Network Application Performance Modelling

Melisa Koorsse, Lester Cowley, André Calitz
Department of Computer Science and Information Systems
University of Port Elizabeth, PO Box 1600, Port Elizabeth, 6000
Tel: +27 (0)41 504 2326, Fax: +27 (0)41 504 2831
Email: csbmk@upe.ac.za, csanlc@upe.ac.za, csaapc@upe.ac.za

Topic: Network Planning; Sub-topic: Planning Issues

Abstract—Application performance measurement (APM) is the analysis of network application service performance from a user perspective. APMs such as application response time and availability, are important indicators of how long application users have to wait for requests to be processed. The purpose of this paper is to provide network managers with an inexpensive tool and technique to quickly and easily determine the effects that network changes may have on the performance of a particular network application. This paper discusses APM, describes a simple methodology for developing an application performance model and presents application performance analysis and prediction methods applied to a real-world case study.

Index Terms — Application Performance Measurement, Capacity Planning, Performance Management.

I. Introduction

Businesses today rely on the support of computer and telecommunications networks to achieve economic and business goals [10]. Critical business operations therefore require high performance levels from business applications to ensure user (employee and customer) satisfaction. It is important when upgrading or installing large networks, for the continued performance of application services to be maintained by thorough planning, using capacity planning methodologies [8,9]. This ensures essential network and application performance, resulting in satisfied users. A *capacity planning methodology* is a step-by-step process to determine the most cost-effective network topology and system configuration [2].

This paper will introduce network application performance measurement and different capacity planning methodologies. Using these methodologies, a methodology and cost-effective model for modelling application performance will be presented. The application of the methodology and model to modelling the performance of application services on the University of Port Elizabeth (UPE) customer network[1], will further be discussed.

II. Network Application Performance Measurement

Network application performance measurement (APM) emphasises the quality of the user experience [6]. *Network application performance* is how well the application is performing from a user point of view, measured in terms of transaction speed and server availability [16].

Frogner [6] and OPNET Technologies [17] have identified several factors that influence network application performance, namely: application characteristics, database access, network architecture, protocols and server configurations.

Two different types of data that need to be collected when modelling application performance are network centric data measures, such as latency and utilisation, and user centric data measures, such as response times and application server availability [13].

The simulation of network application traffic by traffic models, such as the Multifractal Wavelet Model (MWM), which can model the fractal nature of network traffic [14], was investigated using response time and delay data from the UPE customer network. Analysis revealed that this data was not self-similar in nature, and therefore the MWM could not be applied. The research also found that other traffic models, such as Markov processes and Poisson models [1], would not be suitable for use in the development of the application performance (AP) model, after determining that it is not necessary to simulate the AP traffic and then apply changes to the simulation. The changes could be applied directly to the collected traffic data using methods such as capacity planning, discussed in the following section.

III. Capacity Planning Models / Methods

Capacity planning is a process to model and determine the most cost-effective network configurations that will result in the best network performance [5,15,16]. A capacity planning methodology is an outline of this process [2].

Gunther [8] discusses the reluctance of network managers to use traditional capacity planning methods, because most network managers will rather risk the project not meeting performance management criteria, than jeopardise the project deadline or budget.

Several papers discuss performance capacity planning methodologies. Three of the papers, which discuss methodologies that provide logical and useful phases applicable to the UPE customer network, have been selected, from which a methodology suitable for capacity planning application performance will be synthesised in this paper.

The first capacity planning methodology (Fig. 1) is presented by Granville [7], where he provides a general overview of a process to manage network and application performance that can be applied when adding new applications to an enterprise network, planning network upgrades, etc.

[1] "Customer network" is a term used by Telkom, a South African telecommunications company. All parts of a Telkom customer's network managed by Telkom, are part of the customer network.

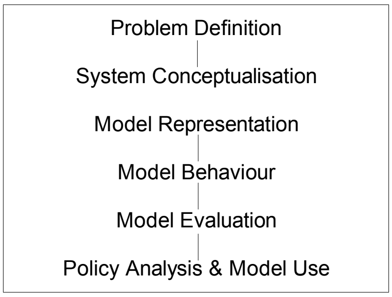


Fig. 1. Granville's methodology.

The steps of Granville's methodology consist of defining the problem, then identifying important influences within the system. This is followed by coding the model (model representation) and analysing and evaluating the resulting computer simulation tool (model behaviour and evaluation). The final step involves testing alternative system designs.

Step 3 of Granville's methodology converts the model into computer code, however, the preceding steps provide no indication of the formulation of a model. The first two steps collect data that can be used for the model formulation, however, the methodology does not clearly define this process in the steps outlined by Granville.

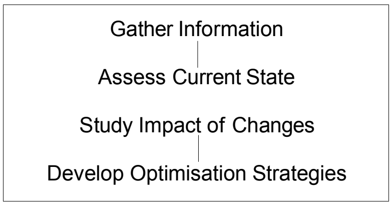


Fig. 2. Zaidi and Blum's capacity planning methodology.

Zaidi and Blum [19] discuss capacity planning strategies offered as a consulting service by a network services company.

The steps of Zaidi and Blum's methodology, Fig. 2, include gathering information on users, transactions and applications on the network (step 1), then assessing the network by understanding the behaviour of each network element and visualising any bottlenecks (step 2). Validated baseline models are used together with statistical analysis to study the impact of changes in utilisation and end-user response times (step 3), from which a strategy to optimise the particular network change, is developed (step 4).

Zaidi and Blum's capacity planning methodology is one part of an integrated and thorough process, that includes performance baselining, application impact analysis, modeling and simulation and service level engineering, performed as a professional performance management service. This process requires the thorough completion of each phase to ensure successful planning; however this could be time-consuming.

Almeida and Menasce's capacity planning methodology [2], provides a guideline for planning the capacity of a client/server system. The use of three models is required in the methodology presented – the workload model captures resource usage and workload intensity information, the performance model predicts response times, utilisations and throughputs and the cost model explains expenditures due to hardware, software, telecommunications and support.

Almeida and Menasce provide a step-by-step approach that can be applied to the UPE case study. Their methodology, however, requires three detailed models, whereas for the research methodology, the performance model with relevant actions from the workload model, such as workload characterisation, will be adequate to predict application performance, quickly and efficiently.

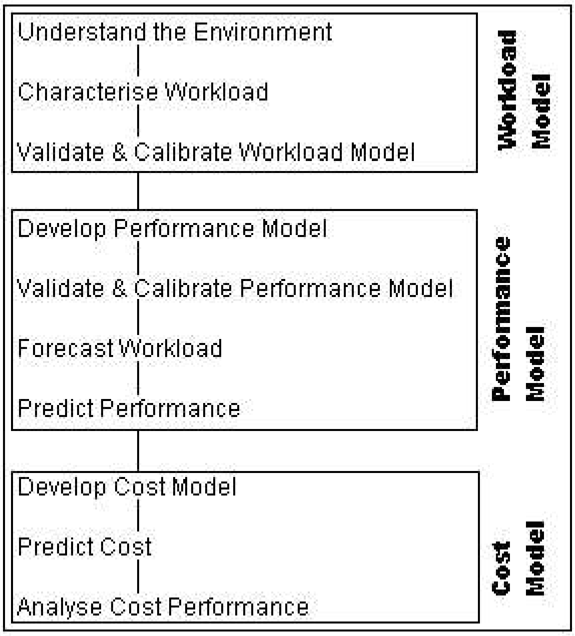


Fig. 3. Almeida and Menasce's methodology.

The research presented aims to provide a methodology for network managers to easily develop a simple model to do simulation analysis, when time and financial expense for using professional network performance consultants is not considered viable for the particular network change. The requirements of the methodology are that it is easy and simple to use, inexpensive, will not be a time- and labour-intensive process, yet provide efficient, validated feedback to the user. Separately, none of the three methodologies presented above meet these requirements, by either being too time-consuming or lacking steps that a simple model will need to be efficient. Therefore, the best phases of these methods have been identified and used to formulate a methodology which is simple to use, yet produces a model that can provide useful predictions. The proposed methodology, which can be used as a guideline for the development of application performance models on a customer network, is as follows:

1. ***Monitoring Current Network Situation [2, 7, 19]:*** Assess the current network topology, gain an understanding of the network components and identify areas of concern. Information on users and network components should be gathered. An assessment of how the business operates and identification of important transactions, should also be done. Data variables for the analysis of network and application performance, should be identified. Interviews with users of the application service should be conducted to gauge the users' views on the application's performance.
2. ***Analysis and Modelling [7, 19]:*** Data variables identified in Step 1, must be collected and analysed to determine which are suitable for use in the prediction

step. Variables are considered suitable if, for example, a network performance data measure can be used to predict an application performance data measure. The analysis determines the relationship between the variables, using methods discussed in Section V below.

3. ***Model Validation [2, 7]:*** Data collected before and after changes to the network have been made, can be used to validate the model by comparing the predicted model results to the data collected. If the comparison is within an acceptable margin of error, the model is validated and can be used for further analysis. If not, Step 2 should be repeated using the new data collected, as this will adjust the relationships appropriately.
4. ***Prediction [2, 7, 19]:*** Using the validated relationships between network and application performance data measures, discovered in Step 2 and 3, the results of different network scenarios can be predicted. The method used for prediction in this research will be discussed in the Section V.
5. ***Selection [2, 7, 19]:*** Alternative scenarios can be simulated to determine the solution that will produce the results desired. A simple cost analysis can also be included to determine the most cost-effective alternative.

This methodology provides a step-by-step process for network managers to develop an application performance model and use it to simulate application performance for proposed network changes. The methods used to analyse and formulate the model and do prediction will be discussed in more detail in the next sections of this paper.

IV. Modelling & Prediction Analysis Methods

Methods of analysing the data, formulating the model and using it to predict application performance in different scenarios, need to be identified for application in Steps 2 to 4 of the methodology outlined above.

This research aims to provide a simple method for network managers to quickly and easily run simulations to obtain only the data required when considering application performance [8].

Several commercial performance monitoring, analysis and prediction tools (e.g. Teamquest Lite, OPNET IT Guru and NetPredictor) are available to perform the analysis and prediction of data. The user enters current network information – some tools even have network self-discovery capabilities – and network changes, and the tool uses analysis modelling methods to provide a simulation of the network changes with the results visually output in the form of 2D graphs. The problem is that these tools are expensive and can be complex to use.

A literature review of capacity planning and performance prediction methods revealed that seldom literature revealed detailed workings of model formulation and use. However, the following techniques have been identified:

- Granville [7] discusses different scenarios of network change, such as adding a new application service or adding more users. An analysis of current network characteristics, such as application or user information associated to the change, for each scenario is done. An informed assumption of the possible effect a change will have on the network, is made from this analysis [7]. In certain cases, such as adding a new application service, the application service is run in a controlled network environment, to collect data for the analysis process.
- Ding and Newman [3] consider workload characterisation and prediction a critical basic step in capacity planning and performance modelling. Workload characterisation traditionally involves determining the different business activities and mapping each of these activities to a process on the network, from which it is possible to determine when performance problems may occur [2,3]. The use of queuing theory to describe queues that arise at system resources, such as CPUs, disks, routers and communication lines, is related to workload characterisation [2]. The level of detail required for queuing networks depends on the reasons for building the model, the availability of detailed information and the effect of the particular component on overall performance [2].
- Gunther [8] discusses the need for simple, accurate models instead of the precision of traditional capacity planning methods, requiring months just to verify a simulation. The "guerilla capacity planning" suggested by Gunther is a method for determining application scalability, formulating a simple capacity model using the following formula:

$$S(\alpha,\beta,N)=N/\{1+\alpha[(N-1)+\beta N(N-1)]\}$$

where α represents contention delays, β is associated with additional delays, such as time to fetch a cache miss, and N is a data variable, for example, the number of virtual users. This formula can easily be entered into a spreadsheet and, using built-in linear regression tools, α and β, can be determined. The basic "guerilla steps" are:

1. Measure the throughput as a function of the load N.
2. A sparse data sample of at least four load points should be collected.
3. Calculate α and β using a regression fit to this data.
4. Use these values to predict complete application scalability using the formula above.

Gunther's method [8] shows that, not only can simple analysis tools, such as Microsoft Excel, Mathematica, MathCad and others listed by him, be used to assist in sizing data, but more complicated queuing theory models are not necessary in providing users with a technique to rapidly forecast capacity requirements.

- Kaminski and Ding [12] outline a technique that relates business measures to capacity planning and identifies the relationship between business metrics of interest and available system performance metrics. Business metrics of interest (BMI) are real world business activities that drive the business workload [12]. Candidate BMIs should be identified and their relationship to workload performance data determined. Examples of candidate BMIs are order lines per hour or database commits per hour, which can be determined through interviews. The candidate BMI and the workload can be graphed over time, to determine any relationship between the two measures. However, the metrics may not have the same units, therefore scaling factors for each can be calculated and graphed to provide a better comparison. The scaling factor X' can be calculated as follows, where X is the candidate BMI data:

$$X' = (X/(Xmax - Xmin)) - (Xmin/(Xmax - Xmin))$$

A correlation coefficient can be calculated which indicates whether a linear relationship between the two data measures exists. The correlation coefficient value can range between -1 and 1. A correlation value of 1 indicates that an increase or decrease in one of the values results in an increase or decrease, respectively, in the other value and it is this relationship between the data values that is of significance for the prediction step. The correlation value can be calculated using the CORREL function in Microsoft Excel or other spreadsheet programs with built-in statistical functionality.

If a good correlation exists between the two measures, then the measures can be graphed using:

$$Y = aX + b$$

the straight line equation, where a is the slope variable and b is the y-intercept value. From this representation of the data, prediction analysis can be done by plotting a line through a data point where the user is happy with performance - the dashed line in Fig. 5. Fig. 5 – taken from Kaminski and Ding [12] - indicates the prediction step, where the new estimate for a happy user is O2, U2. To determine the % increase in utilisation – which is not the same as the % increase in Orders (referring to the graph), the following formula can be applied:

$$(U2/U1) - 1 = growth\ percentage$$

where

$$U2 = a(O2 - O1) + U1$$

This method of performance prediction is simple to use and implement, easily providing the user with the information required, as illustrated in the examples presented by Kaminski and Ding [12].

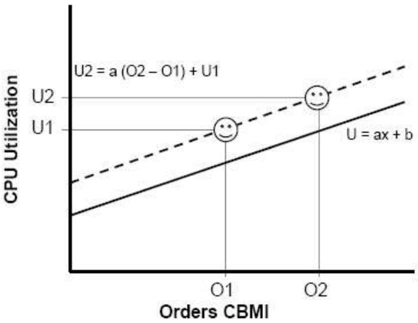


Fig. 5. Calculating workload growth .

Re-emphasising the aim of this research – a simple model, that is easy and fast to develop, not time-consuming or expensive, is required for network managers to be aware of the effects of network changes on application performance.

The AP model will follow the method discussed by Kaminski and Ding [12] closely. This method and the “guerilla capacity planning” method are similar in that both fit a regression line to the data to determine a relationship and both try to ensure simplicity in capacity planning. Granville's method requires a separate analysis for each scenario, and without network data, the results can not be effectively interpreted using visualisation tools, such as graphs or animations. Workload characterisation will be restricted to identifying what activities or tasks are performed in the application by users – making it part of Step 1 in the research methodology outline. Queuing models will not be included as part of the research model as the aim is to model network and application performance and not the performance of the specific component. It is noted that queues will have an effect on the network performance, but this will be reflected in the data measurements.

The AP model developed will be outlined in the next section.

V. Outline & Application of Research Model

The AP model will be presented as a step-by-step process or algorithm, explained by the application of each step to data of the ITS (Integrated Tertiary Software) [11] application service running on the UPE network. ITS is the student administration system used extensively by tertiary education institutions throughout South Africa. Fig. 6 represents a simplified hierarchical layout of the UPE application service network under consideration. The ITS system at UPE is accessed regularly by users completing administrative tasks. Sections I and II of this paper discuss reasons why the assurance of ITS system performance is important to the network managers; the reason for investigating the ITS application service in the research.

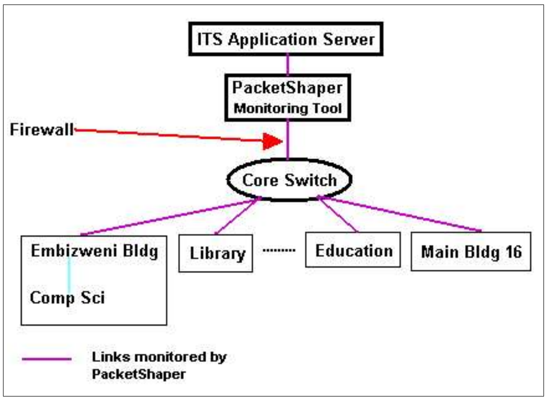


Fig. 6. Simplified hierarchical view of the UPE network.

Network and application service data is collected by Packeteer's PacketShaper monitoring tool [18] – indicated in Fig. 6. One of the reasons for selecting this tool was that it could distinguish between different applications running on the network.

The data collection used for the first analysis of the model, presented in this paper, was collected over a three week period during the university term in March 2004. At least three weeks of data collection provides suitable analysis of weekly trends as well as sufficient overall data to ensure the accuracy of any data relationships.

A. Step 1: Collect data

Key business performance metrics are identified for collection in Step 1 of the research methodology. In Section

II it was identified that network- and user-centric data need to be collected. User-centric data corresponds to application data and similar measures to the following should be collected for the application service under investigation: network, server and total delay values, round-trip-times and total transactions, as well as any other measures reflecting application performance. Network-centric data collected should include network utilisation and transaction information of the network traffic.

Application performance is compared to network performance and not business activities, as identified by Kaminski and Ding [12], as network and application data can be collected simultaneously by a data collection tool, while a time-consuming workload characterisation is required to relate business activities to application performance.

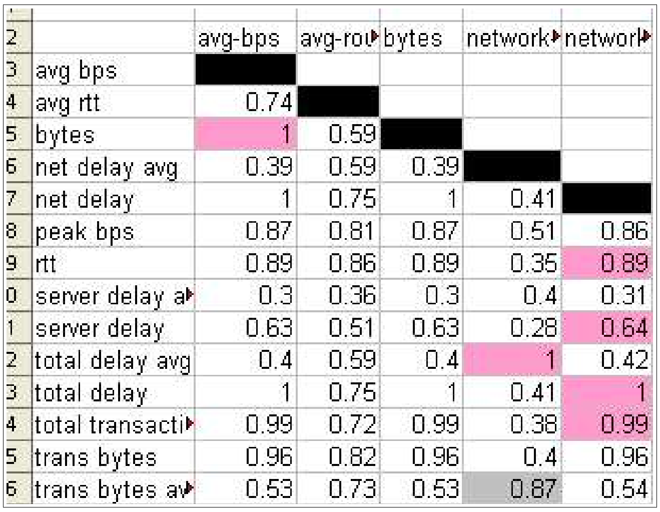
In the ITS case study using PacketShaper, the following performance data measures were identified for collection (the data was collected in hour intervals):

- Network performance data: Average bits per second (bps), bytes, peak bps, total transactions, bytes per transaction and average transmitted bytes; and
- Application performance data: Average round-trip-time (rtt), network delay average, network delay (msec), rtt (msec), server delay average, server delay (msec), total delay average and total delay (msec).

The number of kilobytes or packets could have been collected; however, due to the one-to-one relationship between these two measures and the number of bytes, it was considered not necessary.

B. Step 2: Calculate Correlation Coefficients

Correlation coefficients are important in determining whether a relationship exists between two data measures, particularly a network and application performance measure. Fig. 7 shows how Microsoft Excel can be used to calculate the correlation coefficients between all pairs of data listed above, using the built-in CORREL function. It is important to use a large enough data sample when calculating the correlation values to ensure that the correlation value of the sample is a representative value for the data as a whole.



2		avg-bps	avg-rou▸	bytes	network▸	networl▸
3	avg bps	■				
4	avg rtt	0.74	■			
5	bytes	1	0.59	■		
6	net delay avg	0.39	0.59	0.39	■	
7	net delay	1	0.75	1	0.41	■
8	peak bps	0.87	0.81	0.87	0.51	0.86
9	rtt	0.89	0.86	0.89	0.35	0.89
0	server delay a▸	0.3	0.36	0.3	0.4	0.31
1	server delay	0.63	0.51	0.63	0.28	0.64
2	total delay avg	0.4	0.59	0.4	1	0.42
3	total delay	1	0.75	1	0.41	1
4	total transacti▸	0.99	0.72	0.99	0.38	0.99
5	trans bytes	0.96	0.82	0.96	0.4	0.96
6	trans bytes av▸	0.53	0.73	0.53	0.87	0.54

Fig. 7. Calculation of correlation coefficients in Microsoft Excel.

All pairs with correlation values greater than 0.6 were identified for further analysis, from the calculations. 0.6 was chosen to ensure that no data pairs of interest in further analysis would be discarded [4]. Of particular interest, are pairs consisting of one network and one application performance measure, as this indicates a relationship between the two.

The correlation coefficients in the analysis done on the ITS data indicated strong relationships between the network performance measure: total transactions, and the application performance measures: network, server and total delay (msecs) and round-trip-time (msecs). These relationships can be verified in the next step.

C. Step 3: Data Analysis

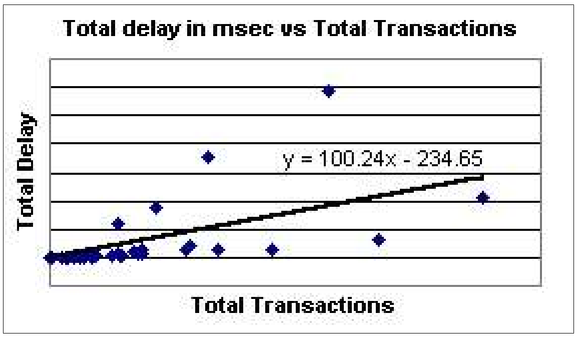


Fig. 8. Scatter plot of total delay (msec) versus the number of total transactions.

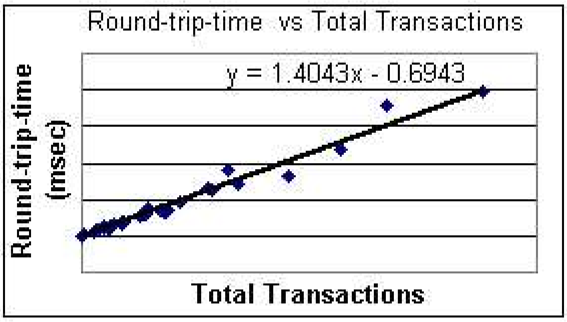


Fig. 9. Scatter plot of round-trip-time (msec) versus the number of total transactions.

All pairs of data with suitable correlation coefficients can be graphed using a scatter plot. As mentioned, pairs containing an application and network performance measure are more significant. The closer the correlation coefficient is to 1, the more accurate a linear regression line through the data points will be. This is indicated in Fig. 8 and 9, where the graphs represent correlation coefficients of 0.6 and 0.98, respectively. Microsoft Excel provides linear regression functionality – producing an equation representing the line. If the majority of the data points lie on or close to the line, the line is a fairly accurate representation of the relationship between the points, and can be used for prediction, explained in the next step.

D. Step 4: Prediction

The graph and line equation calculated in Step 3 above, can be used for prediction by applying the following steps:

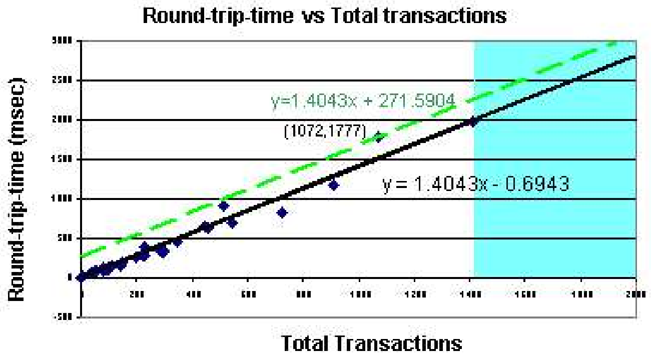


Fig. 10. Prediction technique.

1. Extend the graph further than the data points to allow prediction beyond the current network data (shaded area in Fig 10).
2. Calculate a new line (dashed line in Fig. 10) going through an outlier, with an acceptable application performance value above the line, and the same slope as the original line (solid line in Fig. 10). Calculating the new line through an outlier overestimates the predicted value slightly to provide a margin for error.
3. Input the expected value for the network measure into the equation of this new line (equation of dashed line in Fig. 10), to determine the predicted value for application performance.

Although this prediction method has not been tested and validated, it is promising. The UPE network and ITS system are currently undergoing changes – switch and ITS version 11 to 12 software upgrade, respectively – which provide an opportunity for the analysis and, especially, prediction methods to be validated. The validation process, discussed in Section III, will require data of network and application performance for a month before and a month after the change, to be collected. The analysis methods will be applied to the data before the change and the resulting model will then be used to predict what the application performance will be after the change is applied. The predicted application performance will be compared to the data collected after the change. If the predicted values are not within an acceptable margin of error to the measured values, the results will be used to adjust the model. If the error margin is acceptable, the model is validated and can be used for further prediction.

VI. Conclusion

Tools and methods for the capacity planning process exist to ensure a detailed analysis of all aspects of network planning. A methodology and model, which is easy to use, inexpensive, allows rapid analysis to be done in a common spreadsheet tool, and provides a prediction analysis technique that is easy to understand - no new skills have to be learnt for the modelling process to be carried out – is outlined in this paper to allow network managers to capacity plan a network change, with the specific focus on application performance.

The research will continue with validation of the modelling methods discussed, using the UPE network and the ITS system as a testbed. Once validated, the model will be implemented as a system tool that can be used and evaluated by network managers.

References

[1] A. Adas, "Traffic Models in Broadband Networks", Georgia Institute of Technology, *IEEE Communications Magazine*, July 1997.
[2] V.A.F. Almeida and D. Menasce, "Capacity Planning for Web Performance", Prentice Hall PTR, 2001.
[3] Y. Ding & K. Newman, "Automatic Workload Characterization", *CMG00 Proceedings*, 2000.
[4] Y. Ding, C. Thornley and K. Newman, "On Correlating Performance Metrics", BMC Software, Inc, 2001.
[5] "Capacity Planning in Today's Economy", Enterprise Management Associates Inc., Product View, 2003.
[6] B. Frogner, "Performance Monitoring, Prediction and Optimization Using Model-Based Approach", NetPredict, Inc., 2002.
[7] P. Granville, "Introduction to Network Performance: Modelling and Simulation of Systems", 1997.
[8] N. Gunther, "Hit-and-Run Tactics Enable Guerilla Capacity Planning", *IT Pro*, July /August 2002.
[9] H. Hlavacs, G. Kotsis, and C. Steinkellner, "Traffic Source Modeling", Technical Report No. TR-99101, Institute of Applied Computer Science and Information Systems, University of Vienna, 1999.
[10] "Network Performance & Capacity Planning: Techniques for an E-Business World", IMB Global Services, 1999.
[11] ITS website, http://mail.its.c.za/newsletter/start.php
[12] R. Kaminski & Y. Ding, "Business Metrics and Capacity Planning", *CMG03 Proceedings*, 2003.
[13] P. Korzeniowski, "The Emerging Frontier: Application Performance Measurement", Response Networks Whitepaper, April 2002.
[14] W.E. Leland, W. Willenger, M.S Taqqu and D.V. Wilson, "On the Self-Similar Nature of Ethernet Traffic", *ACM SIGCOMM '93*, Computer Communication Review, Pg. 203-213, Sept. 1993.
[15] G. Mattahil, "A Reference Model for Network Evolution", Strategic Advisory Group, White Paper, June 2003.
[16] "Network Capacity Planning Service", NCR, Presentation created 27 September 2000.
[17] "Deploying and Optimizing Networked Applications", OPNET Technologies White paper, 2003.
[18] Packeteer website, http://www.packetwise.com.
[19] O. Zaidi and R. Blum, "Network Capacity Planning: Applying Best-in-Class Methods, NetKnowledge Webinar hosted by International Network Services (INS), 10 Decmber 2003.

Melisa Koorsse received her BSc and BSc Hons degrees in Computer Science and Applied Mathematics from the University of Port Elizabeth, in 2001 and 2002, respectively. She is currently doing her MSc in Computer Science and Information Technology at the University of Port Elizabeth, researching the modelling of application performance on networks.