

Modelling Approach for VoIP Traffic Aggregations for Transferring Tele-traffic Trunks in a QoS enabled IP-Backbone Environment

Jörn Seger

Faculty for Electrical Engineering and Information Technology
Department of Electronic Systems and Switching
University of Dortmund
joern.seger@uni-dortmund.de

Abstract

This paper gives a theoretical approach to model an aggregated flow of VoIP connections on packet basis which is created due to tele-traffic parameters.

There have been many attempts to model a single voice over IP stream and their aggregation and there are also well explored theories on circuit switched telephone networks. Here we want to expose a way, how to combine these approaches to have a theoretical model for simulating e.g. effects of additional voice over IP traffic on a backbone network.

In this paper we will examine different voice codecs respecting the packet flow. To respect Voice Activity Detection (VAD) or Silence Detection (SD) in modern voice codecs like G729 Annex B, G.723.1 Annex A and NeVoT, spurts and gaps have to be described mathematically as well. The aggregation of flows will be modelled on statistical tele-traffic data, that are globally known or measured for the link, which should be transferred from a circuit switched to a packet switched network.

To get a realistic traffic flow on the link, the flow aggregation must be modelled due to the Service Level Agreement (SLA) between the customer and the Internet Service Provider (ISP) e.g. to simulate violations (out of profile packets) due to the given SLA.

1 Introduction

Within the InterMON scenario, it is necessary to have an estimation of voice over IP traffic within an ISP-backbone network. For example: Assume, a company, which would like to route its voice traffic, between two locations. Formerly this was done via leased line. Now this company wants to transfer their traffic as Voice over IP through an ISP network. Within this scenario, an ISP may like to sim-

ulate, whether it is able to offer this service immediately or if it needs some traffic engineering efforts to provide it.

The approach, presented in this paper will be able to create a voice over IP traffic stream with the knowledge of tele-traffic theory. This traffic is mostly given by statistical estimations or measurements.

There have been many efforts to model calls within a telecommunication network on statistical data and also many ideas, how to model voice over IP traffic within a packet switched network. The aim of this paper is to summarise these approaches and to combine them in the form shown in figure 1.

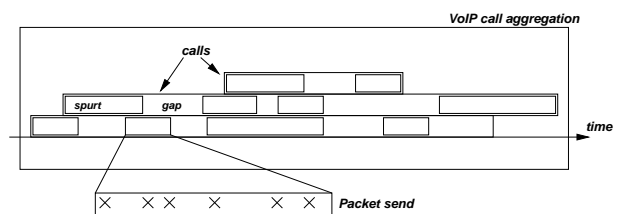


Figure 1. Voice over IP aggregation modelling

To follow this idea, it is necessary to divide the problem of packet flow modulation into three sub-models:

1. model for constant bit rate voice over IP streams
2. model for talk-spurts and silence gaps
3. model for traffic circuits on a link

2 Simple Voice over IP Traffic

Packet streams in voice over IP connections are almost coded in constant bit rates. Therefore packet interval and packet size will be constant. What has to be considered is,

that different codecs produce different intervals and different packet sizes, even if the effective bandwidth is the same.

For example a PCM-codec (G.711) produces a 64kBit stream. Every packet may at least contain 8 Bit, which is the amount of bits for one sample and must then be sent every 125 μ sec to get a bitstream of 64kbit/sec (like it is in ISDN).

The network overhead (IP+UDP+RTP) for a packet with 8 bit (one byte) payload is about 40 byte. Furthermore in this scenario 8000 packets must be transmitted within one second. Because of this inconvenient instance, more than one sample is transferred in one single packet (normally from 30 to 160). The exact amount depends on the used codec. This means, every stream will send 22 to 100 packets per second. To solve the problem of huge headers comparing to the payload, header compression (e.g. cRTP) is introduced. It will compress a 40Byte header to 2 or 4 bytes. But this compression is not widely used, because of incompatible hardware.

What should be considered, too, is that samples will be delayed, while they are collected during waiting for transmission. This period is called endogenous delay, which is constant for every voice over IP stream and could only be shorted on the expense of shorter packet interval and therefore a bigger packet overhead.

Table 1 shows some example codecs information like bandwidth, how many packets will be sent in one second and how many samples will be transferred within one packet. The important data for modelling a bitstream are the interval, which is the time between two packets and the packet size, with and without header compression. The last field reflects the probability, that this codec is used on this connection.

Codec	voice BW $\frac{kBit}{sec}$	samples per packet	interval time (msec)	raw packet- size	compressed packetsize	probability of usage
G.711	64	80	20	120	82	0.6
G.729	8	20	20	60	22	0.3
G.723.1	6.3	30	38.46	70	32	0.1

Table 1. Example list of different codecs and their properties

On modelling a stream, first of all the codec will be chosen due to the probability of usage. After that, a packet stream has to be modelled due to the probabilities of the codec and the modelling in this stage has been finished.

Additionally remarkable on our approach is, that even traffic that follows a variable bit rate, could be modeled with only slide differences in producing packets.

3 Voice Activity Detection (VAD)

More recent voice codecs like G.729B, G.723.1A and NeVoT have the ability to detect talk-spurts and silence gaps within a conversation [5],[6]. During the silence time, the codec stops to transfer data, to save bandwidth.

To model this behaviour, the traditional approaches like [4] assume that the length of spurts and gaps follow an exponential distribution. However, in [1] it is shown, that this assumption does not work any more. In most cases, the spurt is slightly more "heavy-tailed" than exponential, whereas the gap distribution deviates strongly from an exponential model.

In [2], the distributions of different voice samples are tested with the Kolmogorov - Smirnov (KS) goodness-of-fit test against different distributions. Due to this test, the authors stated, that the Gamma distribution fits better for the spurt (ON-)period and the Weibull distribution fits better for the gap (OFF-)period.

In our approach, the stream, that was created in section 2 will be suppressed during the silent OFF-period, whereas the stream passes through during the ON-period (spurt time).

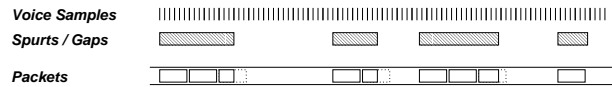


Figure 2. Packet creation based on voice activity detection

4 Aggregated Flows on real Links

To simulate flows on a link, tele-traffic theory seems to be the most promising approach, because the theory of calculating tele-traffic circuits is well known and could easily be adjusted to our approach ([7]). Most telephone equipment even integrate measurement tools, like investigation of average hold time and average inter arrival time.

Roughly spoken, we will use tele-traffic theory to determine the call inter arrival time and the call hold time for a link. The main difference between this packet based approach within quality of service enabled diffServ networks and a circuit switched network is the non blocking character of a link. There is no "natural" detection for the caller that there are no resources left on a link, because the bandwidth will be divided among all connection flows.

4.1 Tele-Traffic Theory

As mentioned before, there are two different sub-models within the tele-traffic model:

- inter arrival time model
- call holding time model

To simulate the raised calls on a link, we will use discrete event simulation. That means, the change of the state process will be determined via calculation of the next arrival or departure time (as shown in figure 3)

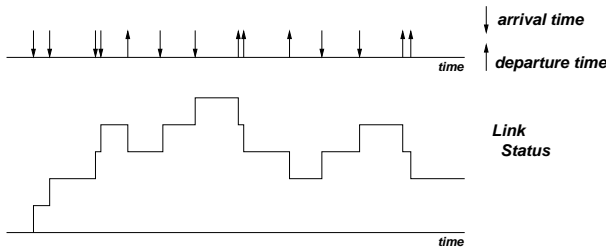


Figure 3. Simulation of arrival and departure time

To find a mathematical model to describe inter arrival and holding time of a call, the traffic must be analysed in more detail:

Call Holding Time

Almost all holding times(b) are distributed exponentially, so we could assume

$$P(b > t) = e^{-\frac{t}{t_m}}$$

as the probability that a call will end within the time t . t_m defines the Average Holding Time (AHT). To calculate this AHT from measurements, traces should be taken on a 15 minutes or 60 minutes basis (ITU recommendation). Within this period, the AHT will be calculated as follows

$$AHT = \frac{\text{total call time}}{\text{amount of calls}}$$

The AHT might be dependable on the time of day and is assumed not to be dependable on the state and the arrival process.

[3] stated, that an acceptable AHT is generally assumed to be 180 to 210 seconds in many business environments.

Inter Arrival Time

The inter arrival time (IAT) is generally assumed to be exponentially distributed. This assumption will work in many environments. However in scenarios where sources are limited and a random distribution could not be implied,

the tele-traffic must be measured to find different distributions. It is argued in [3] that for example an outbound tele-marketing company, has an inter arrival time that will follow a hypo-exponential distribution.

Therefore inter-arrival time is dependable on external influences and must be adjusted to the task.

Difference of packet simulation

In difference to usual traffic theory, this model does not only want to calculate the busy hour traffic, to estimate a fixed amount of switched circuits. The advantage of our approach is that the packet simulation could be integrated into a major environment to forecast the network load. Another advantage is that this traffic can change through time due to the statistical information of the communication data that should traverse the network. Table 2 shows the average inter arrival time between two calls (call arrival is assumed to be exponentially distributed) and the average holding time, that is sampled within one day.

Hour	9:00	10:00	11:00	12:00	13:00	15:00	...
AIT (in sec)	5.65	3.64	2.32	8.84	9.21	6.88	...
AHT (in sec)	325.2	345.4	123.74	165.54	322.54	79.86	...

Table 2. Average interval time between two calls and average holding time

4.2 Combination

To simulate the whole packet flow, the process starts with the calculation of the next arrival time of a packet. When this time is reached, several things will be triggered immediately:

- tele-traffic
 - call holding time (T_H) will be determined due to the given distribution.
 - time until next call (inter arrival time, T_A) will be determined due to the given distribution
- Constant bit rate packets
 - determination of codec, due to the "used codec application probability"
 - start of packet generation with interval T_I
- spurts and gaps
 - spurt time (T_{sp}) will be determined due to the Gamma distribution
 - gap time (T_{gap}) will be determined due to the Weibull distribution

- every new determined spurt and gap will be concatenated to a "voice stream" until the call ends (due to call hold time)

Figure 4 shows the time events during one call. This proceeding of discret time simulation allows on- and offline production of packets. In an offline scenario, the packet will be stored within a database with an additional timestamp. In an online scenario, the packets could be produced, when the next packet creation event arises. In this scenario, the time has not necessarily to be "real time".

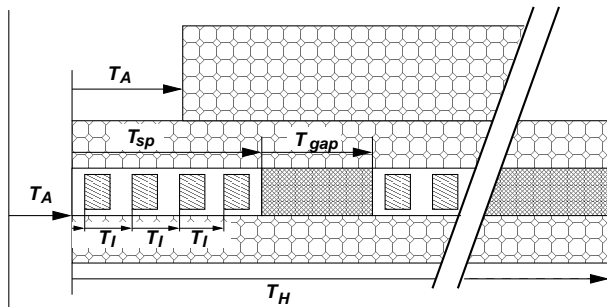


Figure 4. Timing diagram and its connections to packet creation

4.3 Aggregation on SLA-Definition

ISPs normally define Service Level Agreements (SLA) on a connection between the ISP-backbone network and the customer. This agreement should restrict a specified service to both, the costumer and the provider. Mainly this restriction is defined, by delay, packet loss, bandwidth and the borderrouter queueing system.

In this scenario, packet loss and delay is mainly out of scope. Interesting for simulation is the bandwidth and the queueing system and as a result the probability of out-of-profile packets.

E.g. if an ISP provides a connection with a fixed rate, the queueing system will probably be FIFO with size L.

A more flexible way, to provide especially a VoIP connection class is via token bucket filter. This bucket is filled with tokens within a fixed rate (usual depending on the bandwidth). Every token represents a defined size and every packet, that needs to be sent will consume as much tokens as correspondent to its size. If there is no token available, the packet has to be queued and will be sent, when there are enough token in the bucket again.

To estimate the probability p_0 of out-of-profile packets, [1] proposed a leaky bucket, which will be equivalent to the token bucket approach, except the queueing delay.

5 Further research

The next step in this approach is the implementation. We will integrate the modulation module into the NS-2[?] environment to be able to use it within the InterMON scenario and after that, the model has to be tested against real systems and may be changed in some specification details.

If this model has approved reliable, further work will be the examination of the results to be able to simplify the modelling due to some unimportant factors. Also the Grade of Service (GoS) has to be integrated into this approach, to have a more qualitative analysis for the VoIP-stream.

6 Conclusion

On real time communication, the changeover from circuit switched to packet switched networks has just started. Many providers are at an early state of integration of real time traffic into their backbones. Furthermore there is enough bandwidth available at the moment, so congestions control and traffic engineering is not the most important task today.

But in future, competition will force ISPs to lease as few lines as possible and to use them optimal. To do so, simulations are very helpful to determine future capacities and to make best offers on a very fast market.

The approach, presented in this paper is one step to achieve this simulation and it might bring more efficiency to the IP-Backbone network.

References

- [1] Wenyu Jiang, Henning Schulzrinne; *Analysis of On-Off Patterns in VoIP and Their Effect on Voice Traffic Aggregation*. Department of Computer Science, Columbia University 2000
- [2] Boris Bellalta, Miquel Oliver, David Rincon; *Capacity and Traffic Analysis of Voice Services over GPRS Mobile Networks*. Technical University of Catalonia, University Pompeu Fabra, Spain 2002
- [3] Cisco Document Server; *Traffic Analysis for Voice over IP*; Posted Sept 2002 <http://www.cisco.com/univercd/cc/td/doc/cisintwk/intsolns/voipsol/ta_isd.pdf>
- [4] Paul T. Brady; *A model for generating on-off speech patterns in two way conversation*; Bell System Technical Journal; 48(9):2445-2472; September 1969
- [5] International Telecommunication Union; *Coding of speech at 8 kbit/s using conjugate-structure algebraic-code-excited linear-prediction annex b: A silence compression scheme for g.792 optimized for terminals conforming to recommendation v.70. Recommendation G.729B*, Telecommunication Standardization Sector of ITU, Geneva, Switzerland, November 1996
- [6] Henning Schulzrinne; *Voice communication across the Internet: A network voice terminal*; Technical Report TR 92-50; Dep. of Computer Science; University of Massachusetts; Amherst; Massachusetts; July 1992
- [7] Winfried Schuberth, *Verkehrstheorie elektronischer Kommunikationssysteme*; 1. Auflage; ITT Austria GmbH; Wien; 1982