A Flow Management Scheme for Multimedia Traffic for High-Speed Network Based on Statistical Weibull Distribution

A. Sharifinejad
The University of New South Wales, Sydney NSW 2052

Abstract
To avoid network congestion it is very important to know the characteristics of the source transmission behaviour. Having a prior knowledge of source behaviour leads us to better source scheduling and prevention of burst overlapping. Here we have investigated the distribution of cell interarrival for a number of different video sources. Also the effect of cell loss for some multimedia applications such as video can be catastrophic. A new proposed algorithm in this paper confines the effect of cell loss to the individual losses.

Introduction
Multimedia consists of a combination of information such as text, data, images, video, audio and graphics presented to the user in an integrated fashion. In the absence of asynchronous transfer mode (ATM), multiple communication technologies were required for transmission of voice, video and data. That meant the requirement of a separate network for each of these applications and additional hardware and software as well as operational cost. However, the appearance of ATM eliminated the need for application dependent networks. Videoconferencing, video on demand, telemarketing, and distance education are some of the multimedia applications which can be run over ATM networks. ATM is a multiplexing and switching technology designed for simplicity, flexibility, versatility and scalability where simplicity is defined in terms of lower complexity in scheduling operations, flexibility in accommodating and satisfying a diverse set of delay requirements, versatility and efficiency in utilizing network resources for a large number of connections and scalability to allow further network growth and heterogeneous network interconnectivity. Therefore, ATM excels where it is desirable for applications with different performance, quality of service, and technical requirements to be performed on the same network. Applications such as multimedia with video at its core consume most of the transmission bandwidth and require a high-level quality of service (QOS) guarantee from the network. ATM is an excellent candidate for transmission of multimedia applications because it is capable of flexibly and dynamically allocating transmission bandwidth as well as connectivity.

Videos, on the other hand, at the heart of multimedia need to be compressed before transmission because of a large amount of redundant information exists in video. So, due to the extremely high bandwidth requirements of uncompressed video streams relative to voice and data, many coding algorithms have been developed for the improvement of video compression. By utilizing many different techniques based on human psychovisual perception of images, colour sensitivity of eyes, strong correlation in spatial and temporal domains etc., video compression removes the redundant information from the video streams. Universally adopted and most commonly used video compression technique, MPEG, consists of different types of frames namely I, P and B, of varying bit rates. A number of these frames are grouped together to form a group of pictures (GOP). Video bit rate characteristics can simply be affected by varying the number and type of frames in GOPs. The bit rate for generic GOP pattern changes abruptly from frame to frame, the I-frame has the highest bit rate among the three frame types, and on the other hand, B-frame usually has the smallest rate. This implies that compressed videos are highly bursty in nature. Hence, the bursty nature of videos which constitute most of the bits in multimedia streams, necessitate managing the traffic efficiently when transmitted over high-speed networks, in particular over ATM network. ATM network technology enables the network to take advantage of the bit rate variations of individual sources through statistical multiplexing. Other advantages of ATM network technology, such as high trunk speed, low bit error rate, flexible service type (bandwidth on demand), and multiplexing capacity make it very suitable for transporting multimedia with QOS guarantees. Statistical multiplexing of bursty video sources in an ATM network improves the bandwidth utilization of the network and lowers the service cost. Furthermore, the multiplexing of multimedia streams reduces the burstiness and bandwidth requirements of the aggregate traffic. However, coincidence of the peak of many video streams may result in congestion in the network and degrade the quality of the video. Traffic management and congestion control are significantly important to guarantee the quality of multimedia. Moreover, in multiplexed transmission, efficient bandwidth allocation among the multiple multimedia sessions will improve network resource utilization. The task of managing the multimedia traffic over ATM networks, thus, is to regulate the traffic so as to obtain high network utilization, avoid network congestion, and provide acceptable QOS at the lower service cost.
Sensible traffic shaping and rate control, active congestion control, fair bandwidth allocation, adaptive flow control and video's stringent delay tolerance are some of the technical challenges to be considered in managing

multimedia traffic over ATM networks. One way of managing the network much more efficiently is by smoothing the entering traffic. A traffic smoothing scheme has to be used at the source or user network interface (UNI) to reduce the burstiness of video signals. In addition, traffic smoothing can enhance the multiplexing gain of a multiplexer. Authors in [YSS92] determined that traffic smoothing increases the link efficiency to about 80% for traffic with short burst repetition period. However, during smoothing video traffic, strict delay constraints are imposed since a video signal is delay-sensitive [OLT92]. In order to develop a successful video traffic smoothing algorithm, the knowledge of burst duration, shape and distribution or simply characteristics of generic video traffic is crucial. To address traffic smoothing issues particularly for multimedia streams, many researchers proposed many different algorithms based on vastly classified characteristics of video streams. Besides traffic smoothing, flow control can be used as a technique to regulate the traffic rate between the sender and the receiver, so that a fast/slow transmitter will not result in overflow/underflow at the receiver. Overflow and underflow are one of many affects to degrade the quality of video. Flow control generally works in conjunction with buffer management or/and a feedback system [ZA98]. Different congestion control algorithms can be incorporated in different ATM service categories depending upon the parameter types and service constraints provided by that category. Two issues are investigated in this paper, firstly the distribution of interarrival of the cell at any particular node (i.e. Switching device) which helps us to be able to do sources scheduling in order to avoid burst overlap and consequently network saturation and collapse. Secondly, the development of a new algorithm to localize the effect of cell losses to those individual losses.

Control Parameters

Multimedia transmission can generally take place over CBR, VBR and ABR. For a CBR connection, the peak cell rate (PCR) is negotiated during connection setup and is guaranteed by the network for the duration of the connection, resulting in high transmission quality and service. However, because of the bursty nature of compressed video, the PCR may not be fully utilized at all times, resulting in low bandwidth utilization and high service costs. A VBR service negotiates the PCR and the sustainable cell rate (SCR) during connection setup, both of which are guaranteed throughout the entire duration of the connection. Current video transmission systems use several connections to transfer scalable video information. Each connection has a different QOS parameter set and traffic definition. Generally, the base layer of coded video, which is more important than the enhancement layer, is transmitted separately on a more reliable connection in terms of cell loss rate.

Congestion Control Schemes

Congestion in the context of networking is defined as a phenomenon where the amount of traffic injected into the network is more than the available network capacity, resulting in possible overflow of network buffers and consequently loss of information and degradation in providing service quality. In ATM networks, different traffic classes with different flow characteristics and QOS requirements are statistically multiplexed to increase the utilization of network bandwidth. Due to the high-speed nature of the ATM environment, the bandwidth control algorithm employed must be both simple and efficient. Simplicity would permit practical implementation of the algorithm while efficiency would minimize the bandwidth required to guarantee the different QOS requirements. There has been much research on determining methods to evaluate the required bandwidth of aggregate sources of ATM traffic [CLW96], and many solutions have been suggested. Some are based on sophisticated mathematical and statistical analysis, and they all need to address the characteristics of the sources. Thus, modelling of ATM sources has also received much attention [HKH97]. Despite this major effort, the problem of determining the bandwidth requirements of connections and of the aggregated traffic at entry and intermediate nodes in the network by taking into consideration the effects of multiplexing, source and traffic characteristics, cross traffic, and buffer sizes largely remains an unresolved issue. All of the Traffic Management techniques can be classified into either using open-loop or closed-loop mechanisms. The open-loop scheme is based on designing and configuring the system independent of any network status factors to avoid the occurrence of congestion. In contrast, the closed-loop scheme in ATM networks is based on periodic feedback from the network, source or destination, and of particular interest, force the ABR sources to change their transmission rate according to network requirement and availability status. The traffic controllability by this technique is an efficient method at the instances of receiving the feedback information. But the closed loop technique does not operate in real time mode. To tackle this issue based on open loop technique, proposals such as the use of larger group of pictures (GOP) in MPEG video coding to reduce the level of video burstiness have been made [AG98]. The following figures demonstrate some of the implemented techniques.

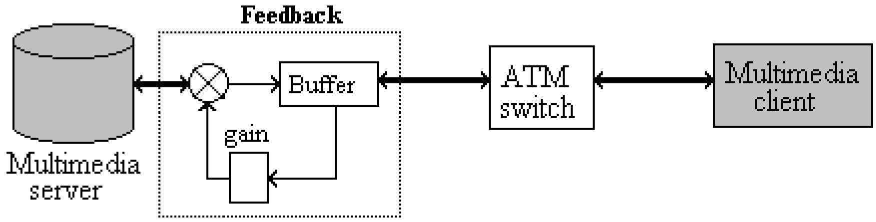

Figure 1, Source rate control with feedback from buffer

When congestion control is implemented at the client/server end, the multimedia server adjusts the video source/server rate based on the level of network congestion and bandwidth availability with the feedback information received from the network (i.e. transcoding). This is slightly different from rate control at the end system in that congestion control at the end system is aimed at helping the network to recover from congestion, whereas rate control is an end-to-end process and prevents the slow processing client from being drowned by a fast transmitting server (i.e. flow control). However, there is no distinct line between two mechanisms as many proposed algorithms integrate many controlling features for better network performance. The servers generally use the feedback from the network as an indication of network congestion level. Authors in [AG00] proposed an alternative method to use a video transcoder in the context of transcoding scheme to reduce the rate of the video bit stream according to the traffic condition based on subjective acceptable minimum video quality. Video transcoding is a further compression of the compressed video such that the bit rate is dynamically controlled according to the channel availability, at the price of poorer picture quality, an additional hardware requirement and cost, and a larger delay. Alternatively, the buffer occupancy at the server can also be used as a measure of the level of congestion in the network. The buffer occupancy can be used to control the transmission rate from the video server [KMR95] [KMM95]. Feedback from the server buffer has the advantage of a much smaller delay when compared to the delay in obtaining feedback from the network. However, it is important to note that server buffer occupancy does not reflect the true network status at any time.

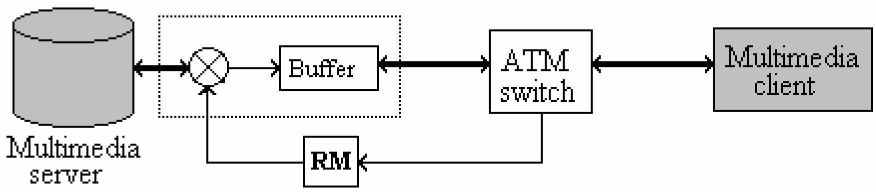

Figure 2, Source rate control with feedback from the network

Congestion control can be done at the network level by switches dropping video cells [K97]. The process of dropping or skipping picture frames helps to reduce transmitted cells when the available bandwidth is temporarily insufficient. The following two methods have been proposed [WSB01] as way for the network to drop video cells; 1) Random Cell Discarding (RCD), 2) Priority Cell Discarding (PCD). In the RCD scheme, the network discards video cells regardless of their importance. Although this method is easy to implement, it may seriously degrade the picture quality if the cells of I-frames are dropped, because I-frames are used as the reference frames, and therefore, any loss in these frames may end up in losses and degradation into subsequent P and B frames. In PCD scheme, video frames are encapsulated into ATM cells with different discard priorities for different frames [K97]. In other words, the cell dropping is based on priority levels and the priority levels are distinguished by the type of video frames.

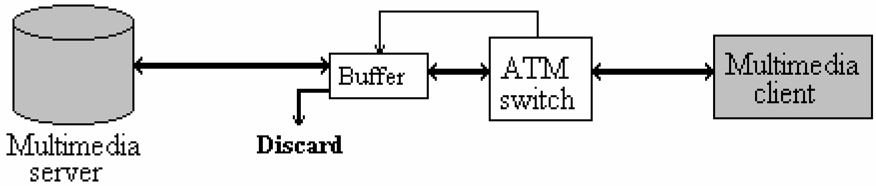

Figure 3, Cell discard scheme depending upon network status

Interarrival distribution

Our investigation and analysis implicate that the cell interarrival time distribution for transmitted cells carrying our four MPEG video sequences maybe at best described by Weibull distribution (Figure 4b, 5b, 6b and 7b). Weibull probability density function is described by equation 1:

$$f(t) = \frac{\beta}{\alpha}\left(\frac{T-\lambda}{\alpha}\right)^{\beta-1} e^{-\left(\frac{T-\lambda}{\alpha}\right)^{\beta}} \qquad (1)$$

Where α = Scale parameter, β = Shape parameter, λ = Location parameter

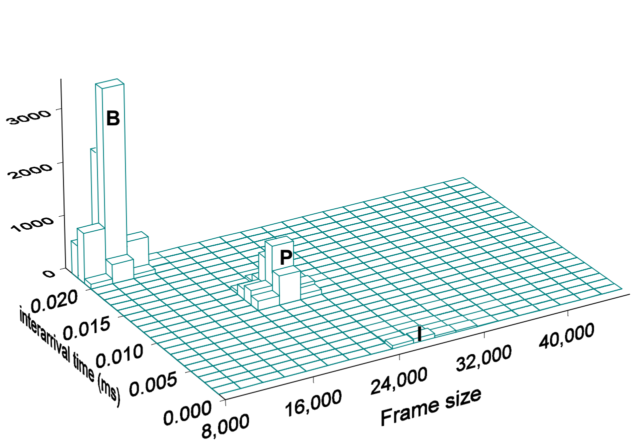

Figure 4a

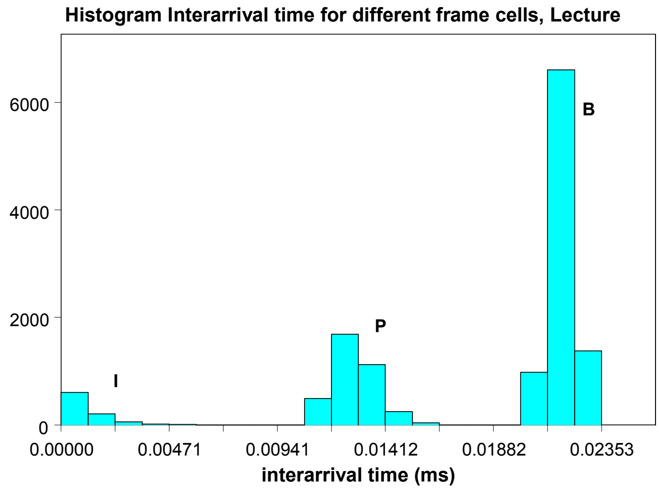

Figure 4b

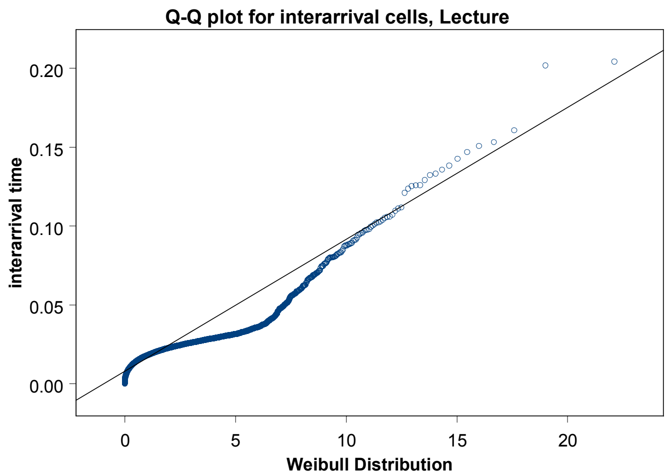

Figure 4c

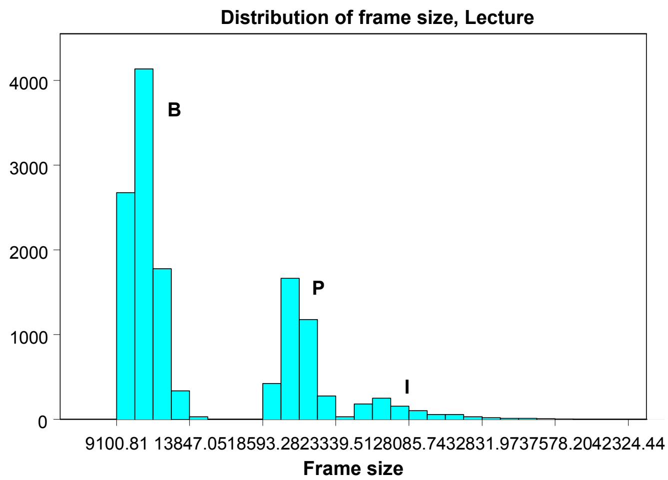

Figure 4d

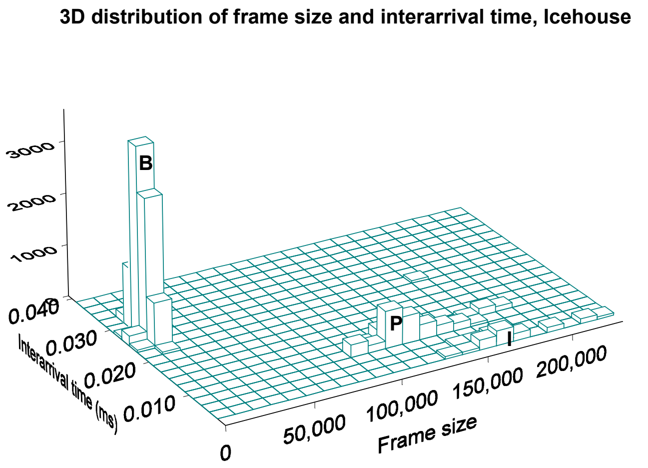

Figure 5a

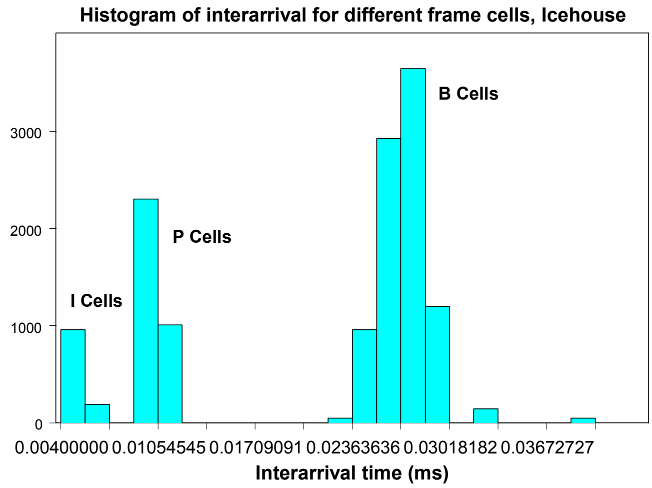

Figure 5b

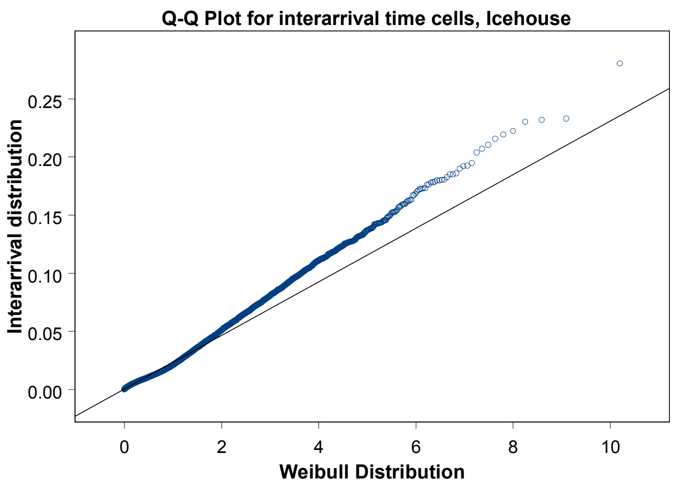

Figure 5c

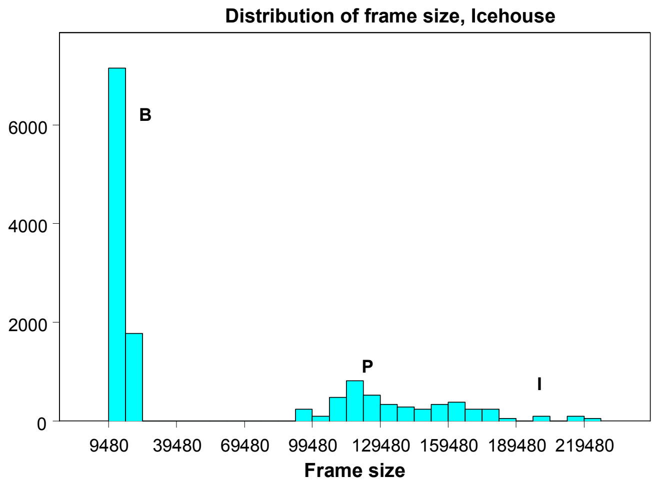

Figure 5d

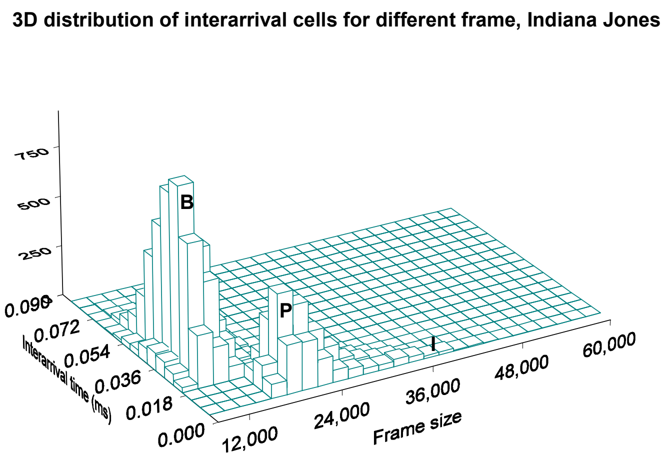

Figure 6a

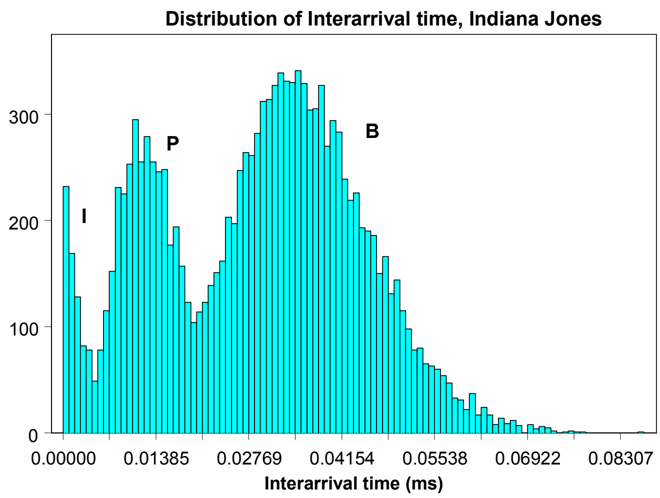

Figure 6b

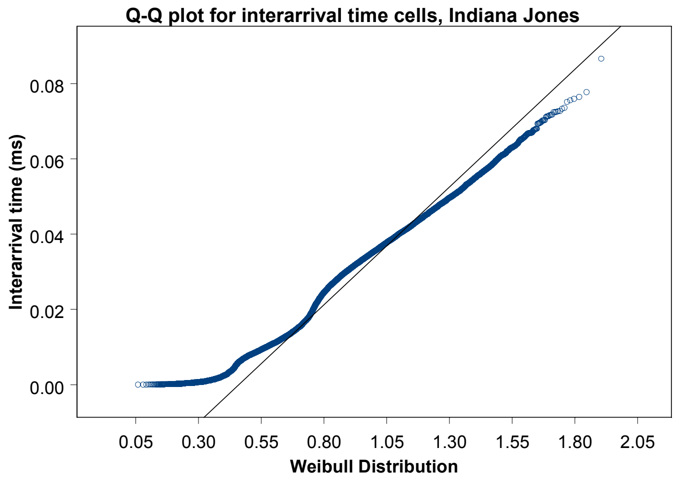

Figure 6c

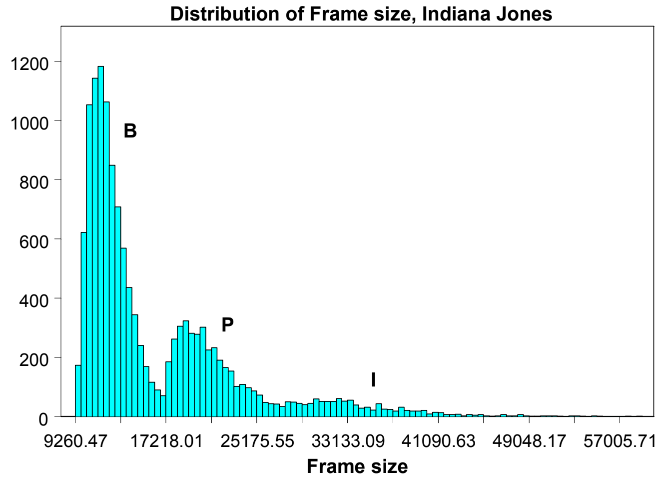

Figure 6d

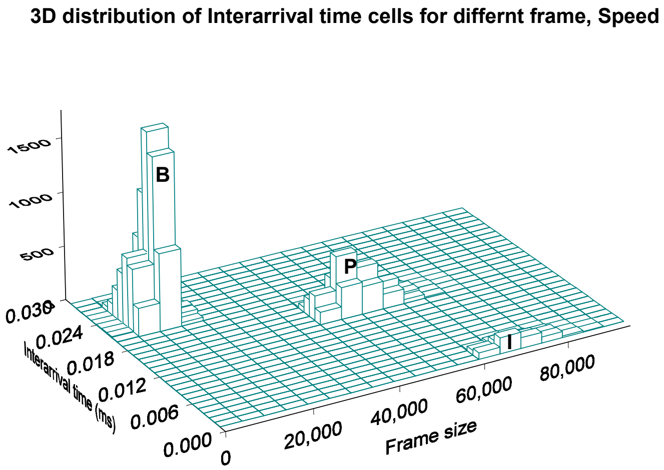

Figure 7a

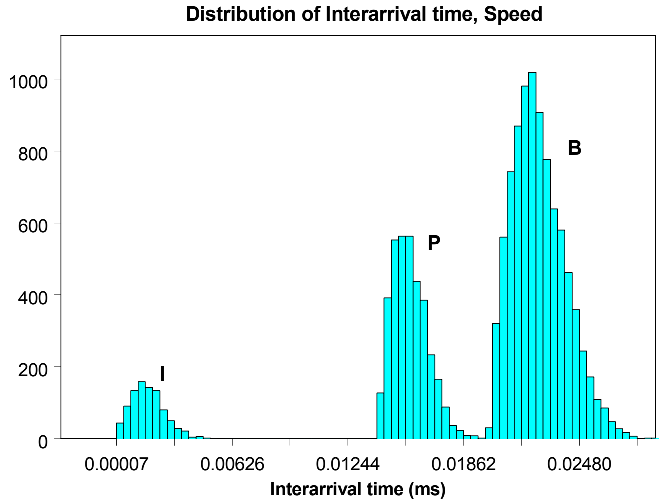

Figure 7b

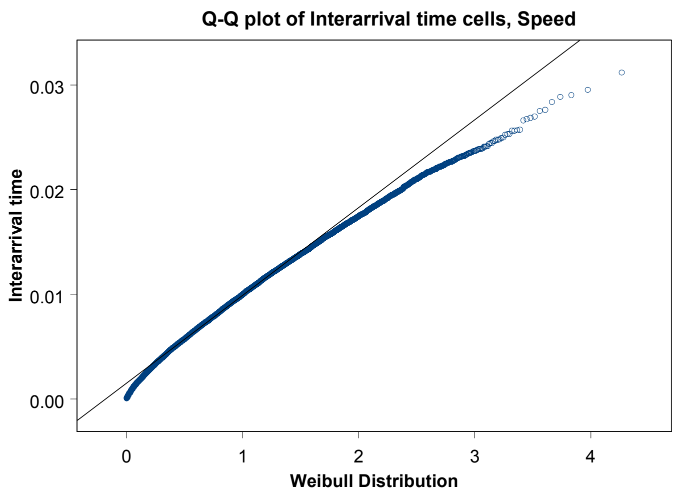

Figure 7c

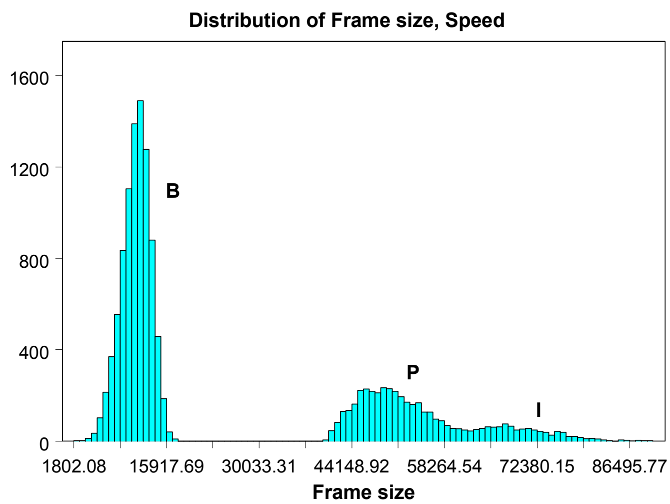

Figure 7d

As the figures show there are burst of cells at short period of time for each video sources. In order to smooth out the traffic, a source scheduling managed by a local server can be very effective.

The Cell Packing Algorithm and Localize Loss Algorithm (LLA)
Packetization of information may have a great impact on the quality of transmitted video in the case of cell losses caused by congestion control cell dropping policy, channel noise, etc. If the frame information stream were packed into a series of cells in a seamless frame-by-frame based manner by which a cell may carry part of the current block and some of the next neighbouring block, any cell loss would have catastrophic outcome on the quality of video since the loss of one cell could have a multi-block damaging impact. So to stop this progressive loss of information through a frame due to loss of single cell, a limiting packing architecture is introduced here to localize any potential cell loss to an individual block region. In our packing proposal as illustrate in Figure 8, the packing is done on the block-by-block basis. That is, the start of every block of a frame needs to be packed into a leader cell where this cell is identified by its bit-flag being set to zero. The rest of the block information is packed into subsequent trailer cells where these cells' flags are set to one until the start of the next block. In case the block information is not enough to fill up the whole 48-byte cell, the rest of that cell will be padded with zeros as so to enable the packing of the start of the next block into a new cell. The rationale in this developed algorithm is that because the smallest identifiable element by its header in a frame is a block, and the header of each block is packed into a new cell, therefore, the magnitude of corruption in a frame due to the loss of any cell carrying any part of a block from that frame would be confined to the size of that block.

In our algorithm, the slices of a frame will be packed alternatively by different priority levels in which the first slice of the frame (which is always accompanied by the frame header) is given lower discarding priority in regard to its counterpart slices. On the contrary, if the lower discarding priority were given to the slices which are not carrying the frame header, the whole frame would be dismissed at the decoder when the cell carrying the frame header was dismissed during network congestion. It should be noted that, during severe network congestion, the allocated channel rates to the ABR sources are set to the MCR values.

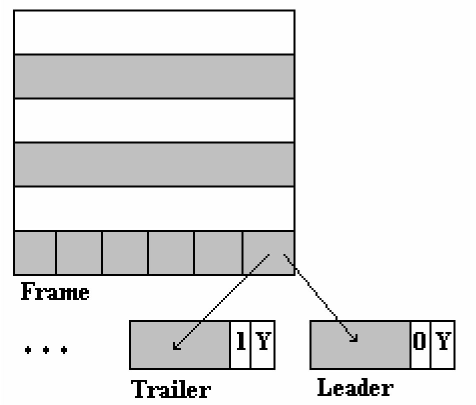

Figure 8
Each block packed into one or more cells

The white slices shown in Figure 8 have a lower discarding priority in contrast to the shaded slices because the white slices in addition to the slice header carry the frame header.

The Results and Discussion
In this simulation experiment, the video streams consist of one high priority VBR source to simulate the bursty background traffic and four ABR videos with different picture contents. The video sequences used in this simulation program are with 57 minutes duration with the total number of 85500 frames each (i.e. 5700 GOP) at 25 frames per second with the frame size of 320×240 and macro-block size of 16×16 and 8×8. The VBR source has a mean bit rate of 1.5 Mbits/s and peak bit rate of 3.85 Mbits/s, while the ABR source has a constant bit rate of at 1.5 Mbits/s. The service rate of the ATM switch was set to 5 Mbits/s.

The simulation was carried out by the Ronin video sequence chosen as VBR stream and the other four Lecture, Icehouse, Indiana Jones and Speed as ABR traffic. The following table shows the characteristics of each video.

	Lecture	Icehouse	Indiana	Speed
Size of file (KB)	4935	4892	5074	3605
Number of I frames	5700	5700	5700	5700
Number of P frames	22800	22800	22800	22800
Number of B frames	57000	57000	57000	57000
Average size of I (KB)	9.3	9.2	9.7	9.7
Average size of P (KB)	5.6	5.5	6.0	6.2
Average size of B (KB)	4.3	4.4	4.4	4.8
Percentage of the total size of I-frames in video	13%	14%	14%	14%
Percentage of the total Size of P-frames in video	27%	26%	29%	30%
Percentage of the total Size of B-frames in video	60%	60%	57%	56%
GOP Pattern	IBBPBB	IBBPBB	IBBPBB	IBBPBB
Transmitted I packets	602,857	605,246	606,754	607,284
Transmitted P packets	1,253,195	1,175,681	1,284,220	1,321,003
Transmitted B packets	2,856,365	2,862,401	2,672,973	2,641,047
Bit stuffing (KB)	223.5	248.7	509.1	225.4
% Redundant data	4.51%	5.06%	10.05%	6.23%

Table 1

Comparison of test video sequences before and after packetization

The simulation was executed for three cases of random cell discard, frame-based cell discard as proposed in [SG98] and slice-based cell discard which developed here as a better performing algorithm to address traffic management and congestion control for multimedia transmission over ATM's ABR service. Table 2 gives the results for the measured errors for continuous stream cell packing and leader/trail cell packing mechanisms.

Random cell discard

MSE Average /80 frames	L/T pack (long)	Continuous pack (long)	L/T pack (medium)	Continuous pack (medium)	L/T pack (short)	Continuous pack (short)
Lecture	1107	12856	875	9925	424	7548
Icehouse	1257	13168	948	10218	490	7752
Indiana Jones	1381	13737	790	10142	477	7601
Speed	1294	12951	618	10084	465	7587

Table 2

The MSE value for randomly cell-discarded video sequences

The table 2 reveals, leader/trailer cell packing mechanism, a major improvement in MSE values over the continuous packing mechanism is also observed. Since the leader/trailer packing mechanism localizes the loss of cells to an individual block size in a frame, it is not surprising to see this mechanism has a better performance over the untrained continuous cell packing scheme whose single cell loss may end up with the corruption of entire slice or even several slices.

The performance evaluation of each algorithm was also investigated subjectively and depicted in Figure 9 for the blocks of 16×16. As clearly the Figure demonstrates a superior video quality with respect to untrained cell packing algorithms. While I-frames are used as the reference frame for other frames in MPEG, any loss or damage to these frames would cause penetration of partial and/or complete loss and damage to any subsequent frames. With the deployment of error concealment technique in conjunction with our algorithm, the image can be reconstructed and reproduced at the decoder with a much better video quality in comparison to untrained method.

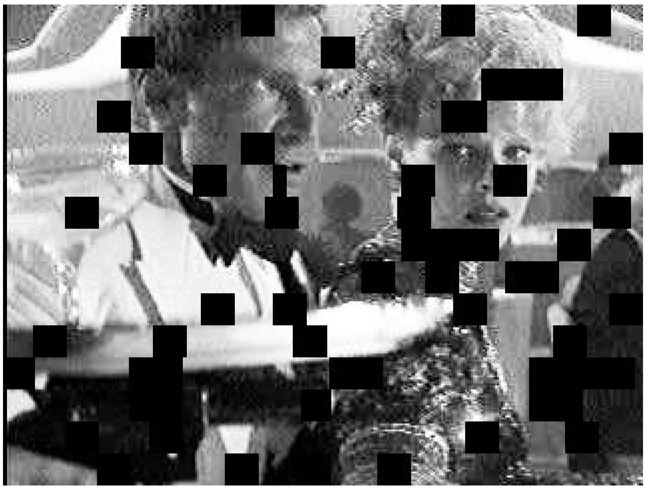

Figure 9a, Header/trailer cell packing

Figure 9b, Continuous cell packing

Conclusion

The distribution of interarrival cells for better congestion control and traffic management of high-speed network such as ATM was investigated here. A new cell packing algorithm for video applications were also introduced as a result, a superior video quality obtained in comparison with untrained cell packing algorithm. To reduce the effect of cell loss on the reconstructed video, a leader/trailer cell packing mechanism was proposed in which the start of every block in all slices required to be packed into a new cell and the flag of one bit field was set to zero to indicate that this is a leader cell and carries the header of block. The rest of the block was placed in the subsequent cells called trailer cells with their one bit flag field set to one. This packing mechanism facilitates to limit the impact of cell loss to individual blocks, unlike continuous cell packing technique in where the corruption or dismissal of a single cell generally has dramatic results in corrupting a whole slice or more.

References

It will be provided in the camera ready version of the paper.