# ANALYTICAL MODELING OF A LARGE LOCAL AREA NETWORK - PART I: INTERNET TRAFFIC CHARACTERIZATION

**_Mathew L. Luhanga and Cuthbert E. B. M Amos_**
*Department of Telecommunications Engineering, University of Dar es Salaam,*
*P.O Box 35131, Dar es Salaam, Tanzania*
*Emails: luhanga@udsm.ac.tz, amomsuya@udsm.ac.tz*

## Abstract

*The goal of both IP network operators and the end users is to get the highest performance from the system for a given cost. This makes Performance a key criterion in the design, procurement, and use of computer and communication systems.*

*In order to address problems associated with performance degradation of operational communications systems, over the last decade, traffic engineering techniques have emerged in an attempt to optimize communication systems' performance and ensure more efficient use of their resources. One of these techniques is analytical modeling.*

*Analytic performance models are an excellent tool for quickly evaluating the performance of operational or new systems. They are also well suited to comparing the performance of several alternative designs. However, analytical models can only be developed once detailed knowledge of characteristics of traffic carried by a network is available. In Part I of this paper, traffic characterization of traffic carried by the largest Local Area Network (LAN) in Tanzania, University of Dar es Salaam Network (UDSMNET) is carried out. In Part II of this paper, an analytical model based on the Discrete Time Markov Modulated Poisson Process is proposed and validated for performance analysis of IP networks.*

***Key words:*** *Traffic characterization, Traffic Engineering, Long Range Dependence*

## INTRODUCTION

The engineering of IP networks is facing a number of difficult challenges, User behaviour access to multimedia content on the web and deployment of new applications result in significant fluctuations of traffic in IP networks. This leaves network operators in difficult situation of trying to tune the configuration of their networks to adapt to changes in traffic characteristics. Consequently the performance optimization of IP networks, especially public internet backbones, has become important (Awduche *et al.,* 2002; Feldmann, *et al.,* 2001).

UDSMNET is the largest LAN in Tanzania. In 2005, it was comprised of 3,500 network hosts serving a community of about 12,000 staff and students. Figure 1 shows the architecture of the UDSMNET.

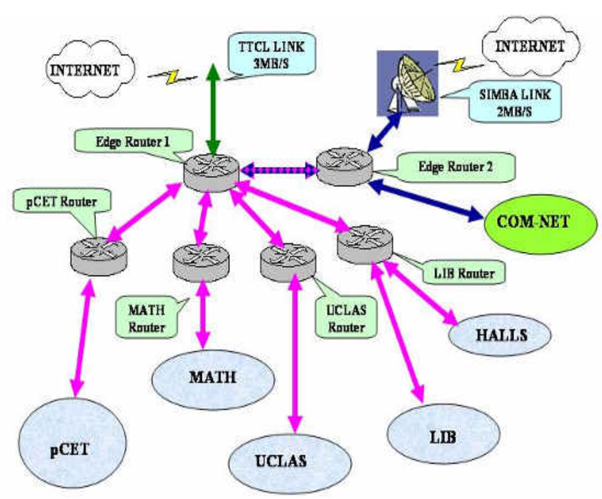


**Figure 1. The UDSM Network Backbone**

Since the inception of UDSMNET, no study has been done on the network to evaluate and optimize its performance. Results reported here were obtained from the first detailed study on UDSMNET that was aimed at developing tools that would help in performance evaluation, optimization and network planning (Amos, 2006).

## TRAFFIC MODELLING

### The Rationale

The performance of any IP network is dependent on the architecture of the network and on the characteristics of the traffic carried over the network. Traffic engineering is thus an important input into the performance evaluation of IP networks. This study was concerned with the development of an analytical model of the UDSMNET. To achieve this, measured traffic data was used to identify and characterize the traffic on the UDSMNET. Part I of this paper deals with the characterization of traffic carried on UDSMNET.

### Traffic Characteristics on UDSMNET

The UDSMNET was used as a test-bed network for all traffic measurements. Details on the physical structure, services offered and traffic types on UDSMNET are available elsewhere (Amos, 2006).

### Traffic Characterization

Traffic traces on UDSMNET were measured and analysed. Since UDSMNET is a client network (Amos, 2006), only traffic traces collected at the egress side of the UDSM-NET edge router1 connecting the UDSM-NET and the TTCL access router are considered.

The traffic measurement setup is shown in Figure 2 and an example of some of the traces obtained is shown in Figure 3. Figure 3 is a daily graph of five minutes average values of the observed traffic.

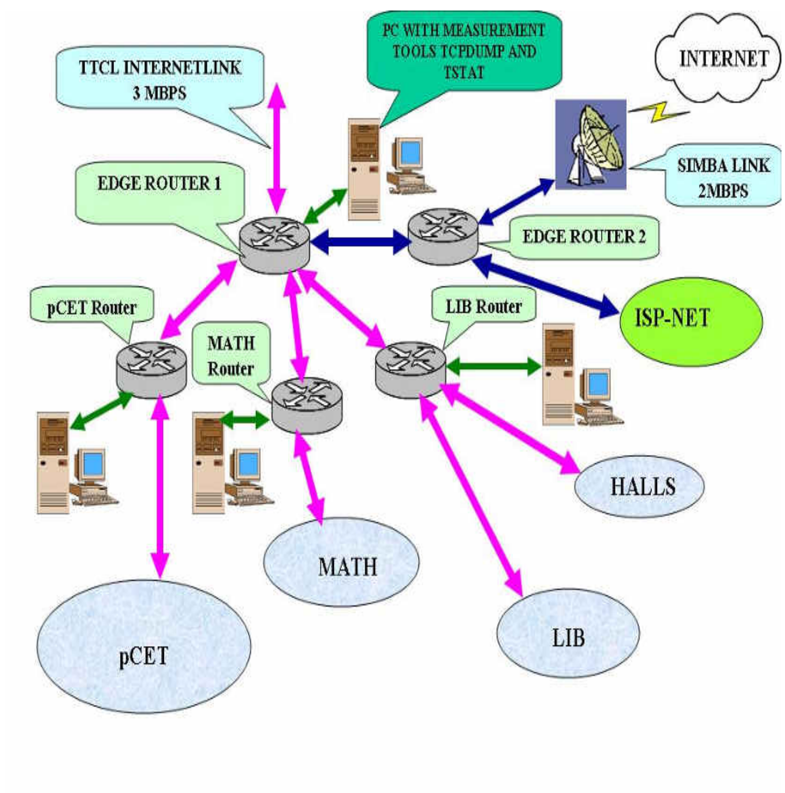


**Figure 2. Traffic Measurement Setup**

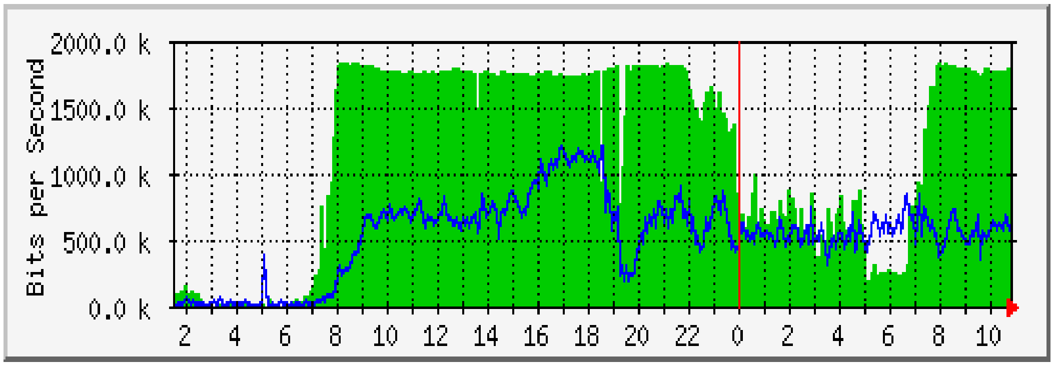


**Figure 3. Measured Traffic Traces. Note: the horizontal axis represents time the time of the day in 24 hours format**

Among all possible metrics that can be derived from the traffic traces, the packet arrival time process and the packet counting process were selected as being most representative of the trace characteristics as far as modeling was concerned. The packet counting process was obtained by counting the number of packet arrivals in an interval of T sec. For this work values of T=0.1 sec and T=1.0 sec were used. The choice of T is made such that the loss in detail due to sampling interval, T, is compensated by analyzing a much longer trace (Amos, 2006). Traffic traces were obtained at an edge router of UDSMNET thus capturing all traffic between the UDSMNET and the Internet (Amos, 2006).

A typical example of characteristics of the measured traces is shown in Figure 4 and Table 1. General observations were that packets due to the TCP protocol contribute over 50 per cent of the measured traffic. In addition, application layer protocols that work over TCP (e.g. HTTP, SSH and SMTP) account for above 20 per cent of the traffic. Detailed characteristics for traffic traces obtained on different dates are available elsewhere (Amos, 2006).

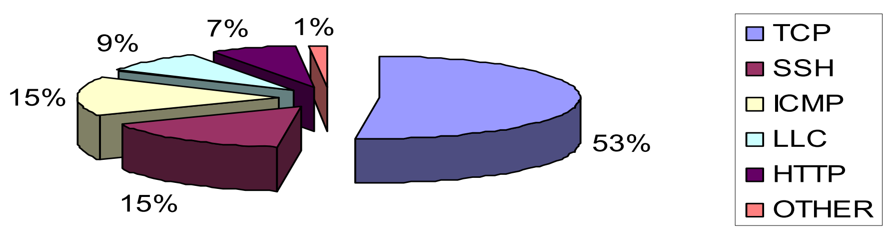


**Figure 4. Protocols of the UDSM-22/23 FEB05.DAT Trace**

**Table 1.** Statistics of the UDSMNET DATA measured on 22/23 FEB 20005

| | |
|---|---|
| Trace Name | UDSM-22/23 FEB05.DAT |
| Capture Date | 22nd Feb 2005 to 23rd Feb 2005 |
| Total Capture Time | 84891.170 seconds |
| Total number of packets | 1,040,236 packets |
| Average Rate | 12.254 packets/sec |
| Average Packet Size | 340.371 bytes |
| Average bytes/sec | 4170 |
| Average Mbits/sec | 0.033 |
| Bytes of traffic | 354066205 bytes |

An appropriate analytical model can only be chosen once the traffic characteristics are known. To meet this goal, the traffic traces were analysed to determine the presence of Long Range Dependence (LRD) (Crovella, *et al.,* 1997) an important parameter in Internet traffic analysis. A wavelet based Hurst parameter estimator known as an AV ESTIMATOR (Abry, *et al.,* 1999) was used to study the LRD of the measured traces. Wavelet analysis explores the scaling phenomenon of the signal and produces a time-scaled view of the signal under study (Abry, *et al.,* 2002). Because of this fact, we refer to the analysis using this tool as "Scaling Analysis". In this study, the counting process was used as the input to the analysis tools used for studying the presence of LRD behaviour in the measured traffic traces.

Software tools, TCPDUMP (Tcpdump.org, 2005) and AWK (Aho, *et al.,* 1998) were used to extract from trace files the packet arrival process and create text files that could be analysed using MATLAB (Mathworks Inc., 2005). This was achieved by reading each trace file using TCPDUMP with a filter made from AWK code such that only the arrival time process was displayed and then saved as a text file. AWK is a general purpose programming language that is designed for processing text-based data, either in files or data streams (Aho, *et al.,* 1998). The name AWK is derived from the surnames of its authors, Alfred Aho, Peter Weinberger, and Brian Kernighan (Aho, *et al.,* 1998).

A file converting tool, named FILE CONVERTOR was developed using MATLAB code and the tool was used to convert the packet arrival time process text files into corresponding counting process text files (Amos, 2006).

**Table 2** shows a sample trace file captured by TCPDUMP. The columns in Table2 represent the following:

| COLUMN NO: | CONTENT |
|---|---|
| 1 | Packet number since counting process began |
| 2 | Packet arrival instant in seconds since beginning of counting process |
| 3 | Source IP address |
| 4 | Destination IP address |
| 5 | Protocol associated with packet |

Note that TCPDUMP can also record the interarrival time process.

**Table 2.** A Sample of TCPDUMP Capture File Showing the First Five Columns

| No | Time | Source | Destination | Protocol |
|---|---|---|---|---|
| 1 | 0.000000 | 196.44.164.3 | 10.101.193.67 | TCP |
| 2 | 0.000068 | 196.44.164.3 | 10.101.80.99 | TCP |
| 3 | 0.000144 | 196.44.164.3 | 10.101.201.139 | TCP |
| 4 | 0.000219 | 196.44.164.3 | 10.101.61.81 | TCP |
| 5 | 0.000293 | 196.44.164.3 | 10.101.169.193 | TCP |
| 6 | 0.000368 | 196.44.164.3 | 10.107.178.209 | TCP |
| 7 | 0.000860 | 64.86.50.210 | 10.45.248.250 | TCP |
| 8 | 0.001246 | 64.86.50.210 | 10.45.145.46 | TCP |
| 9 | 0.001982 | 63.218.13.200 | 196.44.164.128 | HTTP |
| 10 | 0.005003 | 196.44.164.128 | 65.54.195.188 | HTTP |
| 11 | 0.007756 | 206.65.183.68 | 196.44.164.128 | HTTP |
| 12 | 0.008042 | 196.44.167.30 | 62.72.65.76 | TCP |
| 13 | 0.009980 | 196.44.162.188 | 205.188.156.249 | TCP |
| 14 | 0.010109 | 196.44.162.188 | 209.198.87.161 | TCP |
| 15 | 0.011432 | 196.44.167.123 | 216.15.227.229 | TCP |

**Table 3a** shows a sample of the packet arrival process in seconds while Table 3b shows the counting process for T=0.1

**Table 3a.** The Arrival Process for the UDSMNET DATA measured on 5th March 2005

| Arrival Time Process in Seconds |
|---|
| 0.000000 |
| 0.000068 |
| 0.000144 |
| 0.000219 |
| 0.000293 |
| 0.000368 |
| 0.00086 |
| 0.001246 |
| 0.001982 |
| 0.005003 |
| 0.007756 |
| 0.008042 |
| 0.00998 |
| 0.010109 |
| 0.011432 |
| 0.013292 |
| 0.013541 |
| 0.013796 |
| 0.015333 |
| 0.015376 |

**Table 3b.** The Counting Process for the UDSMNET DATA measured on 5th March 2005

| The Counting Process for T=0.1 |
|---|
| 132 |
| 124 |
| 166 |
| 201 |
| 164 |
| 288 |
| 147 |
| 164 |
| 287 |
| 267 |
| 293 |
| 115 |
| 151 |
| 244 |
| 186 |
| 288 |
| 228 |
| 161 |
| 248 |
| 137 |

**Hurst Parameter Estimation From Measured Traces**

The LRD phenomenon can be characterized by the Hurst parameter (Park *et al.,* 2000). In this section the presence of LRD in the measured traffic is analyzed by estimating the Hurst parameter from the measured traces. We use the semi-parametric estimation tool referred to as the "AV ESTIMATOR" (Abry *et al.*, 1999). The choice of AV ESTIMATOR is justified elsewhere (Hae-Duck Joshua, 2002).

**The AV Estimator**

Let $r(k)$ be the autocorrelation function of the packet counting process $X(t)$ and $\Gamma(v)$ its spectral density, not a function in the frequency domain as in Fourier series, but a function in a time-scale wavelet domain. In this representation the sinusoidal functions of Fourier series are replaced by wavelet basis functions (Roughan *et al.,* 2000).

It can be shown that $\Gamma(v) \sim c|v|^{-\alpha}$, $v \to 0$, where $c$ is a constant and the parameter $\alpha$ (the scaling exponent) is related to the Hurst parameter $H$ by: $H = (\alpha+1)/2$ (Abry *et al.,* 1999).

Wavelet analysis explores the scaling phenomena of a signal and produces a time scale view of a signal under study (Abry *et al.,* 2002).

The estimator (implemented in MATLAB) takes as its input the counting process of a trace generates a Logscale Diagram (LD) and a goodness of fit curve. To establish the presence of LRD, one looks for alignment in the LD, which is a log-log plot of the variance estimates of the discrete wavelet transform coefficients representing the traffic process, against scale, complete with confidence intervals about these estimates at each scale. The scale is represented by $j$ and the logarithm of the variance estimate by $y_j$. Traffic is said to be LRD if, within the limits of the confidence intervals, the log of the variance estimates fall on a straight line, in a range of scales from some initial value $j_1$ up to the largest one present in data. The slope of the straight line is an estimate of the scaling exponent, $\alpha$, which lies in (0, 1) (Abry *et al.,* 1999).

The range of scales $(j_1, j_2)$ over which a scaling phenomenon exists varies. For the LRD process, the upper count scale $j_2 = \infty$, but the lower count scale, $j_1$, where the LRD "begins", must be chosen (Abry *et al.,* 1999). In this case, prior to estimation an analysis phase is necessary to determine the lower count scale at which the LRD "begins", and to see if LRD is present at all. The region over which a scaling phenomenon exists is known as the "***Scaling region***" (Abry *et al.,* 1999).

In the analysis phase, the generated Logscale Diagram is examined to find a lower cutoff scale $j_1$, and upper cutoff $j_2$, where alignment (a straight line) is observed. These cutoffs should be experimented with to find a range where the regression fits the confidence intervals plotted on the Logscale Diagram well. Initial values must be given in the argument list to the AV ESTIMATOR, but these can then be changed interactively. In addition, a goodness of fit statistic 'Q' is outputted to help with the choice of scaling range, and is plotted in the title of the Logscale Diagram. Q is the probability of observing the data given that the expectations of the variance estimates at each scale really do follow the defining linear form of linear regression. A Q value greater than 0.05 are acceptable. However, depending on the data under consideration it may happen that the value of Q is always less than 0.05. In such cases, Q values can only be used in a relative way to compare different choices of the scaling range (Abry *et al.,* 1999).

Experimentation with the number, N, of vanishing moments of the wavelet is also needed in the analysis phase in order to ensure that the wavelet details are well defined. The value N=1 is sufficient for LRD processes (Abry *et al.,* 1999) and therefore it was used in this study.

The goodness of fit curve which is a graph of the goodness of fit measure $Q(j_1)$ against the lower cutoff scale $j_1$, enables the user to quickly determine the beginning of the scaling region (Abry *et al,* 1999). Taking Figure 5 as an example, it is observed that $Q(j_1)$ as a function of $j_1$ improves dramatically at $j_1=5$. This is thereby recommended as the beginning of the scaling region (Abry *et al.,* 1999). Figure5 was obtained after carrying out scaling analysis on of the UDSMNET traces.

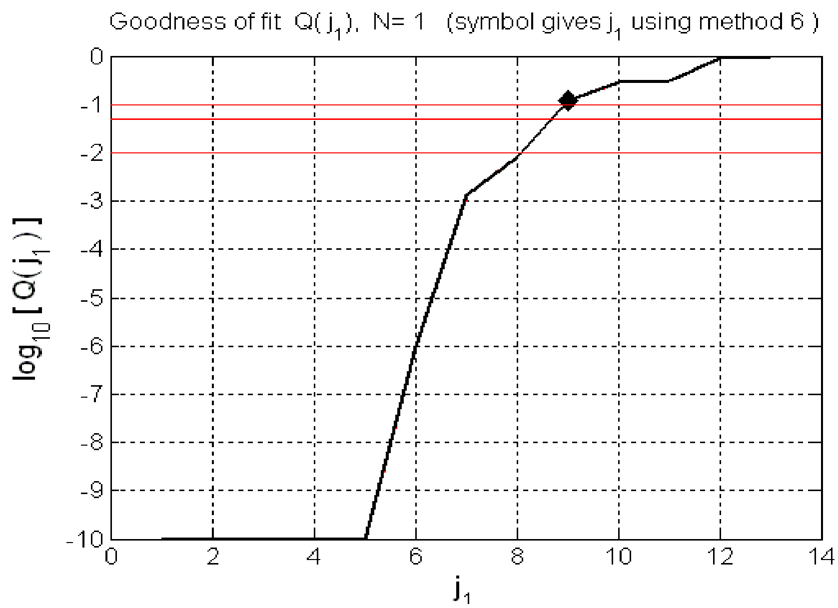


**Figure 5. Goodness of fit curve**

**Scaling Analysis**

An example of the result of the scaling analysis carried out on one of the traces of UDSMNET is shown in Figure 6. Figure 6 shows that the trace exhibits the LRD behaviour and that the scaling region begins at $j_1$ as previously determined.

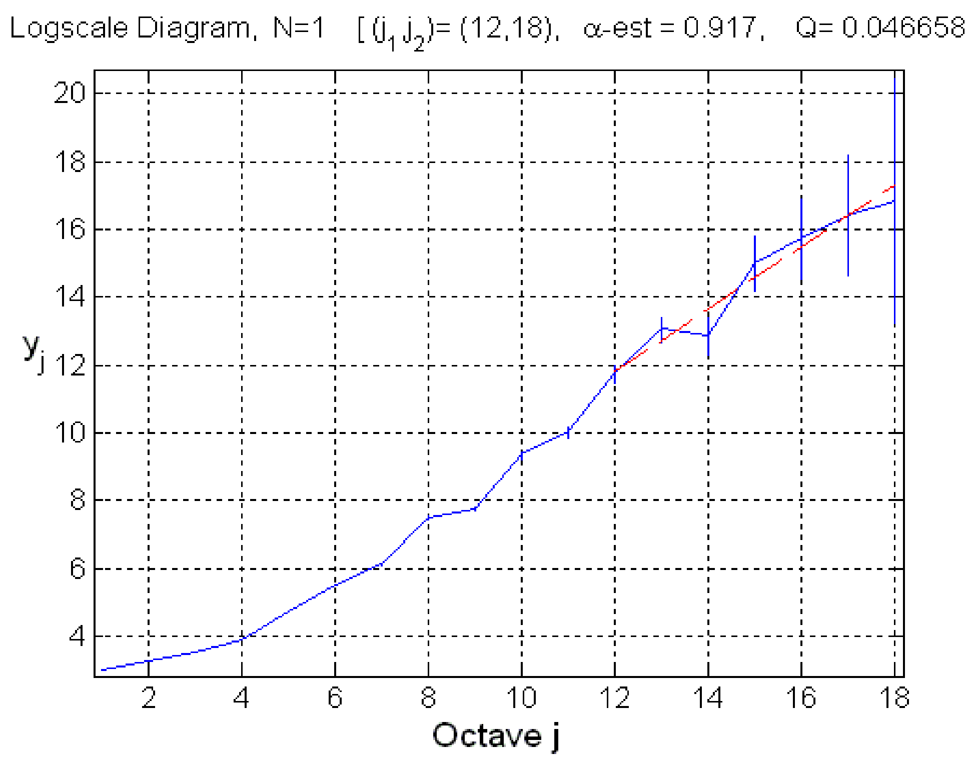


**Figure 6. Scaling Analysis. UDSMNET DATA TRACE measured on the 22nd and 23rd February 2005**

## CONCLUSION

The analysis of measured traces of traffic carried by UDSMNET has revealed that the traffic exhibit Long Range Dependence characteristics. This opened a way to using existing analytical models to get a mathematical model of the traffic. Development of the mathematical model is reported in part II of this paper.

## NOMENCLATURE

| | |
|---|---|
| $c$ | Wavelet scaling parameter often referred as the LRD size |
| $j_1$ | The lower cut off scale of the AV Estimator where LRD begins |
| $j_2$ | The upper cut of scale of the AV Estimator where LRD ends. |
| $(j_1,j_2)$ | The range of scales over which a scaling phenomenon (LRD) exists when using the AV Estimator |
| N | The number of vanishing moments of the AV Estimator |
| Q | Goodness of fit statistic of the AV Estimator |
| Q $(j_1)$ | The goodness of fit measure of the AV Estimator |
| $r(k)$ | The autocorrelation function of the packet counting process |
| $X(t)$ | The packet counting process (Note: this is a time series) |
| $y_j$ | The Logarithm of the Variance Estimate when using the AV Estimator |
| $\alpha - est$ | Estimate of the dimensionless scaling exponent $\alpha$ |
| $\alpha$ | Dimensionless scaling parameter that describes the intensity of LRD |
| $\Gamma(v)$ | Spectral density of the packet counting process |

## LIST OF ABREVIATIONS AND ACRONYMS

| | |
|---|---|
| **IP** | Internet Protocol |
| **LAN** | Local Area Network |
| **LRD** | Long Range Dependence |
| **TCP** | Transmission Control Protocol |
| **UDSMNET** | University of Dar es Salaam Network |
| **HTTP** | Hyper Text Transfer Protocol |
| **SSH** | Secure Shell |
| **SMTP** | Simple Mail Transfer Protocol |

## REFERENCES

Abry, P., Flandrin, P., Taqqu, M.S., and Veitch, D., (2002) *Self-Similarity and Long-Range Dependence through the Wavelet Lens.* In Doukhan, et al., editors, *Theory and Application of Long-Range Dependence*, Birkhuser, Boston, USA, ISBN 0-8176-4168-9.

Abry, P., and Veitch, D., (1999) *A Wavelet Based Joint Estimator for the Parameters of LRD,* IEEE Transactions on Information Theory, Vol.45 No.3.

Aho, A. V., Kernighan, B.W., and Weinberger, P. J., (1988)*The AWK Programming Language*. Addison-Wesley, ISBN 0-201-07981-X.

Amos, Cuthbert E.B.M., (2006) *Performance Analysis of IP- Based Networks:A Case Study of the University of Dar Es Salaam Local Area Network.* Unpublished M.Sc. Thesis. Computer and Systems Engineering Department, Faculty of Electrical and Computer Systems Engineering, University of Dar es Salaam.

Awduche, D., Chiu, A., Elwalid, A., Widjaja, I. and Xiao, X., (2002) *Overview and Principles of Traffic Engineering*, RFC 3272., pp3-21.

Crovella, M. E and Bestavros, A., (1997) *Self-Similarity in World Wide Web Traffic: Evidence and Possible Causes,* IEEE/ACM Transaction on Networking, Vol.5, N.6, pp.835-846.

Feldmann, A., Greenberg, A., Lund, C., Reingold, N., Rexford, J., and True, F., (2001) *Deriving Traffic Demands for Operational IP Networks: Methodology* *and Experience*. IEEE/ACM Transactions on Networking, pp. 265-279.

Hae-Duck Joshua Jeong, (2002) *Modeling of Self-Similar Teletraffic for Simulation.* Unpublished PhD Thesis in Computer Science. University of Canterbury, New Zealand.

MathWorks, Inc. (2005), *Matlab Manual.* URL: [http://www.mathworks.com/access/helpdesk/help/techdoc/matlab.shtml/.], 2006.

Park, Kihong., and Willinger Walter.,(2000) *Self-Similar Network Traffic and Performance Evaluation*. John Willey & Sons, Inc. New York.

Roughan, M., Veitch, D and Abry, P., (2000) *Real-Time Estimation of the Parameters of Long Range Dependence,* IEEE/ACM Transactions on Networking, Vol. 8, No 4.

Tcpdump.org, (2005), *Tcpdump Synopsis.* URL: [http://www.tcpdump.org/tcpdump_man.html/.], 2006.