students determine the squared difference between the projected value and the actual value for each of the months 1-8. This is called the "squared error". It is implemented in the above table in the lines labeled "Difference" (which calculates the difference between actual vs. projected values) and "Squared Difference" (which squares the "Difference" values). For each of these "Squared Difference" lines, students should sum the squared errors in order to determine the total squared error for each forecasting method.

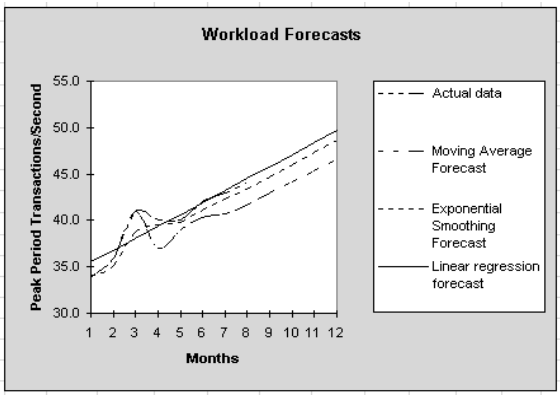

Figure 3: Graphing the various workload forecasts

The worksheet for forecasted projections should also include a line graph that displays, for each month, the projection based on each forecasting method. A sample is shown in figure 3. Note that the actual data ends at month 8, while the projections all extend out to month 12.

4.3 The System Performance Forecast Sheet

Based on the workload projections, students will calculate and graph expected system performance measurements, including system utilization, average queue length, and average response time. The data provided above should result in the worksheet illustrated in figure 4.

Note that all data should be *locked* EXCEPT for the average service time. This value, which indicates the average amount of time a transaction needs the system for, is modifiable by the user. It is a measurement of the power of the system (kind of a composite figure that represents CPU speed, network bandwidth, disk IO speed, etc.). In the above example, this figure is set at 20 milliseconds. The number below this is the average transactions per second that can be processed by the system (i.e. the average service rate), and is calculated from the average service time. Then, the average service rate will be used in further calculations. This is the μ value described above.

Widgets 'R Us	System Performance Forecasts			
Average Service Time (msec)	20			
Average Service Rate (tps)	50			
	Sept	Oct	Nov	Dec
Moving Average				
Transaction Forecasts	43.0	44.2	45.3	46.5
System Utilization	0.9	0.9	0.9	0.9
Average Queue Length	6.1	7.6	9.7	13.3
Average Response Time	0.1	0.2	0.2	0.3
Exponential Smoothing				
Transaction Forecasts	44.6	46.0	47.3	48.6
System Utilization	0.9	0.9	0.9	1.0
Average Queue Length	8.3	11.4	17.4	35.0
Average Response Time	0.2	0.2	0.4	0.7
Linear Regression				
Transaction Forecasts	45.8	47.1	48.4	49.6
System Utilization	0.9	0.9	1.0	1.0
Average Queue Length	10.9	16.1	29.4	139.0
Average Response Time	0.2	0.3	0.6	2.8

Figure 4: System performance measurements

Figure 4 also shows, for each forecasting approach, the projected values for the transaction arrival rates. These projected values should be obtained by using the link paste options, and should come from the values generated in the Workload Forecast sheet. These are the λ values discussed above. From these, students can use the formulas they learn from Little's Law to determine system utilization, average queue length, and average response time based on each forecast.

After calculating these values, students should construct a graph for these values. For example, the average response time graph based on the above numbers will look like figure 5. Note that with the numbers given in the data sheet, the linear regression forecast gives a worst-case scenario.

4.4 Questions

After constructing the spreadsheet, students should be encouraged to use it for analyzing various scenarios. To facilitate this, the assignment can include questions, such as the ones shown below:

1) Note that the worst-case scenario (based on the linear regression projections) for the example above indicates that the response time will shoot up to 2.8 for the month of December. Suppose that the

Service Level Agreement guarantees that the average response time will never exceed 0.5 seconds per transaction. What is the least amount

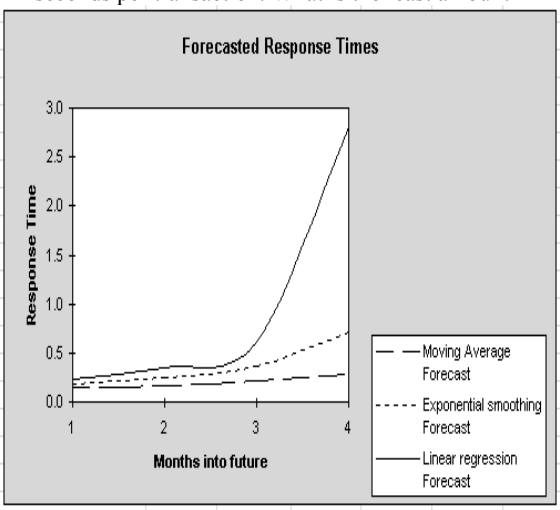

Figure 5: Graphing the expected system performance based on various workload forecasts

of improvement to the system value that is needed to keep December's average response time at a maximum of 0.5? (To answer this, students can play around with the Average Service Time value until they get December's response time projection of 0.5).

2) This model does not deal well with situations in which the transaction rate exceeds the service rate. Students can play around with the transaction data numbers until they see the results in the system performance forecast where transaction rates exceed the service rate. What does this do to queue length, response time, and utilization figures? Recommend a way to make the system performance forecast more accurately and realistically reflect the effects of cases where $\lambda > \mu$.
3) This model does not account for seasonal fluctuations. Assume that the holiday season should add more to the projections that the normal growth rates that are being reflected in the first eight month transaction figures. How might one modify the Workload Forecast formulas to account for heavier than normal growth in the November and December months?
4) In terms of capacity planning, the need to upgrade a system is often indicated by the inflection point of a system performance curve. Based on the worst-case system performance curve above, when will the company need to purchase upgraded hardware?

5. CONCLUSION

IT management often involves making predictions of future demand for computing resources based on known and forecasted trends in business activity. Capacity planning requires the ability to predict future workloads, relate these workloads to system requirements, and translate the system requirements into performance measurements such as response time and throughput. MIS students should be exposed to these topics, since they will be facing these issues when they get into the work world.

The above-described course content and spreadsheet assignment is an attempt to meet these pedagogical needs. The course content includes discussion of business transactions and collection of data related to periodic transaction rates. It also involves presentation of forecasting methods such as moving average, exponential smoothing, and linear regression, and applies these techniques to predictions of future business transaction activity. Finally, the course discusses system performance metrics including system utilization, queuing theory, response time, and throughput. Students put all of these ideas into practice by developing a spreadsheet that models the effect of business transaction trends on system performance predictions, with implications for capacity planning and system acquisition decisions. This assignment challenges the students to consider all the relevant topics related to capacity planning, and also gives rigorous practice in spreadsheet development. [1]

6. REFERENCES

Anderson, D.R., D.J. Sweeny, and T.A. Willams, 1994. An Introduction to Management Science: Quantitative Approaches to Decision Making, 8th Edition. West Publishing Co., St. Paul. NM.

Frenzel, C. W, 1999. Management of Information Technology, 3rd Edition. Course Technology, Cambridge, MA.

Little, J.D.C., 1961, "A Proof of the Queuing Formula *L = λW"*, *Operations Research*. Vol 9. pp 383-387.

Menasce, D.A., V.A. Almeida, and L.W. Dowdy, 1994. Capacity Planning and Performance Modeling: From Mainframes to Client-Server Systems. Prentice Hall, Englewood Cliffs, NJ.

Sevcik, K.C., and I. Mitrani, 1981, "The Distribution of Queuing Network States as Input and

[1] The solution for this assignment can be obtained from the author, who can be contacted at mitrimx@jmu.edu.

Output Instants.", Journal of the ACM, 28:2. pp 358-371.

Vijayan, M.J, 1997. "Service Pacts Ease Conflict," ComputerWorld, p14.

Vincent, D.R., 1988 "Service Level Management," EDP Performance Review, April, 1988, p3.

Michel Mitri is joining the faculty in the College of Business at James Madison University as an Associate Professor of Computer Information Systems, beginning in the Fall 2001 semester. Prior to this he has served as Associate Professor of Computer Information Systems at Eastern Michigan University, where he was on the faculty since 1992. Dr. Mitri's teaching interests include programming, web development, database design, artificial intelligence (AI), decision support systems (DSS), and management of information systems. His research interests include DSS, AI, and their applications for teaching and learning. Mitri earned a PhD in computer science from Michigan State University in 1992, and a BA in psychology from University of Michigan in 1976.

Information Systems & Computing
Academic Professionals

STATEMENT OF PEER REVIEW INTEGRITY

All papers published in the Journal of Information Systems Education have undergone rigorous peer review. This includes an initial editor screening and double-blind refereeing by three or more expert referees.

ISSN 1055-3096