PERFORMANCE STUDY OF END-TO-END RESOURCE MANAGEMENT IN ATM GEOSTATIONARY SATELLITE NETWORKS

Güray Açar [1]
Catherine Rosenberg[2]

acarg@ecn.purdue.edu
[1]Imperial College, Dept of EE., London, UK

cath@ecn.purdue.edu
[2]Purdue University, Dept. of ECE., West-Lafayette, USA

ABSTRACT

This paper deals with end-to-end resource management in geostationary (GEO) satellite networks. Bandwidth on Demand (BoD) is central to end-to-end resource management in these systems. In this paper, using simulation analysis, we study the impact of BoD reservation parameters on the Quality of Service (QoS) received by the connections, the resource utilization, and the Grade of Service (GoS) of the network. These results are important since they give us a lot of insights on how the various parameters of the BoD impact the performance.

INTRODUCTION

In this paper we concentrate on GEO based satellite networks. They play an ever-increasing role in the public and private internets, due mostly to their large geographic coverage, inherent broadcast capabilities and fast deployment. See [2] for more details on the systems, issues and solutions.

In this paper, we will focus on bent pipe satellite systems. A typical system will have thousands of individual users connected to the satellite through Satellite Access Units (SAUs). The system also comprises a gateway (GW) that connects the satellite network to other networks and is in charge of most of the signaling and management functions in the satellite network. SAUs send traffic to the gateway via the satellite through a MF-TDMA (Multi-Frequency Time Division Multiple Access) return link. The satellite acts as a mere repeater. The gateway sends traffic to the SAUs via the satellite using a very high-speed broadcast forward link.

Satellite networks are multiple access systems with limited transmission capacity compared to terrestrial networks. Therefore, end-to-end resource management for such systems is key to deliver acceptable QoS to users while providing adequate efficiency. Bandwidth on Demand (BoD) is central to end-to-end resource management. It is defined as a set of MAC (Medium Access Control) protocols and algorithms that allow connections to request resources on a demand basis, while the connections are already in progress, in an environment where many bursty connections share a common medium access link. We have designed a method integrating a BoD process with a Call Admission Control (CAC) scheme for an ATM geostationary satellite network ([3]). In this paper, we present further studies and results on this method that show the importance of using reservation carefully. More precisely, we discuss performances in terms of delay and efficiency using a large-scale simulation program.

In a bent pipe system, terminals (i.e., SAUs) use BoD to request bandwidth on the return link, which is the scarce resource to be managed efficiently and fairly. Typically, each SAU will periodically request some resources, (i.e., a number of Time-Slots (TS)) on the return link using signaling. A BoD controller based in the GW would collect these requests and share the resources among the requesting SAUs. More precisely, the BoD process consists of the following five phases:

Phase 1: During this phase, each SAU computes the resource requirements for individual ATM VCs (Virtual Connection) or for groups of ATM VCs (BoD Entities). In [1] we presented two *RRE (Resource Requirement Estimation)* algorithms to perform this first phase.

Phase 2: It consists in signaling the resource requirements in the form of *Resource Requests* (RR) from the SAUs to the BoD controller.

Phase 3: The third phase is crucial. The BoD controller has to compute the fair and efficient allocation of the return link resources (i.e., the Time-Slots) to the VCs (or BoD entities). This results in the creation of the *Burst Time Plan* (BTP). An algorithm to perform the fair and efficient sharing of resources is presented in [3].

Phase 4: It consists in signaling the response from the BoD controller to the SAUs (broadcast of the BTP).

Phase 5: This last phase is performed by each SAU that has to share the TS it has received among its different connections. This is an internal scheduling phase.

The next section will briefly describe the BoD and CAC integrated process. Emphasis will be given to phases 1 and 3. Then we will present a brief description of our

simulation program and show the results of our simulation analysis. Finally, we will present our conclusions.

BoD AND CAC INTEGRATED PROCESS

Note that not all connections will use BoD. Connections with strict delay constraints will only rely on statically allocated resources assigned at call set-up, because of the low responsiveness of BoD in a GEO system. The time interval between RR signaling and the reception of the corresponding BTP is the *response time*. With the BoD controller being in the gateway, the response time will be at least 500 msec, because of the 125-msec propagation delay between the SAU and the GEO satellite.

Phase 3: In the following we will assume that a request is sent per VC using BoD, i.e., we do not assume aggregation. Each such VC is assigned a Static Resource (SR) and a Booked Resource (BR) by the satellite network at call set-up. These parameters can be zero depending on the type of the VC (i.e., Variable Bit Rate (VBR) or Unspecified Bit Rate (UBR)). SR is the amount of return link resources (a number of TS per frame) that is statically allocated to the connection. On top of SR each connection using BoD can request additional resources via RR signaling. BR is the amount of resources that is *booked* for the connection. If in a given period, the RR for the connection is less than its BR the BoD controller fully grants the RR. Otherwise the connection is allocated its BR, and a fair share of the available best-effort capacity. Equation-1 illustrates the relation between RR_j and BR_j for a VC_j, where y_j represents the amount of resources allocated to VC_j having requested RR_j.

$$\text{If } \left\{ \begin{array}{l} RR_j \le BR_j \Rightarrow y_j = RR_j \\ RR_j > BR_j \Rightarrow y_j = BR_j + x_j \end{array} \right\} \qquad \text{Equation-1}$$

In a given period, after the BoD controller has allocated to each VC_j the minimum of RR_j and BRj, it can compute how much available capacity C_A it has left. It can then compute how many more TSs VC_j would receive assuming that $RR_j > BR_j$. In Equation-1 x_j is the share of C_A that VC_j would receive. We proposed in [3] a method based on Game Theory to share C_A fairly. Very briefly, assuming that there are N BoD connections and M non-BoD connections in the current period, the x_j's are solution to the following optimization problem:

$\max \prod_{j=1}^{N} \alpha_j$ subject to:

1. $\alpha_j \le \max\left(0, \left(RR_j - BR_j\right)\right)$ For $\forall j$
2. $\sum_{j=1}^{N} \alpha_j \le C_A$

with $C_A = C - \sum_{j=1}^{(N+M)} SR_j - \sum_{j=1}^{N} \min\left(RR_j, BR_j\right)$, where C is the total return link transmission capacity.

Phase 1: The resource requirement estimation (RRE) phase is another essential component of the BoD process, and is explained in the following (see [1]for details). Let *i* be the number of frames during one response time. The BTP corresponding to a RR computed in the k^{th} frame will arrive and be effective after a complete response time, i.e., in the $(k+i)^{th}$ frame. This is the *target frame*. In other words, the objective of the RRE algorithms is to estimate the amount of resources that the connection will require **during the *target frame*,** which is one response time (i.e. *i MF-TDMA frames)* after the moment the resource requirement is computed. Our RRE algorithm aims to compute the number of ATM cells that will **certainly** be ready to be transmitted in the *target frame*. Accordingly, there are two basic **assumptions** behind our RRE algorithm:

i. there will be no cell arrivals from the moment RRE is invoked till the end of the target frame,
ii. all past (i-1) RRs will be fully granted by the BoD controller.

The RRE algorithm must be such that the SAU and the BoD controller are somehow synchronized in terms of the *request-reservation* process. There is a need for a *memory element* that will remember those RRs that are not fully granted by the BoD controller. We have developed two approaches to deal with this problem [1]. We describe the more promising here, which is called the RRE Algorithm with memory at the BoD Controller.

If the BoD controller cannot fully grant a RR(k), which was sent in the k^{th} frame, by a given VC, it keeps in memory what it was not able to allocate, say t(k), and adds it to the next request coming from this connection (i.e., RR(k)=RR(k)+t(k-1)). In other words, if the BoD controller can only partially accept a RR, it remembers that it *owes* the remaining part and will try to allocate it in the next frame. The SAU, on the other hand, remembers how much resource the BoD controller owes, and avoids re-requesting this amount of resource. In order to implement this approach we need two variables per connection. The first one, t(.), represents the memory of the BoD controller and is held at the BoD controller, while the second one, p(.), is held at the SAU and is the vision that the SAU has of the BoD controller memory. These variables are necessary to deal with RR losses due to transmission errors.

$t(k) = \left[t(k-1) + RR(k) - N(k+i)\right]^{+} \qquad k \ge 0$ where $t(-1) \equiv 0$

Note that RR(k) is the k^{th} RR arriving at the BoD

controller, and N(k) is the number of TSs allocated to the VC for the k^{th} frame. The response to RR(k) will be received by the SAU in the $(k+i)^{th}$ RR period, corresponding to N(k+i). Similarly,

$$p(k) = [p(k-1) + RR(k-i) - N(k)]^{+} \quad k>i \text{ where } p(i) \equiv 0$$

As long as there is no RR losses the equation below must hold. Note that there is a need for a mechanism to re-establish the request-reservation synchronization between the SAU and the BoD controller when a RR is lost.

$p(k)=t(k-i) \qquad k>i$

Then our RRE algorithm is represented by Equation-2 below.

$$RR(k) = \left[q(k) - N(k) - (i+1) \cdot SR - \sum_{r=1}^{i-1} RR(k-i+r) - p(k) \right]^{+} \qquad \text{Equation-2}$$

where q(k) is the number of cells of the connection waiting in the SAU buffer at the beginning of the k^{th} frame.

BoD & CAC Integration and QoS

In ATM networks, the objective of the CAC is to limit the number of connections within the network so that each connection receives sufficient amount of network resources to meet its guaranteed QoS requirements.

In multi-service networks, users have different traffic characteristics and different QoS requirements. Managing such networks and offering differentiated QoS to different traffic classes require some segregation among service classes. What we call segregation is the ability for a network to protect the QoS of each class from the behavior of the others. In terrestrial network, the switches (or the routers) are responsible for performing segregation. In general, we expect a switch to be able to segregate between service classes (e.g., between VBR and UBR) and to segregate within a class between different connections. Different schedulers are being used to typically allocate the output link capacity among several traffic classes.

Just like a scheduler in a terrestrial switch, the BoD process is the resource manager in the satellite network in charge of sharing the return link capacity among different traffic classes. SR and BR are the means by which the users of the network can be guaranteed QoS. Using SR and BR, the BoD process allows us to segregate not only among service classes but also among the connections within the same service class. For instance, in a network with non-real time VBR and UBR connections, we may allocate non zero SR and BR for VBR connections while UBR connections would not be allocated any static or booked resources. SR and BR will not only favor VBR connections by making sure that some of the return link capacity is dedicated to them, but also guarantee various levels of QoS to different applications within the VBR service class by reserving different amount of resources for each connection. Note that if a connection does not need its SR for a period of time, only other connections within the same SAU can use the corresponding TSs while for BR, any other connections in the system can use the corresponding TSs.

The method for CAC and BoD integration, which is explained in [3] in detail, ensures that a VC_j will be accepted into the network if it can be allocated its SR_j+BR_j. This is expressed by

$$SR_j + BR_j + \sum_k SR_k + \sum_k BR_k \le C \qquad \text{Equation-3}$$

where C is the total capacity of the satellite return link. For reasons to be explained later, Equation-3 will be modified later.

The SR and BR assignment to connections do not only determine the QoS guaranteed to the connections, but also determine the maximum number of connections with guaranteed QoS that can be *supported by* the network. In other words, SR and BR assignment to the BoD connections is an issue that has great impact on both the QoS guaranteed to the connections and the GoS.

The choice of the right values for SR and BR for a connection is a complicated issue. A connection can request (and very probably will get), using BoD, much more than its BR thanks to statistical resource sharing among all the connections. On the other hand, SR is a fixed amount of resources that is allocated to the connection every frame. Therefore the same amount of resource, say *x*, provides a different QoS to a VC depending on whether it is allocated in a static-only fashion (i.e. SR=x, BR=0) or a booked-only fashion (i.e. SR=0, BR=x). SR and BR could also have different impacts on the GoS of the network as the resources that are statically allocated to some connections cannot be statistically shared by other users.

SR and BR can be envisaged as the tools provided to the network operator in order to fine-tune the trade-off between the QoS guaranteed to connections and the GoS of the network. In the next section, we will present the results of our simulations. We will present the impact of SR and BR on the service segregation in the satellite network, the delay characteristics experienced by the connections, and the amount of resources (i.e., Time-Slots) wasted by the connections.

SIMULATION AND PERFORMANCE ANALYSIS

Simulator Program

In our simulations, we assumed only one connection per SAU, where each connection is an ATM cell stream that is generated by a 2-state MMDP (Markov Modulated Deterministic Process) with PCR (Peak Cell Rate) equal to 192 kbps. The mean burst length (i.e., mean ON period) is equal to 200 msec and the mean inter-burst time (i.e., mean OFF period) is varied to attain different SCR (Sustainable Cell Rate). Also note that the Maximum Burst Size (MBS) of each connection is kept equal to 1024 msec (i.e., 512 cells) by means of a leaky bucket.

We assumed that we had 4 MF-TDMA carriers and a frame of 32 TS. Each TS is one msec long, and can carry one ATM cell. Accordingly the maximum useful transmission rate is 384 kbps, and one TS per frame corresponds to a useful transmission rate of 12 kbps.

In our simulations we assumed that the processing time at the BoD controller, the transmission time for signaling the RRs and broadcasting the BTP will add up to a total of 76 msec. Under these assumptions the response time is 576 msec. Hence there are i=18 frames within one response time.

Note that SR, BR, SCR, and PCR are all expressed in terms of number of TSs per frame. The total of SR and BR for a connection will also be expressed in terms of TSs per frame, and denoted as TTS (Total Time Slots).

We define two types of connections, type-1 and type-2, which are identical in their traffic characteristics. The only difference is that type-1 connections are not assigned any SR or BR while type-2 connections are guaranteed some QoS via SR and/or BR assigned to them. We assume that there are N1 type-1 and N2 type-2 connections in the network. The total number of connections in the network, N=N1+N2, is kept constant at a value to ensure a network load (i.e. total mean cell arrival rate/total capacity). Note that we have kept the load under control to be able to compare results of our studies. In reality, our CAC scheme cannot limit the number of type-1 connections as they are not assigned any SR or BR and the CAC is only based on these parameters. Therefore there is no way to control the network load.

Service Segregation

Each connection has a SCR=2 cells/frame, and type-2 connections are assigned BR only. Note that SR is kept zero in this part of the study. The network load is 0.9; hence there are 58 active connections in the network (i.e., N=58).

First we measured the mean queuing delay in the system when we have only connections of type-1. Then we introduced type-2 connections with a given TTS value (i.e. TTS=BR, because SR=0), and measured the mean queuing delay for both type-1 and type-2 connections. The number of type-2 connections in the network was increased from 1 to the maximum possible number, which is computed using the CAC scheme expressed by Equation-3. The results are presented in Figure-1 that illustrates the segregation, in terms of mean queuing delay, between type-1 and type-2 connections provided by non-zero TTS. The more we increased TTS for the type-2 connections the shorter their mean queuing delay became. In return, the mean queuing delay experienced by type-1 connections increased.

Note that increasing TTS beyond a certain point has almost no effect on the mean delay while it has a large impact on the maximum number of type-2 connections (hence the GoS of type-2), and that it is impossible to reduce the mean queuing delay lower than one response time in the network (i.e. 576 msec). Applications that require a shorter mean queuing delay must be assigned a non-zero SR.

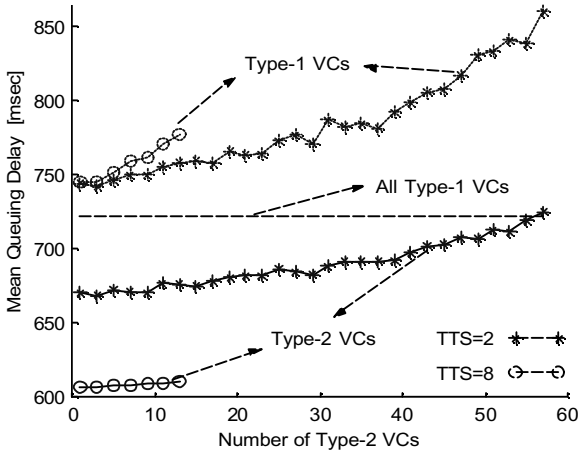

Figure 1. Service segregation illustrated

Figure-2 shows the mean queuing delay that a type-2 connection would experience as a function of the TTS (i.e., TTS=BR, SR=0) under various network loads. Note that the number of type-2 connections is at the maximum possible value. It is seen in this figure that the mean queuing delay experienced by a type-2 connection is very much dependent on the current network load for low values of BR. As we increase BR, the mean queuing delay and its variation with respect to the network load will reduce. However, as we mentioned before, increasing BR means reducing the number of type-2 connections in the network. Here we observe again the trade-off between the number of type-2 connections (i.e., the GoS) and the mean queuing delay experienced by each type-2 connection.

Mean Queuing Delay and Resource Waste

Further reduction in the mean queuing delay for type-2 connections can only be achieved by assigning SR to them. In this part of our simulations we present the relation between the mean queuing delay and TS (Time Slot) waste because of non-zero SR assignment.

The network load is kept constant at 0.9. Each type-2 connection is assigned both SR and BR, where SR+BR=TTS. The mean queuing delay is measured for type-2 connections and is shown as a function of the percentage TS (Time Slot) waste in Figure-3 for SCR=2, 4 and 8 cells/frame. TS waste occurs when a non-zero SR is allocated to a bursty connection since we have assumed 1 VC per SAU.

In Figure-3 a somehow surprising result is presented. For high values of TTS we observe the expected trade-off between TS waste and mean queuing delay. For high values of TTS we are able to reduce the mean queuing delay by increasing SR provided that we are content with the increasing resource waste.

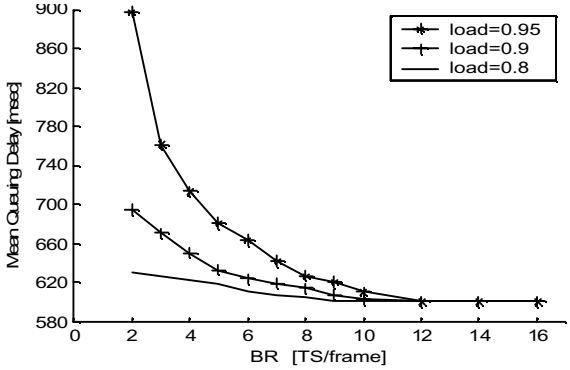

Figure 2. Impact of BR alone

However, for low values of TTS, increasing SR does not necessarily decrease the mean queuing delay. Indeed in those figures, increasing SR while keeping (SR+BR)=TTS, **increases** the mean queuing delay experienced by type-2 connections for TTS values that are smaller than a critical value. We define here the *Critical TTS* as the lowest TTS value for which increasing SR **always** causes reduction in the mean queuing delay for type-2 connections. If TTS is less than the Critical TTS, increasing SR may increase the mean queuing delay for type-2 connections rather than decreasing it. On the contrary, if TTS is greater than the Critical TTS, increasing SR will always reduce the mean queuing delay for type-2 connections.

For a type-2 connection increasing SR will have two contradicting side-effects:

1. The reduction in the mean queuing delay, because every time a burst arrives at the buffer a number of cells will leave quickly using the statically allocated TSs.
2. The increase in the mean queuing delay. The more resources are assigned statically, the less statistical resource sharing takes place among the connections and hence the probability that the BoD controller will not fully grant some RRs will increase.

Depending on the traffic characteristics of the connections, the TTS (i.e., SR+BR) assigned to the type-2 connections, and the current network load, one of the side-effects listed above will prevail, and the mean queuing delay will either decrease or increase.

In Figure-4 we present the variation of the ratio (Critical TTS/SCR) with respect to the burstiness of the connections (i.e., PCR/SCR). It is clearly seen in this figure that as the burstiness of the connections decreases the critical TTS approaches the mean cell arrival rate of the connection. As the burstiness of the connections increase the (Critical TTS/SCR) ratio increases almost linearly. A brief observation of Figure-4 reveals that for a connection with a burstiness factor of 8, the Critical TTS is 6 times the SCR of the connection. That is, the connection must be assigned a TTS greater than or equal to 6 times the connection's SCR if we want to be sure that the mean queuing delay experienced by the connection will decrease by increasing SR/BR ratio.

The second side-effect of increased SR, which is the increase in mean queuing delay because the statistical resource sharing is reduced, can be mitigated by limiting the amount of resources that can be statically allocated to connections. This implies a modification of our CAC scheme, which was expressed in Equation-3. The modified CAC scheme is shown in Equation-4 below.

$$\left\{ \begin{array}{c} SR_j + \sum_k SR_k \le \varepsilon \cdot C \\ SR_j + BR_j + \sum_k SR_k + \sum_k BR_k \le C \end{array} \right\} \quad \text{Equation-4}$$

where $0 < \varepsilon < 1$

In Equation-4, we propose a 2-stage CAC scheme. The first stage checks if the total amount of statically allocated resources to be assigned to all connections exceeds the upper limit on the amount of resource that can be statically allocated to connections, which is denoted by $\varepsilon \cdot C$. The second stage of CAC, which is identical to the CAC proposed earlier, checks if the total SR and BR to be assigned to connections exceeds the capacity of the network. The right value to be chosen for ε is a further research topic.

As it is seen in Figure-3, for SCR=2 cells/frame, which represents the most bursty connection, the steady reduction in mean queuing delay occurs for only those TTS values that are higher than or equal to 12 TS/frame.

More than 80% of the total TSs allocated to a type-2 connection are wasted in order to reduce the mean queuing delay from 600 msec to 100 msec. As the burstiness of the connections decreases the TS waste percentage corresponding to a mean queuing delay value decreases as well. Note that, even for SCR=8 cells/frame, which represents the least bursty connection, almost 30% of all TSs allocated to a type-2 connection are wasted for a mean queuing delay value of 100 msec for type-2 connections.
We have also generated queuing delay histograms. However, due to space limitations we could not present them in this paper.

CONCLUSIONS

Satellites are multiple access systems with very long propagation delay and scarce transmission capacity compared to terrestrial networks. BoD is necessary for the efficient and fair sharing of satellite resources, and for QoS support. SR (Static Resource) and BR (Booked Resource) are the means with which the BoD process guarantees various levels of QoS to connections and the CAC guarantees GoS to some of the connections.

Our simulations illustrated that SR and BR can be successfully used to segregate between different service classes. However, the choice of the right values for SR and BR is a complicated issue because of the impact it has on the QoS that the connections receive and the GoS of the network. The network designer has the challenging task of fine-tuning SR and BR assigned to connections in order to find the optimum trade-off between QoS and GoS.

REFERENCES

[1] G. Açar and C. Rosenberg, "Algorithms to compute for Bandwidth on Demand Requests in a Satellite Access Unit," *Proceedings of 5th Ka Band Utilization Conference*, Taormina, Italy, October 1999.

[2] *IEEE Communications Magazine*, Vol. 37, No. 3, pp. 8-72, March 1999.

[3] C. Rosenberg, "End-to-end Resource Management for ATM On Board Processor Geostationary Satellite Systems," *Proceedings of 4th Ka Band Utilization Conference, Venice, Italy, pp. 481-488, November 1998.*

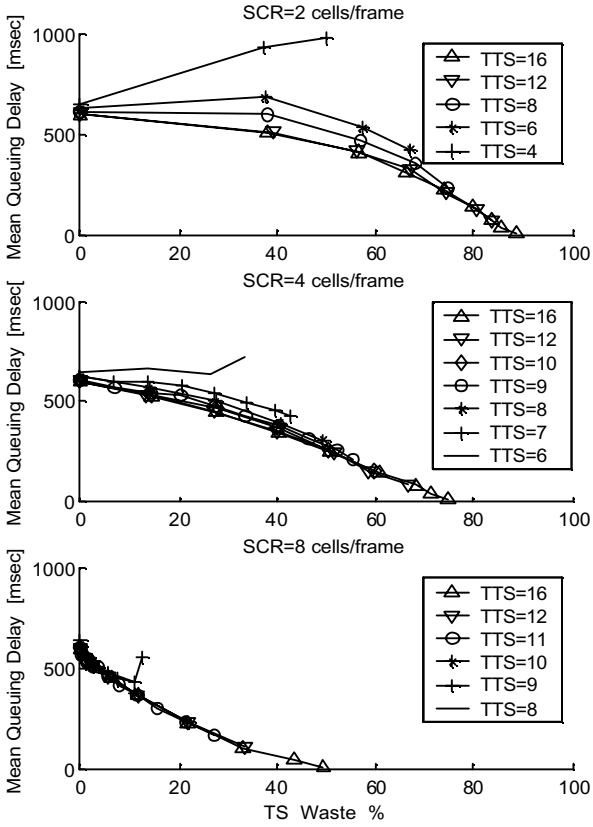

Figure 3. Impact of SR and BR with varying burstiness

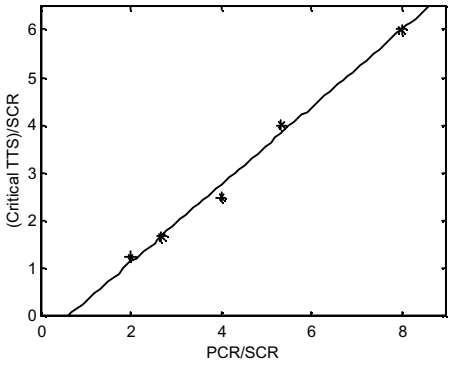

Figure 4. (Critical TTS/SCR) vs Burstiness