CHAPTER-II

METHODS ON MANAGING STREAMING TRAFFIC AND QoS

2.1 Introduction

Traditional web applications, differing client based access patterns, computational complexities and bandwidth overhead associated with streaming request have increased traffic intensity on the Internet. IP networks require guarantee of Quality of Service parameters for transmission of real time media data and web traffic. Media transmitted over IP networks require highly reliable QoS since protocols such as RTP [102], RTCP [103], TCP [36] and UDP [25] with qualifying techniques such as error checking, packet acknowledgment, re-transmission for loss and sequencing of media packets does not support in providing QoS.

This chapter discusses on the nature of traffic network, their limitations, and related schemes adopted for providing QoS. A detailed study of network traffic and various congestion-based situations are analyzed. An analysis of congestion avoidance schemes like RED [99], BLUE [67], Drop-Tail [57], and queue management schemes are discussed. DiffServ [74], IntServ [90] are two most important schemes used to manage real-time traffic. These schemes are also discussed.

2.2 Network Traffic Nature and Characteristics

To study and understand the scalable nature of media traffic over large domain IP networks, an experimental setup is proposed. The setup is tested for short time-interval of a call session and also for long time duration of an audio play-out. Traffic management schemes and its scaling methods are studied.
2.2.1 Survey on Nature of Network Traffic

Real-time traffic is found to be bursty and multi-scalable in nature. Network traffic is studied from the huge variations in time-range of traffic and its impressive range of bandwidth, from kilobytes to terabytes per second over large backbone of network. The Internet carries all types of traffic, where each type has different characteristics and requirements.

Fig 2.1 shows the time invariant nature of tele-traffic obtained over an IP-Telephone setup. This setup is used to analyze an end-to-end delay of media transfer over a call session. To study, analyze the facts and facets of a media call session over a long distance Internet setup, an "active probe" method is equipped. The call session is a conference call interconnecting the locations in Coimbatore (INDIA) with Denver (USA) using Dishnet Hubs and local ISPs. Two experiments were conducted on 7th Jan 2002 at 10:25am IST and 1:30pm IST respectively for identifying the delay time. The first experiment is conducted for short range series of one call duration while the second experiment is for long call duration of time-interval.

![Fig-2.1 Average Call Delay Time](image)
1. The Delay and loss measurements are collected by means of probes sent over the transfer path.

2. Subjective quality measures, which cause various impairments such as Play-out Delay, End-to-end Delay, and Packet Loss are being noted.

3. Isolated packet loss was considered.

4. The test was carried out on Internet backbones, since long distance calls carry delay.

5. Call serviced over IP network and switched telephone networks carry additional delay.

6. High traffic carrying paths are considered, which carry added delay.

From the setup, it was found that in Internet telephony 72% of packets reach the receiver taking an average of 250ms. If enough bandwidth is available, best-effort service fulfills all of these requirements, but when resources are not available, real-time traffic suffers from congestion. The solution for streaming media over IP is to classify all traffic, allocate priority for different applications and make reservations. A snap-shot (Fig-2.1) of the graph shows an average delay of 270ms for a single call to be received during time-interval from the start of a call to end of a call of ten seconds duration. The spikes in the graph show the bursty TCP based media packets gathered from the local server in Coimbatore. Increased burstiness result in lower levels of resource utilization leading to lower quality of service. In media streaming networks, quality refers to available bandwidth, transfer delay, and packet loss. The impact of scale invariance leads to need for packet scheduling methods, traffic congestion control as well to policies for fairness and pricing.
Fig-2.2 shows an increase in delay for varying traffic intensities gathered for the time duration of 900 seconds, the network exhibits delay of an average of 200ms. While for long time intervals due to queuing delay and other network delay, the time taken to receive the packet increases by an average of 200ms to 250 ms. Hence for long time interval the current network setup does not provide the expected quality.

2.2.2 Scaling Network Traffic

Network Traffic has been analyzed and studied earlier [1,2] under consistent improvement by numerous research works. The complexity [27] and richness [34] of packet – switched telecommunications traffic attribute to the “bursty” nature [34] and scaling behaviour of Internet traffic. Guerin and Peris [47] has identified that any increase in burstiness lead to lack of resource utilization for a fixed quality of service hence increasing cost of time and resources. Any increase in media traffic requires buffer to be allocated based on demand of incoming media.
The key motivation of traffic scaling is, it helps in investigating the type of traffic nature as well the root causes of traffic generated. The philosophy of scale invariance phenomenon can be studied using time series $\phi(t)$, where $\phi$ is the link capacity of the network over the specified time 't'. Generally in terms of network performance, variability is always an undesirable feature of traffic data. Patrice Abry [1] had studied the nature of media traffic using the method of multi-scaling. The results prove that media traffic is always bursty in nature, and stochastic. Minoli Daniel [34] and Claffy [27] proved that increased burstiness results in lower resource utilization for a low quality of service.

2.3 Traffic Policing Methods

Streaming over the Internet is characterized by Best-Effort service [16], since the Internet traffic management protocols (TCP and UDP) provide no hard-time guarantees for timely delivery of data, hence loss-less schemes [65] are more difficult to adapt to quality. Increase in congestion may deplete the buffer and the playback of the media packet sequence must be stopped for some time to allow more packets to be added into buffer. In order to reduce the risk of buffer underflow, an adaptive scheme is necessary. Adaptive nature of traffic can be incorporated by policies for fairness or pricing. Integrated Service Architecture (IntServ) and Differentiated Service Architecture (DiffServ) incorporate policing methods to control traffic and deliver QoS.

2.3.1 Integrated Services Approach (IntServ)

A simplest approach for introducing quality of service in the Internet is based on the Integrated Services (IntServ) Architecture [100,102]. The IntServ
architecture provides guaranteed service using per-flow reservations, limiting its scalability. Since this thesis work concentrates on delivering quality, IntServ and DiffServ architecture models are discussed in depth. Extensive analyses show that the issue of scalability concerning a large domain network deployment was not effective. IntServ provides an end-to-end QoS solution, by way of end-to-end signaling, state-maintenance (for each RSVP-flow and reservation), and admission control for each network element. Internet community proposed the IntServ for providing deterministic service guarantee for bandwidth demanding applications.

The scalability problem in IntServ arises mainly because networks manage reservations for each session individually and as well for per-flow scheduling at links. The key design principle in IntServ is to avoid any virtual circuits and proceed with a connection-less approach of current IP. The per-flow mechanism in IntServ does not scale to backbone networks where there are multiple number of flows to be monitored. To address this problem, George and Illia [44] have proposed a scheme where flows are aggregated and handled collectively, so that state size and processing power in routers are reduced. In the aggregated version of RSVP [18], per-flow routing state is replaced by per source-destination routing state. Thus IntServ method provides a significant improvement in per-flow approach, but the problem is that if flows are established only along shortest paths to enable flow aggregation, bandwidth will be consumed on these paths, leading to high call-blocking rates.
2.3.2 Differentiated Services Architecture (DiffServ)

In Differentiated Services Architecture [49,74] scalability is achieved by reducing complexity, maintaining minimum per-flow state information in core routers and pushing unavoidable complexity to the network edges. Therefore, individual flows belonging to the same service are aggregated inside the network, thereby eliminating the need for complex classification or managing state information per flow in interior routers.

To facilitate an effective end-to-end QoS on an IP-network, the Internet Engineering Task Force (IETF) has defined Differentiated Services as an alternative to IntServ. IntServ follows the signaled-QoS model, where the end-hosts signal their QoS need to the network. DiffServ works on the provisioned-QoS model where network elements are set up to service multiple classes of traffic, with varying QoS requirements. Differentiated Services provide a simple, scalable, and relatively easy deployment in a predominant best-effort Internet setup. In addition, within differentiated services there is significant emphasis on allowing for meaningful end-to-end services to be provisioned across multiple, separately administered network clouds and on keeping the consequent business models as simple as possible. DiffServ addresses the clear need for relatively simple and coarse methods of categorizing traffic into different classes, also called class of service (CoS), and applying QoS parameters to those classes. To accomplish this, packets are first divided into classes by marking the type of service (ToS) byte in the IP header. A 6-bit bit-pattern (called the Differentiated Services Code Point [DSCP]) in the IPv4 ToS Octet or the IPv6 Traffic Class Octet is used to establish end-to-end QoS setup.
The real-time service will enable IP networks to provide quality of service to multimedia applications. Resource Reservation Protocol (RSVP), together with Real-time Transport Protocol (RTP), Real-Time Control Protocol (RTCP), Real-Time Streaming Protocol (RTSP), provides a working foundation for real-time services. Integrated Services and Differentiated Services allow applications to configure and manage a single infrastructure for multimedia applications and traditional applications. It is a comprehensive approach to provide applications with the type of service needed and acceptable quality.

2.4 Queue Management Methods

The capacity to hold packets in router or a gateway for a specific time-interval is of vital importance in avoiding congestion and minimizing packet loss as well the delay of network including play-out delay. Enhanced queuing methods to route packet over the network are adopted, which ranges from static queuing schemes to dynamic queuing algorithms.

2.4.1 Queuing Methods

Various queuing schemes are employed to manage traffic flow.

a) **First In, First Out (FIFO)** [31] : This is the simplest queuing method where outgoing packets are serviced based on incoming packet order. Best Effort Scheme in Internet adopts this method.

b) **Priority Queuing** [11] : This method is an improvement on FIFO by having separate queues for different outgoing priorities. The queue with the highest priority is serviced first, and then the lower-priority queues are serviced in sequence. In some cases, the lower-priority traffic can be denied access to bandwidth entirely if there is insufficient capacity.
c) **Fair Queuing** [11] : This method allocates each session flow the required bandwidth in order to avoid complete starvation of lower-priority queues. The algorithms for fair queuing consume more CPU capacity than those for FIFO or Priority Queuing, but there is no prioritization. This method is a variation on FIFO queuing, with fairness added on bandwidth.

d) **Custom Queuing** [21] : This method reserves a portion of link bandwidth for each selected traffic type, based on user requirement. To configure custom queuing, the network manager must determine beforehand how much queue space in the router buffers to reserve for each traffic type, which is considered as a primary disadvantage. This allocation is fixed and does not change in response to changing network traffic patterns. However, if a particular type of traffic is not using the bandwidth reserved for it, other traffic types can utilize the unused bandwidth.

e) **Weighted Fair Queuing** (WFQ) [10] : This method is similar to fair queuing but explicitly gives the different types of network traffic different priorities. WFQ reduces traffic jitter, which is important for successful multimedia delivery, but it produces more predictable round-trip delays. WFQ separates traffic into high and low-priority classes based upon the bandwidth of the individual flows, instead of some other indicator. Low-bandwidth flows, which are assumed to be interactive traffic, are given high priority. High-bandwidth flows are given a lower priority and interleaved. WFQ improves the response time for interactive traffic.

f) **Class Based Queuing** (CBQ) [19] : This method does not classify packets and all packets are treated equally without any discrimination. CBQ classifies the traffic and categorizes into classes. This type of
definition of service rendered by the network is called "Class Based Service Model". Fig 2.3 explains Custom Queuing.

Real time traffic such as conversational voice, videoconferencing, and real time multimedia require very short latency (typically less than two-tenths of a second end to end, including processing at the end stations) and controlled latency variation (Jitter). Number of methods has been proposed which insist on Congestion Alert [99], Effective Congestion Notification [106], Active Link Management [67].

2.4.2 Random Early Detection (RED)

Dynamic algorithms are adopted where adaptive queuing was required. RED algorithm [99] detects congestion early using active queue management technique (AQM). Floyd suggested this method of detection of packets randomly at an early stage of packet loss and inform to end-hosts, which enables them to reduce their transmission rates. With average queue length, when \( \text{min}(q) \) minimum threshold limit is exceeded then packets are dropped randomly or marked with ECN (Explicit Congestion Notification). If maximum
threshold limit is crossed \( (\text{max}_{th}) \), then all packets are dropped or marked. The disadvantage is that the work relies on queue length as an estimator of congestion and requires wide range of parameters to operate with sufficient buffer space.

According to ECN adopted by RED, the aim is to detect congestion early and to convey information to end-hosts. Introducing infinite buffer space in network, allows queue to grow without bound, resulting in increased end-to-end delay. Since IP packet’s Time To Live limit (TTL) parameter is finite, if the packets coming out of router have crossed the TTL setup, then they are to be dropped and retransmitted by the end-host.

Several researchers have studied on congestion control methods such as Early Random Drop of packets [109], Drop Tail. Hashem discusses on shortcomings of Random Drop and Drop Tail Gateways, where if queue exceeds the DROP level, then gateways drop each packet arriving at the gateway with a fixed drop probability. The work also stresses on the drop level to be adjusted dynamically depending upon network traffic. The Congestion Control Survey [110] considers the versions of Early Random Drop as unsuccessful in controlling misbehaving users. This work does not prove to be successful to solve all the problems of unequal throughput given connections with different roundtrip times and multiple congested gateways. IETF [110] suggests that the goals of Congestion Avoidance in Gateways are "uniform, dynamic treatment of streams or flows of low overhead and of good scalability in large and loaded networks".

2.4.3 BLUE

BLUE algorithm [111] is used to remove the increasing packet loss rates caused by an exponential increase in network traffic. The disadvantage
of RED scheme is that it uses the queue length as the indicator of severity of congestion. BLUE algorithm is modification of RED algorithm that uses a single probability to mark packets and uses link utilization history to manage congestion. If the queue is continually dropping packets due to buffer overflow, BLUE increments the marking probability thus increasing the rate of sending congestion notification. Compared to RED, BLUE algorithm is found to avoid congestion effectively but does not focus on chances of minimizing media packet loss at gateways.

Even though BLUE is considered to be an improvement over RED, Feng and Shin [90] had showed that the results are not significantly better, but SF-BLUE [90], an improved scheme of BLUE shows a better result than RED and BLUE. This queue management scheme can identify the rate-limit non-responsive flows using simple state of information.

2.5 Traffic Management Mechanisms for Streaming Networks

In media streaming network, traffic can be classified into VBR (Variable Bit Rate) or CBR (Constant Bit Rate). Media Streaming falls into VBR category since the inter-arrival of media packets cannot be determined at an instant of time. Generally traffic service is provided, based on the type of traffic, which may be considered as Best Effort Traffic, Bursty Traffic or Bulky Traffic, generated by the type of application.

2.5.1 Traffic Servicing Mechanisms

The behavior of traffic is dependent on traffic pattern generated and servicing nature. Queuing at the gateway introduces traffic, when the queue is full, new packets cannot enter into the network hence to be discarded. If the
media packets wait at the queue for a long time interval then "Out of Time" is signaled by the network and old packets has to be dropped. To support media traffic consistently and reliably, a network must be able to provide packet-forwarding latency methods that does not exceed the maximum tolerable level for 'media' conversation.

Guarantee of QoS between end-to-end network stress on maintaining consistent and aggregate traffic at network level within egress nodes and domain level of a network as well with other intermediate resources such as type of application, bandwidth of network, device memory, operating memory requirements in setup are few. Key methods that provide traffic management include the following schemes:

a. **Traffic Conditioning or Shaping**—Traffic entering a network can be conditioned by using a policy control or shaper. A policy enforces a rate-limit, while a shaper smoothes the traffic flow to a specified rate by use of buffers. Mechanisms such as CAR (Committed Access Rate), GTS (Generic Traffic Shaping), and FRTS (Frame-Relay Traffic Shaping) are components of Traffic Conditioning.

b. **Traffic Classification & Marking**—Packet classification features allow traffic to be partitioned into multiple priority levels or classes of service. Packets can be classified in a variety of different ways—ranging from type of application input to NBAR (Network Based Application Recognition) and arbitrary access-control lists.

c. **Link Efficiency Mechanisms**—Streaming video and voice traffic uses the Real-Time Protocol (RTP). IP, UDP and RTP packet headers can be compressed from approximately 40 bytes down to 5-8 bytes. This saves
tremendous amount of bandwidth in case of low speed links, when supporting large number of media streams.

d. **Link Fragmentation & Interleaving** (LFI) allow for fragmenting large data packets, interleaving them with RTP packets, and maintaining low delay and jitter for media streams.

### 2.5.2 Traffic Models

**The Best Effort Service Model** [56] exerts its best effort to deliver the packets in the stream without committing any quantitative service performance bounds. The best effort service model does not have any formal specifications to be followed where packet delivery should be an expected process. In Internet all packets are treated equally without any discrimination or any explicit delivery of guarantees. In this model, network exerts its “best-effort” to deliver packets injected into it without committing to any quantitative performance QoS bounds. This model adapts to any queuing mechanism method such as FIFO (First-In-First-Out), or Drop-Tail method to handle congestion or manage packet loss.

**Bulk transfer traffic** or Bursty Traffic Nature accepts virtually any network latency, including latencies on the order of a few seconds. Within each traffic category, further subdivide the traffic by priority. Top priority traffic receives preferential treatment because of its importance to the enterprise. Priorities are not substitutes for traffic categories, which are absolute, not relative, requirements. Priorities can be used to differentiate among user groups or to differentiate among applications and users within a group.
2.5.3 **QoS Management In Media Streaming Networks**

The current Internet architecture embodied in Internet Protocol (IP) network protocol offers "Best Effort Service" service model. Generally QoS of media system enforces the following factors:

1. Minimized delay for media play-out at receiver end.
3. Effective Traffic and Congestion control.
4. Providing end-to-end control setup among various multimedia devices and setup components.
5. Media Packet Scheduling / Reducing jitter-delay.

Generally, deterministic behavior of the system refers to the adherence of time spans defined in advance for the manipulation of data, such as meeting the guaranteed response time. Speed and efficiency in delivering the quality are the main characteristics of a QoS real time media streaming networks. Research has been carried out in this area for over a decade and has produced commendable work on congestion management and avoidance schemes such as Drop-Tail [78], RED [99], S-F-BLUE [90], which provide acceptable results on congestion avoidance and traffic management issues. These traditional schemes are applicable only for few range of applications such as web traffic, mail transfer, messaging traffic and data request / retrieval traffic while there are no specific QoS schemes existing for managing media applications and real-time media streaming applications on IP network.

Streaming-media solutions targeted at Internet must embrace the notion of graceful degradation, they must be architectured with the expectation that they operate within a continuum of service levels, adjusting quality-resource trade-offs as necessary to achieve just-in-time requirements.
Packets must be delivered in a timely fashion due to the real-time nature of audio and video stream data. Congestion within the network may delay the reception of packets and cause them to miss presentation deadlines. Several techniques deal with congestion as the media is being transmitted.

2.5.3.1 Study on existing QoS Protocols

Good number of QoS protocols have already been in process for media streaming such as RTCP [103], RSVP [29], MGCP [14] / MEGACO [25], which strive to provide QoS in media streaming networks. ITU has as set of standard media protocols such as Real Time Protocol (RTP) [102], RTCP, Real Time Streaming Protocol (RTSP) [101], which has been developed specifically for IP networks to carry streaming media data. The standard packet format used for continuous media traffic – such as audio, video – on the Internet is RTP, which includes sophisticated algorithms to control and manage the traffic placed on the network. This protocol assumes that audio traffic is adaptable, compared to video traffic, which is "bursty" in nature. Schulzrinne has stated that, "in an audio conference the audio traffic is self-limiting because only one or two people will speak at a time" [102].

This protocol does not provide enough support for conference setup where multiple users are called for a session. The distribution of audio streams in the network for such applications is either one to- many or few-to- many (M-to-N where usually M<<N). Various techniques have been proposed for reducing the bandwidth required for such broadcasts, most notably network multicasting [44]. Due to the nature of these applications, some protocol developers have explicitly disregarded the possibility of multiple simultaneous speakers and have focused on supporting static, controlled,
small scale and slow pace dynamic models of interaction. In an interactive small-scale conferencing (such as some IP telephony applications) more than one participant can transmit audio at the same time.

Hence it is admitted that protocols such as RTCP and RSVP can help routers and several traffic shaping algorithms to control traffic, but guaranteed end-to-end IP QoS is very difficult to achieve, while a better QoS level can be delivered. IETF has recognized this as a problem [30] "A Framework for QoS-based Routing in the Internet". This RFC describes some of the QoS based routing issues and requirements and proposes a framework for QoS-based routing in the Internet. "A Framework for Integrated Services and RSVP over ATM" [29] outlines the issues and framework related to providing IP Integrated Services with RSVP over ATM and IP.

2.5.3.2 Issue of Fault Tolerance in Streaming Media

The tolerance limit of media streaming network is defined, as the extent the media network is able to maintain its performance effectively under abnormal conditions. Normal Condition refers to the condition network can work during packet transmission. Tolerance of voice network depends basically upon packet loss and traffic intensity. Such loss could be caused by discarding packets at will in IP networks due to congestion (eg Random Early Detection[99]) or packets of late arrival or buffer overflow within the network. Even if the Internet eventually supports reservation mechanisms or differentiated services, it is likely to be on per-class than per-flow basis. Thus flows are still expected to perform congestion control within their own class.

1. If any packets are to be dropped out then less priority packets are to be dropped.
2. Priority assignment is done initially when packets are created at sender and checked instantly at intermediate routers.

3. Continuous probing of network and neighboring networks should be done, such that failure of a link or a route through which the packets being transferred can be determined and alternate route is selected.

2.6 Schemes on Managing QoS

Real-time audio communication is bandwidth intensive, and has strict requirements for minimum throughput, network delay and jitter. Internet operates on a best-effort basis and cannot guarantee an upper bound on end-to-end delay or lower bound on available bandwidth. These applications can easily cause network congestion hence packet loss and increased delays that can significantly impair media quality. As such applications become more widespread, large number of streams may form a considerable portion of the Internet load. Therefore, overall behaviour of the applications that include large number of streams has significant impact on the Internet traffic and the quality of delivered audio.

2.6.1 Traffic and Congestion

From the network point of view, large scale deployment of streaming media applications results in severe unfairness against TCP-based traffic, responsive traffic and non-responsive traffic leading to congestion collapse [2,78,99], since the network in general does not manage the utilization of its own resources, such as bandwidth, buffer. The main cause for congestion is non-cooperative aspects of resources sharing.

Unfairness refers to bandwidth starvation that unresponsive flows can inflict on well-behaved responsive flows such as TCP flows. TCP is the
dominant transport protocol in the Internet and the current stability of the Internet is mainly due to TCP’s congestion control, which fairly shares link bandwidth between multiple connections and maintains load at the network on useful levels [57,81]. If the multiple real-time streams introduced to the network do not obey the laws of existing traffic on the Internet, and do not respond to signals of congestion as TCP does, then that can lead to network collapse [99].

Common network resources such as link bandwidth, buffer spaces in routers are finite. Chances may exist for all media packets to arrive at the same router to be forwarded at a single output link, while other routers are not utilized. Various methodologies are being preferred such as FIFO schedulers [1] where the packets are just serviced and forwarded in their order of packet arrival. Problems exist, when the buffer space is exceeded and the router starts dropping packets at will, as algorithm RED. Braden [107] and Floyd [104] refer to other methods of congestion collapse. Braden’s method of congestion collapse is a common method of packet loss due to congestion at gateways. While Floyd’s method of congestion collapse is fragmentation-based congestion in which network transmits fragments of packets that will be discarded at the receiver since they cannot be reassembled into a valid packet.

2.6.2 Packet Loss

Digitized audio is transmitted over a network in a series of packets. In order for the receiver(s) to reproduce the original signal, they need to receive all the packets with preserved timing relationships among them. However the Internet does not guarantee when or whether a packet will be delivered to the
receivers. Buffering of the packets in the network node introduces variable delay and can distort the original timing relationship among the packets. In addition to the Delay, Jitter that result from the variable buffering times when the buffer limit is exceeded, can further impair the audio quality. This is the primary cause of packet loss in the Internet [28].

If the bandwidth requirements of such bursts are lower than the available bandwidth, then all the packets will be delivered to the end user successfully. However, during loaded times when there is insufficient bandwidth for these peaks, media will suffer severe packet loss and delays and may be unusable to end users. The technique used to correct the lost packets do not provide effective solution [28], since addition of large amount of repaired data being lost is detected, increases the network congestion, leading to delay. The problem of error-resilient media communication [51] has received significant attention in recent years and a variety of techniques have been proposed to combat packet losses and increase the robustness of communication. Examples of recent work in this area include the method of Forward Error Correction (FEC) [11], dynamic control of prediction dependency using long-term memory [10,11,28], channel-adaptive packet scheduling [54,55,56]. These techniques manage and optimize the dependency across predictive coded packets All of these schemes improve error resilience at the expense of increased bit-rate.

Prior work on modeling the effect of packet losses, identifies quality distortion being proportional to the number of losses that occur [3]. Achim et al [2] carefully analyzed the distortion for a single (isolated) loss (accounting for error propagation, and spatial filtering), and the effect of multiple losses as the superposition of multiple independent losses. With this linear or additive
model, the expected distortion is proportional to the average packet loss rate. This model is accurate when single losses that occur are spaced sufficiently far apart with respect to the intra refresh period, ie when the packet loss rate is low and the losses are not bursty. In [51] the length of a burst loss was shown to have an important effect on the resulting distortion where longer burst lengths generally led to larger distortions. Effect of burst loss was also identified as an important feature for comparing the relative merits of different error-resilient coding schemes.

2.6.3 Delay

The biggest impediment to widespread use of media streaming applications is delay. This is the latency introduced by voice compression / decompression (CODEC) algorithms used by communicating end devices, and internet/transmission delays introduced within an IP network as it routes packets over an end-to-end path.

The ITU standard G.114 states that end-to-end delay should be no more than 130 milliseconds (ms). However, experience has shown that an end-to-end delay of 200 ms is usually satisfactory for 50% of users. Jitter should not be more than 50 ms.

2.7 Summary

To estimate the lossy distortion, which destroys the quality of media data transmitted through the network, the proposed schemes of the thesis work explicitly considers the effect of different loss patterns, including bursty losses and separated (non-consecutive) losses spaced apart by a lag and also considers inter-frame error propagation and the correlation between error frames.