Self-Similarity and Internet Performance

Devarajan Gopal
Professor, Department of Digital Information & Communication
Woosong University, Daejeon, Korea 300-718
Email: gdrajan33@yahoo.com

Taye Abdulkadir
Department of Electrical & Computer Engineering
Faculty of Technology, AddisAbaba University, Ethiopia
Email: tabdulkadir@yahoo.com

In this work, it has been demonstrated through experiments that Long Range Dependence (LRD) in internet traffic is indeed affecting the queueing performance. Queueing performance is being degraded as the self-similarity of the packet arrival process increases. Here, the link buffer capacity and queueing delay exhibits super-linear relationship. This in turn has led to the proposal of resource provisioning strategy to boost the queueing performance. The inherent self-similar nature of the internet made the possibility of carrying voice traffic accompanied by voice quality degradation. Self-similar or variable nature of packet delay process implies voice quality degradation due to inter packet arrival variations. Delay and delay variation (jitter) behave erratically due to the self-similar nature of the packet arrival and delays. As the self-similarity of the packet round trip delay increases, TCP throughput performance is degraded.

Keywords: Long Range Dependence (LRD), Quality of service (QoS), Voice over Internet Protocol (VoIP), Round Trip Time (RTT).

ACM Classification: C.4 (Performance of Systems).

1. INTRODUCTION

Studies on Internet Traffic data (Taye, 2003) confirm the presence of scale invariance in different time scales. So the performance of Network should be dominated by this property. Packet Round Trip Delay process when displayed as a time series also proved to have the scale invariance bursty nature or self-similarity (Taye *et al*, 2002; Borrela *et al*,1997). This finding is essential in building proper Quality of Service (QoS) provisioning for the next generation of delay sensitive real-time multimedia traffic to be carried over the Internet.

Once the existence of self-similarity is asserted in both the packet traffic and delay processes, we focus on the influence of LRD in traffic and delay processes on the performance of Networking, which is approached from *three* possible directions. One is investigation of the effect of Traffic Self-similarity on Queueing performance. As a second approach, performance of Real-Time traffic, in this case Voice over Internet Protocol (VoIP), investigated in relation with the nature of Packet Delays. The third is investigation on the effect of Packet Round Trip Delay Self-Similarity on the performance of Internet transport mechanisms. Here, Queuing Performance with LRD packet traffic

Manuscript received: 17 November 2004
Communicating Editor: Zoran Salcic

is seriously degraded. Numerical and analytical studies based on models catching LRD property of Traffic (Borrela and Brewster, 1997) demonstrate that the tail of a queue length distribution decays much more slowly than the exponential rate; this implies that a packet tends to experience longer delay with in networks on average before it gets to its destination than that predicated by traditional models; such as the Markov Model. The conclusions drawn from these studies have remarkable implications for the design, control of networks, and tuning of protocols. Real time or Constant Bit Rate (CBR) packets travel through the Internet is most affected by the packet delays encountered in the Internet or by the overall performance.

For VoIP to become popular and in the long run to replace Public Switched Telephone Network (PSTN), Quality of Service (QoS) issues need to be resolved. As Internet Protocol (IP) was designed for carrying data, it does not provide real time guarantees but only provides best effort services. VoIP will be successful if one gets best understanding of Packet Delays in the IP network. So for voice communication over IP to become acceptable to users, the delay needs to be minimized to an acceptable value. Investigation on the relation between Internet packet delay nature and VoIP requirements are also included in the paper. Commonly used transport mechanisms is Transmission Control Protocol (TCP), included in TCP/IP, which is the protocol suite used in the Internet. With TCP there are several time out parameters needed to be carefully tuned, the most sensitive one is the Retransmission Time Out (RTO) parameter determined dynamically from packet round trip delay in Internet. The RTO is used to indicate when a packet can be assumed lost in the Network and consequently invoke a retransmission event. Failure in tuning of RTO will result in wastage of bandwidth due to premature time outs, which result in packet re-transmission. With the high variability nature Packet Round Trip Delay process, the Retransmission Time Out (RTO) parameter will be highly affected and bandwidth wastage due to unnecessary re-transmissions is expected.

The paper is arranged as follows: Section 2 covers the work on the effect of Packet arrival self-similarity on Queueing Performance. Section 3 presents concepts on VoIP and how it is affected by the self-similarity of Packet Delays. Section 4 presents investigation on the effect of self-similarity on TCP throughput performance. The last section concludes with the inference drawn.

2. QUEUEING PERFORMANCE ANALYSIS

The dominant component of Packet delay is Queueing delay, which is stochastic in nature. In Internet a packet route generally includes a tandem of many queues, so the packet round-trip delay process is mainly determined by the Queueing performances along the route. Therefore, studying the effect of Self-Similarity of arrival traffic on Queueing performance is studying its influence on Packet Round Trip Delays.

2.1 Experimentation

In this section the experimentation performed to demonstrate the effect of Traffic Self-Similarity on Queueing performance is discussed. We have used a Spreadsheet (MS-Excel based) Queueing simulator, adapted from a freeware Spreadsheet Queueing simulator engine, © A. Ingolfsson and T. A. Grossman. The simulator is modified to accept Interarrival time as an input series, instead of taking it from a fixed probability distribution. The service rate of the Queue is taken from a fixed probability distribution.

For this experimentation, two different data series were used. One is taken from the Average Interarrival time series of the actual collected Internet Traffic data (Taye, 2003). It is calculated from the collected traffic time series by dividing the time resolution (*10 ms*) by the amount of packets, for the entire sequence. That is

$$Interarrival\ Time = \frac{10}{Number\ of\ packets}$$

Then it is normalized to fit the need of our simulator. Since the packet Interarrival is taken from the packet arrival process, it is expected to have long-memory. Indeed, the Hurst parameter estimated for the Interarrival time series has a value of 0.7231.The second data set is taken from a typical Short-Range dependent time series, generated from an Auto Regressive Moving Average (ARMA (4,4)) traffic generator, written in Matlab programming language. The estimated Hurst parameter (H) for the series is 0.3.

2.2 Observation

The simulator runs by changing the interarrival time series from the two data sets *namely* Long Range and Short Range dependant time series. The resulting average Queueing delay for the Self-Similar Inter arrival time sequence is found significantly larger (Value=0.10894) than the resulting average Queueing delay of the Short-Range dependent Interarrival time (Value=0.083). Therefore, it can be concluded that Queueing performance is affected due to the Self-Similar or Long Memory nature of the packet arrival process, that is, Queueing performance degraded as the self-similarity of the Packet arrival process increases.

A more exhaustive experimentation (resolving the limitations of Trace amounts and real-time nature) can be done by using Real Network environment or real-time Network Simulators such as the LBNL-NS[1] as a simulation environment. Park *et al* (1997; 2000) studied the effect of traffic self-similarity on different aspects of Network performance, such as throughput, packet loss, and Queueing delay, using LBNL-NS. We have used the tail index or shape parameter α (H=(3-α)/2)(α is near 1 i.e. H goes near 1) the traffic gets more self-similar) of the heavy-tailed distribution (Pareto Distribution) (Taye, 2003) to control the self-similarity. We did run tests for different link buffer sizes (3KB, 10KB, 28KB, and 64KB) and the experimental results obtained is shown in Figure 1

[1] Laurence Berkeley National Laboratory – Network Simulator – the most commonly used Network Simulator

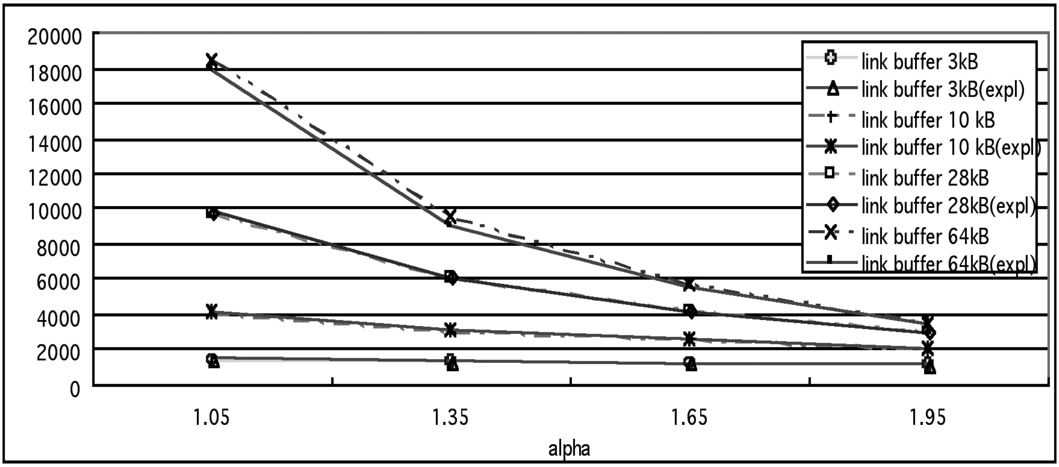

Figure 1: Mean Queue length as a function of α

and compared with that of the results of Park. It may be observed from this figure, the mean Queueing delay increases nonlinearly as the traffic gets more self-similar. Our experimental results are shown in solid line (marked expl). Further, it may be noted here the super-linear relationships between link buffer capacity and Queueing delay. This has led to proposals advocating a resource provisioning strategy due to link buffer's simplistic, yet curtailing influence on Queueing: if buffer capacity is small, the ability to queue or remember is accordingly diminished. This resulted in a different QoS strategy for LRD traffic in such a way that keeping buffer capacity small and increasing bandwidth to boost Queueing performance (Park and Willinger, 2000).

3. VoIP PERFORMANCE ANALYSIS

3.1 Overview of Voice-over-IP

In a Voice-over-IP system, the analog voice signal is typically picked up by a headset-mounted microphone and sent to an audio processor within a PC. The nominal bandwidth required for telephone-type voice ranges from 2.9 Kbps to 13 Kbps (GSM cellular standard) (Davidson and Peters, 2000; Garret and Willinger, 1994). The H.323 standard provides a foundation for audio, video, and data communications across IP-based networks, including the Internet. By complying with H.323, multimedia products and applications from multiple vendors can interoperate, allowing users to communicate without concern for compatibility.

In placing the codec output into packets, there is a trade-off between bandwidth and latency (Leland *et al*, 1994). Codecs do not operate continuously. Instead, they sample the voice over a short period of time, known as a frame. These frames are like little bursts of data. One or more frames can be placed in a single IP datagram or packet, and then the packet payload is wrapped in the necessary packet headers and trailers. Waiting longer to fill the IP datagram reduces overhead, which in turn reduces the true bandwidth needed to send the digitized voice. However, this waiting creates latency at the source, and too much total latency makes for a difficult conversation. The total network latency and jitter (changes in the latency) have a degrading effect upon voice quality. Therefore, real-time voice quality is difficult to maintain over a large wide-area packet network without priority handling. If latency and jitter are too high, or the cost of reducing them is excessive, one alternative is to buffer the codec data at the receiver. A large buffer can be filled irregularly but emptied at a uniform rate. This permits good quality reproduction of voice. Such a buffering technique is known as audio streaming, and it is a very practical approach for recorded voice or audio.

The Internet has become a ubiquitous infrastructure, used by numerous applications having various requirements and that generate traffic that has different characteristics. Voice requirements are stringent, toll-quality real-time communication is needed, which limits the maximal tolerable round-trip delay to 200-300 ms; that is, one-way delay must be in the range of 100-150 ms for adequate performance. Delay and jitter are the important measures of QoS particularly for the voice traffic in the network environment with bursty background traffic.

3.2 Delay

Internet traffic is proved to have bursty nature which is evident from the high variability and correlations spanning large time scales, which is tackled through Self-similarity and Long-range dependence (Garret and Willinger, 1994; Leland *et al*, 1994). The delay process associated with the traffic is also proved to have the Self-similar nature. The high variability encountered both in traffic and delay processes affects the performance of the applications carried over the Internet. Further queueing performance degrades as the self-similarity of the packet arrival process increases. This implies that packet traveling with a self-similar traffic will encounter more delay than the case when the traffic is non self-similar or short-memory. The delay encountered as a packet carried on the Internet has more impact on real-time traffic, such as VoIP. To achieve a reliable real-time communication, issues related to the very nature of the Internet have to be resolved. That way the migration of PSTN to VoIP can be achieved.

For a given voice connection, the only random component of voice delay is the Queueing delay in the network. The increase in self-similarity of the network traffic arrival implies Queueing performance degradation. VoIP carries voice over the Internet, whose traffic and delay processes are self-similar. Therefore, the queueing degradation is implied for VoIP packets, if they are treated in unison with the other Internet packets. The average queueing delay for a given buffer utilization is inversely proportional to the available bandwidth, that is, delay percentiles decrease with an increase in available bandwidth, that resulted in a different QoS strategy for LRD traffic in such a way that keeping buffer capacity small and increasing bandwidth to boost Queueing performance. In terms of traffic characterstics, voice streams have low data rates (in the order of tens of killo bits) and exhibit low burstiness. The approach of QoS strategy for LRD traffic, advocating the use of large bandwidth, is therefore not suitable for managing voice traffic "specific" QoS mechanisms, but suitable for the performance of the whole Internet. Because of its stringent requirement of delay and particular characterstics, voice traffic should be treated differently than other traffic in the Internet. This can be achieved by having voice traffic through a separate queue. If other traffic is flowing on the link, then voice traffic could be serviced using Priority Queueing (PQ), and given the highest priority over all other traffic. At the Internet scale, high priority can be provided to voice traffic in the Differentiated Services framework.

3.3 Jitter

Variable packet delays or jitters in a single voice stream will result in voice degradation. These delay variations are results of the Queueing delays encountered with the different packets that made up the voice stream. Therefore, if the variations in one-way delay are properly resolved, then the jitter problem will be resolved. Other wise, QoS mechanisms should be implemented to counteract the delay variations. The usual implementation to resolve voice degradation due to jitter at the receiver side is to use jitter buffer. This is achieved by holding arriving packets until a later playout time in order to ensure that there are enough packets buffered to be played out continuously.

3.4 Experimentation

Following is experimentation, on how packets (data, voice or video) encounter variable delays as they traverse the routes of the Internet, assuming without any QoS measures. The delay variation or burstiness exists even if the packets follow the same route through out the stream duration, due to the random nature of Queueing delay encountered along the path.

Researches (Taye *et al*, 2002; Garret and Willinger, 1994) asserted the high variability or bursty nature of packet round trip delays. One-way delay, which determines the nature of jitter, is also to have such nature. To have an insight on the nature of the one-way delay, we have used the Traceroute application of the TCP/IP protocol stack. Traceroute lets us see the route that IP datagrams follow from one host to another and estimate the time spent while the packets traverse each hop along the path. There is no need of time synchronization, since the sender machine manages both timings. Although there are no guarantees that two consecutive IP datagrams from the same source to the same destination follow the same route, most of the time they do. Logging one-way delay values for some time and treat the trace as a time series shows how bursty the delay

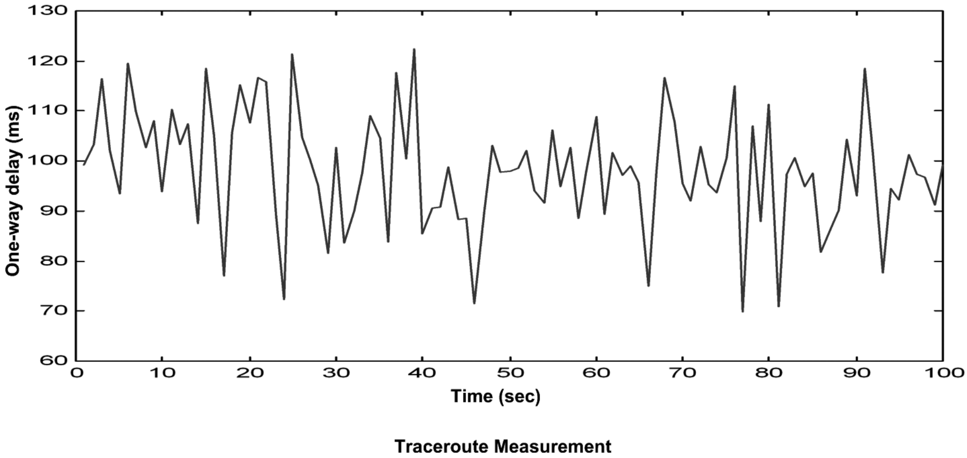

Figure 2: One-way delay variations

variation is. Figure 2 depicts this nature, which is obtained by tracing a path to Ethiopian Telecommunications Corporation Internet DNS server. The significance of this is, even if a packet traverses the same path it will get delayed randomly, therefore, unless voice traffic enjoy exceptional handling, it will suffer voice degradation due to jitter.

3.5. Observation

The inherent self-similar nature of the Internet made the possibility of carrying voice traffic to be accompanied by voice quality degradation, of course, unless proper QoS measures are taken. Delays relating to the Queueing performance degradation in self-similar packet arrivals may get intolerable at the receiving side. Further the self-similar or variable nature of the packet delay process implies voice quality degradation due to inter-packet arrival variations. For paving the way for VoIP gradually replace the existing PSTN, these issues should be resolved. The conclusion is that; to achieve good quality of voice over the Internet, voice traffic should be handled independently of the remaining Internet traffic, through implementations of new features in the existing Internet infrastructure.

4. TCP PERFORMANCE ANALYSIS

In this section, the effect of packet round trip delay self-similarity on TCP's throughput performance, which is captured through the Retransmission Time Out (RTO) parameter, is investigated. It is important to evaluate the impact because TCP is the most dominant *transport* protocol in the current Internet, and its performance depends on the packet round trip delay or RTT (Round trip time). Internet packet delay is Long Range dependent. It is bursty across multiple time scales, which implies that end-user Quality of Service in the Internet is likely to be affected by long period of very large and/or highly variable delays.

4.1 TCP Transport Mechanism

TCP employs an ACK (acknowledgment)-based window control, that is, a TCP sender updates its congestion window size every time it measures RTT (Hagiwara *et al*, 1999). This measurement is taken randomly. Therefore, TCP throughput performance depends on the RTT. With the most basic mechanism in the Internet to detect losses, the sender retransmits a segment if its ACK has not been received in the expected amount of time (this is RTO based on the RTT). If segment losses are detected by RTO, TCP congestions window, which specifies the amount of data the TCP sender can transmit before receiving an ACK is reduced to one segment, which implies RTO based method effectively reduce throughput. Therefore, the goal of a good RTO estimator should be to minimize the number of unnecessary re-transmissions due to failure in proper estimation of RTO. If RTT drastically changes, RTO may not be able to adapt to the RTT and the TCP sender may end up retransmitting segments.

RTO is calculated by using Jacobson's algorithm (Stevens, 2000). That is

$$RTO = SRTT + kRTTVAR \tag{4.1}$$

$$RTTVAR = RTTVAR + h(|SRTT - RTT_{MEAS}| - D) \tag{4.2}$$

with $k = 4$, $h = 0.5$ in many cases, where *SRTT* is a smoothed estimate of RTT and *RTTVAR* is a smoothed estimate of the variance of RTT. Both variables are updated every time an *RTT* measurement RTT_{meas} is taken. *SRTT* is updated using an exponentially weighted moving average (EWMA) with a gain of α.

$$SRTT = (1 - \alpha)SRTT + \alpha RTT_{meas} \tag{4.3}$$

4.2 Experimentation

TCP employs a Retransmission Time Out (RTO) parameter to re-send a packet for which it didn't get an ACK (Acknowledgment). TCP dynamically re-calculates the value of the *RTO* based on the measured Round Trip Time (RTT), which is equivalent to the Packet Round Trip Delays. Due to the high variability or bursty nature of RTT, TCP may invoke premature retransmission time out and waste useful bandwidth on the link. Here, a study that catches the percentage of RTO events-invoked whenever the sender failed to receive ACK packets is dealt with. The relation between the occurrence of RTO and Packet Round Trip Self-Similarity is also to be analyzed.

For studying the relation between occurrence of RTO and self-similarity of packet round trip delays, we used an experiment using the Ping application, which is part of the TCP/IP protocol stack. Ping uses ICMP echo request message to a host expecting an ICMP echo reply to be returned. If the echo doesn't returned in the pre-specified duration (depends on the implementations of the Ping program), the application issues an "IP Request Time Out" message. This may happen in two cases: the first case is when a packet gets lost due to congestion on the Queues along the route, and the second is if a packet encountered a prolonged delay at the intermediate Queues. Unless mechanisms are implemented within the Ping application to track the variability or burstiness of the Packet Round Trip Times and dynamically update the pre-specified duration, the amount of premature "IP Request Time Out" messages get increased along with an increase in traffic.

In a similar fashion, TCP requires ACK packets to arrive from the receiver to determine whether a packet gets delivered. Otherwise it will keep re-sending the same packet. TCP employs RTO to invoke a packet retransmission event, but implements ways for dynamically calculating RTO from randomly measured RTT values. The failure in getting ACK message may happen either due to packet loss or prolonged delay. These cases are similar to the events we catch using the Ping Application. Therefore, having knowledge of the percentage of the "IP Request Time Out" messages gives a rough estimate on how TCP performs on that link.

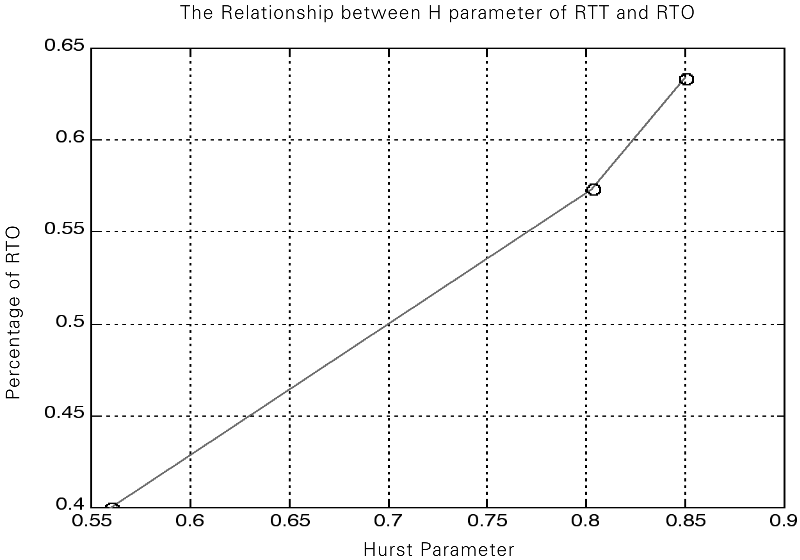

Figure 3: Hurst Parameter versus Percentage of Retransmission

We perform a continuous Ping operation to the yahoo.com server (IP Address: 64.58.79.230). Then the percentage of the "IP Request Time Out" messages out of the replies to the total Ping requests is calculated. By taking out the "IP Request Time Out" events from the collected traces, we rearrange the successful Ping reply Round Trip Times as a time series, then the corresponding Hurst parameter of this time series estimated. The relation between Hurst parameter and percentage of RTO is plotted in Figure 3 which is linear up to H= 0.8 and beyond this it is nonlinear due to self similarity.

4.3 Inference

The relation between the Percentage of "IP Request Time Out" messages and the Hurst Parameter of the Packet Round Trip Value is presented in Figure 3 which is a useful graph to draw the following inferences. The percentage of RTO increases linearly as H increases and hence results wastage in bandwidth. The bandwidth wastage increases significantly due to self-similarity of the packet round trip delay as H tends to unity. This in turn degrades the TCP throughput performance. From this it can be concluded that, as the self-similarity of the Packet Round Trip delay increases, the wastage in bandwidth is due to increase in packet re-transmission.

This experimentation is an indirect measure of TCP's performance through another application of the TCP/IP protocol suite and failure of simulating a real-time environment. In a similar manner to Queueing performance analysis, the limitations can be improved by using Real Network environment or real-time Network Simulators. Hagiwara *et al* (1999) did research (refer Figure 4) on the effect of Packet Round Trip Delay or Round Trip Time (RTT) self-similarity on the performance of TCP using the LBNL-NS simulator to simulate a network environment, with all the links using TCP connections. The TCP senders send files according to the Pareto Distribution (Taye, 2003) to create aggregated self-similar traffic, and RTO estimations are evaluated using different Hurst parameters of RTT to investigate premature time outs. RTO at *H* values near 1 (H≈ 1) is found to be burstier than RTO at *H* values near *0.5*. This result of Hagiwara *et al* (1999) agrees with our findings and the same is presented in Figure 3.

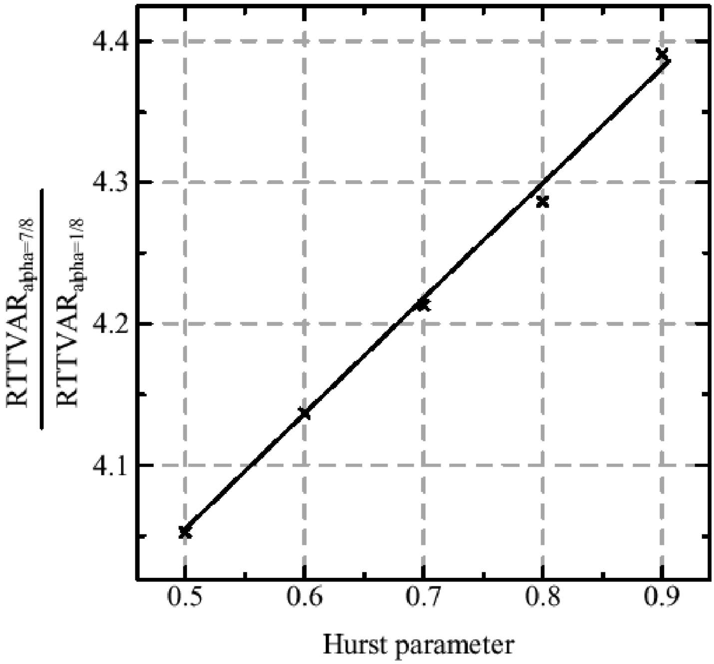

Figure 4: Hurst parameter effect on retransmission of TCP (from Hagiwara *et al*, 1999)

5. CONCLUSIONS

The following conclusions are drawn from the discussions and experimentations covered in this paper.Queueing performance has been affected due to the Self-Similar or Long Memory nature of the packet arrival process, that is, Queueing performance degraded as the self-similarity of the Packet arrival process increases. As the self-similarity of the Packet Round Trip Delays increases, TCP throughput performance is degraded. As the Hurst parameter of the Round Trip Time becomes near to unity, the wastage in bandwidth due to packet re-transmissions is increased. The self-similarity of Internet degrades voice quality in VoIP. Both delay and delay-variation or jitter behaves erratically due to the self-similar nature of packet arrival and delays. To achieve good quality of voice over the Internet, voice traffic should be handled independently of the remaining Internet traffic, through implementations of new features in the existing Internet infrastructure.

REFERENCES

BORRELA, M.S., ULUDAG.S. and BREWSTER.G. (1997): Self-similarity of internet packet delay, *IEEE ICN'97*, January, 513–517.

BORRELA, M.S. and BREWSTER, G. (1997): Measurement and analysis of long-range dependent behaviour of internet packet delay, *IEEE/INFOCOM*, 497–504.

DAVIDSON, J. and PETERS, J. (2000): Voice over IP fundamentals, Cisco Press, Indianapolis.

GARRET, M. and WILLINGER, W. (1994): Analysis, modeling and generation of self-similar VBR video traffic, *SIGCOMM*, Networking Transaction.

HAGIWARA, T., MASHIMA, H., MATSUDA, T. and YAMATO, M. (1999): Impact of round trip delay self-similarity on TCP performance, Osaka University.

LELAND, W. E., TAQQU, M. S. and WILSON, D. V. (1994): On the self similarity nature of Ethernet traffic (Extended Version), IEEE/ACM transactions on Networking.

LELAND, W. E., TAQQU, M.S. and WILSON, D. V. (1994): TCP/IP Illustrated, Volume 1, Addison Wesley, New York.

PARK, K., KIM, G. and CROVELLA, M. (1997): On the effect of traffic self-similarity in network performance, *1997 SPIE*, International Conference on Performance and Control of Network System.
PARK, K. and WILLINGER, W. (Ed) (2000): Self similar network traffic and performance evaluation, John Wiley and Sons.
TAYE, A. (2003): Internet self-similarity, modelling and performance evaluation, a thesis submitted for a partial fulfillment of M.Sc. in Electrical Engineering, Addis Ababa University, Ethiopia.
TAYE, A., MOHAMMED, A. and DEVARAJAN, G. (2002): Internet traffic and packet round trip self-similarity, *Zede Journal of Ethiopian Association of Engineers and Architects*, 19: 53–64.

BIOGRAPHICAL NOTES

Devarajan Gopal

Dr Devarajan Gopal was born in Dharmapuri district, Tamilnadu, India. He received the B.E. in electronics and communication from the College of Engineering, Guindy, Madras University, the M.Tech. in communication and radar engineering from the Indian Institute of Technology, New Delhi, and the PhD at Cochin University of Science and Technology, India. He joined the Defence Research and Development Organization in the Government of India and served as a Scientific Officer/Scientist for about two decades. During this period he was involved in radar, sonar and sonobuoy systems design and development. He received the prestigious National Scientific Award "VASVIK" for his contribution to "Airborne Sonobuoy System and Modular Data Bus" in 1991. He was associated with different universities in India and Ethiopia for curriculum revision at the undergraduate and graduate levels. He has held various positions at Madras University and Addis Ababa University. Currently, he is working as a Professor in the Department of Digital Information and Communication at Woosong University, Korea. His fields of interest are in digital signal processing, computer networks and broadband communication. He is a Fellow of Institution of Engineers. He has a number of publications to his credit.

Taye Abdulkadir

Taye Abdulkadir was born in Ethiopia. He received his BSc degree in Electrical Engineering from Addis Ababa University (AAU) in 1997. He earned his MSc degree in Communications Engineering from AAU in 2003. His research area is in Network Traffic, Self-Similarity, Traffic Modeling and Network Performance. For the last seven years he has worked in Computer Networking (design, installation, performance studies and administration) with different institutions in Ethiopia and also teaches at the Department of Electrical and Computer Engineering, AAU.