Analytical Model for Voice over IP traffic characterization

A.E. GARCÍA¹, K. D. HACKBARTH ¹, A. BRAND², R. LEHNERT²
¹ Telematic Engineering Group of Communication Engineering Dept. ² Institut für Nachrichtentechnik
University of Cantabria
Av. los Castros s/n, 39005 Santander (Cantabria), Tel: +34 42-201549, Fax: +34 42-201488
SPAIN
agarcia@tlmat.unican.es  http://www.tlmat.unican.es

Abstract: - The scope of this paper is to derive a set of simple formulas to dimension virtual paths in IP/ATM packet networks that carry compressed voice traffic. The project further gives an overview of voice over packet technology with a special focus on voice over IP and its most prominent standard ITU-T H.323. After introducing a network model for voice over packet networks, the dimensioning formulas for the communication application are derived for two different levels: call and burst level. The obtained formulas for packetized voice services are validated with the use of the simulation software OPNET.

Key-Words: - Voice over IP, Internet telephony, Traffic modeling, network simulation.

1   Introduction

Voice transmission has more than enough networks typically based on packets like for example IP, ATM or Frame-Relay are winning positions regarding traditional circuit switching networks as PSTN and ISDN.

Basically we can consider two reasons that they have taken to the necessity of carrying out the integration of voice in packet networks. On one hand reduction of the costs comes given by the best use in bandwidth that is made in these networks. This fact is translated in the case of voice in code outlines and efficient voice compression, silences compression, and the possibility of carrying out statistical multiplex.

On the other hand new services appears, specially based on IP, independent to the network technologies and user's platforms, see [1]. We are this way scenarios of such services as multimedia data over single line, videoconference, multicasting, CCA's (Call Center App.), unification of messages, etc.

Crucial points so that companies accept to use of IP voice services it has more than enough they are fundamentally their readiness, qualities comparable to the public service of voice, the reliability, and their manageability [2]. This way, client of business waits some concrete parameters of quality of service [3], at least comparable to that of traditional services.

2   Voice over IP

For technologies of such compressed voice as voice over IP (VoIP), voice over ATM (VoATM) and voice over Frame-Relay (VoFR), different standard exist.

2.1 ITU-T H.323

Denominated "Systems for multimedia communications based on packets", it is part of the family of standard H.32x that specify the services of multimedia communication on different networks, as ISDN (H.320) or BB-ISDN (H.231 and H310).

H.323 is not specific of IP, so that it can be used by IPX or AppleTalk. However, it makes use of IETF standards: RTP (Real Time Protocol) and RTCP (Real Time Control Protocol).

It describes all the entities of a system H.323:

- Terminals: provide full-duplex connections in real time, point to point and point-multipoint, being able to be computers or special devices as videophones.
- Gateway: allows the interconnection with other terminal (H, PSTN or ISDN).
- GateKeeper: provides the admission control and addresses translation services on terminals, gateways and MCUs.
- Multipoint Control Unit (MCU): gives support for multipoint conferences.

3   Network and network layer model

To carry out the modeling of the network we can leave of a scenario where a number of users of a corporate network can establish voice calls by means of the connection through a VoIP gateway and a supplier of VoIP service. Gateway includes a IP router that guides the calls carried out in Internet toward a VoIP gateway of the PSTN network in which the destination telephone is located.

Voice is compressed in client's part using G.729 recommendation with silences compression, what
means that packets will only be transmitted to the gateway and Internet when voice activity is detected. During voice activity a fixed number of audio samples is either packed in RTP, UDP or IP, taking a size of deterministic packets.

Protocols stack would be the corresponding to H323 standard:

<table>
<thead>
<tr>
<th>5: Session</th>
<th>H.323</th>
</tr>
</thead>
<tbody>
<tr>
<td>4: Transport</td>
<td>RTP/UDP</td>
</tr>
<tr>
<td>3: Network</td>
<td>IP</td>
</tr>
<tr>
<td>2: Link</td>
<td>(MLPPP/FR/ATM-AAL 1)</td>
</tr>
<tr>
<td>1: Physical</td>
<td>(SONET)</td>
</tr>
</tbody>
</table>

Table 1: Stack for H.323 overIP terminal

4. Analytic models for voice services

In general, analytic models can be divided in three different scale levels: call level, burst level and cell/packet level. In this study call level will be used to determine the number of virtual connections that they are necessary under a probability of given loss. In the burst level bandwidth will be calculated for each one of these connections, being obtained the width of equivalent band for the voice transport through a virtual channel VBR. If this bandwidth was multiplied by the number of established virtual channels at the call level, it will be a good estimate of the width of global band. This estimation will be validated via the simulation in the next chapter.

4.1. Call level

The VoIP gateway described in the previous chapter can be characterized as a pure loss system because calls will only be accepted as long as there are free ports for the calls available. Thus a traffic model would look like the following:

![Traffic model for call level](image)

Fig. 1: Traffic model for call level

Origin of calls is characterized random when in a very short time interval $\Delta t \rightarrow 0$ three conditions are true:

- **Stationarity**: probability for one call in this time interval tends to $\lambda \Delta t$ independent of $t$ with $\lambda =$ const.
- **Single arrivals**: probability that more than one call arrives in a very small time interval tends to zero i.e. no batch arrivals.
- **Call independence**: each call originates independent from other calls

Assumption of random call arrivals is reasonable if the number of potential callers is large. With this assumption call arrival process with the offered traffic $A = \lambda \Delta t$ is poissanian.

The number of servers (=ports) $s$ is finite, therefore call losses can occur and offered traffic $A$ is the sum of carried traffic $Y$ and lost traffic $R$. The calls shall be served first-in first-out (FIFO).

If also call termination is considered as random, service time can be assumed exponentially distributed, which agrees fairly well with actual telephone conversation times [1].

With that given description of the call level traffic system it can be classified using the Kendall notation as a M/M/s(0), that is Poissonian input process, exponential service (call holding) times, $s$ servers and no waiting room. Traffic type in this system is first-order random traffic also known as Erlang-traffic.

We are only interested in the steady state of the system, so we can derive formulas from global balances of the steady state transition diagram of the M/M/s(0) system:

![M/M/s(0) steady state transition diagram](image)

Fig. 2: M/M/s(0) steady state transition diagram

Erlang formula will be used to determine the number of virtual channels under a given loss probability. This loss probability $P_r$ is equal to the blocking probability $P_b = P(\text{all connections are busy}) = P(i=s)$ because of PASTA property (=Poisson Arrivals See Time Average) of Poisson streams [4] which leads to the Erlang B formula. For the determination of the number of lines $r$ needed for a given loss probability recurrence formula is used iteratively.

4.2. Burst level

4.2.1. On-Off voice model

Multimedia traffic is very bursty in nature and simple models as the Poisson process do not capture the important characteristics of these sources. To model bursty traffic sources different approaches are available many of them using Markov modulated processes (MMP). These are doubly stochastic processes in which each state of $N$ states of embedded markov chain originates another stochastic process. If this originated process is a Poisson process the MMP is called Markov modulated Poisson
process (MMPP), if it is deterministic it is a Markov modulated deterministic process (MMDP).

One special case that is very suitable to model voice sources is the On-Off model. This is a MMDP with only 2 states as shown in the following figure:

**Fig. 3:** On-Off model

According to the On-Off model signal stream of a single source is modeled through an alternating sequence of burst and silence periods. Duration times of burst and silent periods are exponentially distributed with mean $t_{\text{ON}}=1/\alpha$ and $t_{\text{OFF}}=1/\beta$, respectively.

During the burst period packets with fixed length are generated with constant interarrival time $t$. Some studies have proposed $t_{\text{ON}}=0.4$ sec and $t_{\text{OFF}}=0.6$ sec [5], setting the transition rates to $\alpha=1.6$ and $\beta=2.5$.

Steady state probabilities can easily be derived from balance equations of above graph:

$$p_{\text{OFF}}=\frac{\beta}{\alpha+\beta}=0.6 \quad \text{and} \quad p_{\text{ON}}=\frac{\alpha}{\alpha+\beta}=0.4.$$  

In the active state packets are generated with constant speed $v_p=L/t$ (p-peak rate) with $L$ as the packet length and $t$ as the constant packet interarrival time. This is also depicted in the next figure:

**Fig. 4:** On-Off voice packetization

Each generated voice packet with two G.729 audio frames has a fixed length of 20 bytes audio + 40 bytes overhead = 60 bytes, generated every 20 msec. So the peak bit rate of voice source in active state is $v_p=24000$ bps or $v_{pk}=50$ packets/sec. Average data rate is $E(v)=v_p\cdot p_{\text{ON}}=9600$ bps and the standard deviation.

$$\sigma(v)=v_p\times\sqrt{p_{\text{ON}}}\cdot(1-p_{\text{ON}})=11758 \text{ bps}$$

4.2.2. Aggregating On-Off sources

The multiplexing of $N$ independent On-Off sources is outlined in the figure 5. Each source generates $v_p$ bits/sec in it’s active state. Aggregated data rate would be $N\cdot v_p$ in the CBR source case.

**Fig. 5:** Multiplexing of On-Off sources

Because of silence compression, one can profit from statistical multiplexing gain so that the needed server capacity $C$ will be $C=\varepsilon\cdot v_p$ with $\varepsilon$ as the number of equivalent CBR sources.

If $k$ independent voice sources are active with probability $p_{\text{ON}}$, $(N-k)$ sources are inactive with $p_{\text{OFF}}=(1-p_{\text{ON}})$. Further there are $(N$ over $k)$ possibilities for choosing $k$ elements out of $N$. Thus the probability of $k$ active sources follows a binomial distribution:

$$p(k)=\binom{N}{k}\cdot \left(1-p_{\text{ON}}\right)^{N-k} \cdot p_{\text{ON}}^k = \binom{N}{k}\cdot \left(1+\frac{\alpha}{\beta}\right)^k \cdot \left(\frac{\beta}{\alpha+\beta}\right)^{N-k}.$$  

In the mean $E(k)=N\cdot p_{\text{ON}}$ sources are active so the average data rate generated by $N$ sources is $E(v)=v_p\cdot E(k)=v_p\cdot N\cdot p_{\text{ON}}$. Therefore $\varepsilon$ should satisfy the condition $N \geq \varepsilon > E(k)=N\cdot p_{\text{ON}}$.

If server capacity is less than potential maximum source data rate an overload condition can appear. Under this condition server buffer will be filled and if no more buffer space is available packet loss occurs.

This can also be derived from Markov chain of $M(N)/M/N$ binomial source model. If the server provides capacity for $\varepsilon$ sources with $S_u \leq \varepsilon < S_o$ all states from $S_o$ to $N$ cause overload.

**Fig. 6:** $M(N)/M/N$ state transition diagram

Overload probability is calculated as the probability that they are more voice sources active than capacity is available:

$$P_d = P(k > C) = \sum_{i=C+1}^{N} \binom{N}{i} \cdot p_{\text{ON}}^i \cdot (1-p_{\text{ON}})^{N-i} +$$  

with $C$ as the server (=backbone) capacity in number of sources, $k$ as the number of active sources and $N$ as the total number of sources. Note that for pure loss systems without buffers overload probability becomes loss probability of the system. In the following we will assume a pure loss system.

To take into account loss probability as a function of $p_{\text{ON}}$ and the number of sources $N$, server capacity
must be augmented by a correction value $\gamma(P_B, p_{ON}, N)$. This $\gamma$ can be interpreted as a multiplier of standard deviation of aggregated data stream. Corresponding formulas are derived below:

$$C = (1 + \gamma(P_p, N, P_{ON}) + \gamma(N)) \cdot \left(1 - P_{ON}\right) \cdot \left(1 - P_{ON}\right)$$

$$N_{eq} = \frac{C}{\gamma(N)} = \frac{N}{N} = p_{on} + \gamma(N) \cdot \left(1 - P_{ON}\right)$$

(6)

$$v_{eq} = C_{eq} = \frac{C}{\gamma(N)}$$

respectively, total capacity of N aggregated binomial sources in bps, total number of equivalent circuits for N sources in $v_p$ units, equivalent capacity for one binomial source in $v_p$ units, and limiting behaviour of the equivalent capacity $v_{eq}$.

For a great number of sources equivalent capacity of one source converges against its active probability $p_{ON}=40\%$ especially for great loss probabilities that go along with small $\gamma$ values. As an example for 1000 sources under $P_B=0.1$ needed number of equivalent circuits is 420 which gives us a equivalent capacity $v_{eq}$ of 42%. Theoretically $v_{eq}$ slowly converges against $p_{ON}$ with $N/\sqrt{N}$.

4.2.3. Poisson source with waiting system model M/M/1(m)

In the first analytic model a pure loss system was presented which gave an expression to estimate the needed bandwidth of a packet voice backbone. Now we introduce a system buffer of size m to estimate further QoS parameters such as mean waiting time and mean queue length. This will allow a comparison with the values obtained by simulation.

To model the bursty traffic of On-Off sources in the waiting system, a simplification of the source behaviour will be done. The On-Off source in the previous chapter was modeled with exponentially distributed active and silence periods with the mean values $t_{ON}$ and $t_{OFF}$ respectively.

Now the burst length $t_{ON}$ and the burst interarrival time $t_{ON}+t_{OFF}$ will be considered as exponentially distributed. This allows approximating the system as M/M/1(m) and avoids complicated models such as MMPP/E_k/1(m), which are used in literature, see e.g. [4], [5].

The addition of two exponential distributions with mean $t_{ON}$ and $t_{OFF}$ and variances $t_{ON}^2$ and $t_{OFF}^2$ leads to a hypoexponential distribution with the mean value $t_{ON}+t_{OFF}$ and the variance $t_{ON}^2+t_{OFF}^2$.

$$\lim_{N \to \infty} \frac{1}{N} = \lim_{N \to \infty} \left(\frac{1}{N} \cdot \left(1 - P_{ON}\right)\right) = p_{ON}$$

4.2.4. Aggregating the modified On-Off sources

Multiplexing of N modified On-Off sources leads to an aggregated arrival rate of $\lambda = N \cdot \alpha$ bursts/sec. Service rate $\mu$ of the system is a function of the mean burst length E(L) and the server capacity C as shown in the traffic system model for M/M/1(m):

$$\lambda \lambda \lambda \lambda \lambda \lambda \lambda \lambda = \frac{PB C_{eq}}{E(L)}$$

Fig. 7: M/M/1(m) traffic model

Formulas for the QoS parameters of M/M/s(m) and their derivation can be found in any traffic theory book, e.g. [4]. System behaves like a M/M/1 system with infinite buffer when the buffer size is equal or greater than 100.

4.3. Cell/Packet level

In IP/ATM networks signal stream of a MMM or an On-Off source is normally not encapsulated in only one packet but a sequence of packets or cells. Therefore cell level requires more sophisticated models such as semi-markov models, fluid models or Markov modulated poisson models.

These models are mainly required for the design of router and switching equipment. They normally need much processing time and therefore are not well suited for dimensioning of large networks with many nodes.

For reasons stated the cell level will not be considered explicitly. If voice packets in On-Off model or the bursts in M/M/1(m) model are further divided into smaller transmission units such as ATM cells this can be taken into account by incrementing the source data rate in the burst model.

5. Validation of the VoIP models

To carry out the verification of the previous models it is necessary to develop the behavior of the voice sources completely. Most of the programs for alone network simulation have defaulted traffic generators based on typical distributions (exponential, Erlang, etc). This makes necessary to design a model on-off based on generic modules of this type, emulating the behavior of the same one by means of the corresponding machine of finite states.
5.1. on-off source model
As it was indicated previously, on-off source consists of two basic states, activity and silence. It should give the possibility to establish such parameters as the distribution of time among arrivals, silences distribution and distribution of calls duration. Resulting machine of states is directed by three types of events, one to create the packets, and two for the transition among states:

Fig. 8: On-off source process model.

5.2. Gateway model
For the interconnection of voice nodes, it is necessary to define a device that format the voice packets and route toward the address of corresponding destination. Gateway model provides transmission / reception ports for 100 voice nodes, and an interconnection port with other sub-networks, so that its direct connection is facilitated to the dorsal connection, which will be used for the study of statistical multiplex of all the sources.

Gateway has a finite queue of size m voice packets that can be used to adjust the extreme retard to end. Also the routing capacities are very simple, so that different packets are addressed toward the destination node that should be modeled as a drain located in the sub-network.

This type of nodes allows to obtain very interesting statistical figures, e.g. packet loss rate and total received packets. Additionally, by means of an additional parameter, gateway capacity in bits and/or in packets, it is possible to model the global queue of incoming voice packets.

5.3. Network model
Once designed the different nodes, it is necessary to establish the network topology that allows to interconnect them. Selected network model consists on the interconnection of two subnetworks through a dorsal connection. First subnetwork contains the gateway node with access to 100 voice nodes in star. Second subnetwork consists on a single voice node that takes charge of picking up the traffic of the dorsal connection, writes statistics of received packets, and finally it destroys all them.

6. Validation of the analytic models by means of simulation
For the obtaining of corresponding results we will treat network model described before as a block in which input parameters will be the number of users N, the size of gateway queue m, and the capacity of dorsal connection C. However, the behavior of the sources will be fixed for all the simulations (it is fixed the rate of burst for example).

Probabilities corresponding to each one of the states p_ON and p_OFF is also fixed respectively at 0.4 and 0.6, and the time among arrivals is of 0.02 sec. (for packets of fixed size of 480 bits).

The most interesting results are mainly the probability of loss P_B and the mean and variance of delay in the gateway queue. By means of these values it is possible to obtain the minimum of capacity low given conditions of P_B and delay.

6.1. comparison with the binomial model
Graphic results shown next represent the capacities of the dorsal connection for probabilities of inferior losses at 10^{-4}, 10^{-3} and 10^{-2}, obtained by means of simulation and making use of binomial pattern. It is observed that binomial pattern carries out an overestimation of necessary capacity in the connection for all cases, such and like it was assumed in its theoretical analysis.

Fig. 10: Theoretical and simulated backbone capacity (loss prob. < 0.001)
The reason rests in that the size buffer m used by the simulated model disappear in the theoretical pattern. This buffer diminishes the probability of loss substantially, impeding that it leaves of packets that found the busy connection they can be lost. This is translated in a better use of connection. However, an unexpected effect appears, en-to-end delay then is increased.

Fig. 11: Theoretical and simulated backbone capacity (loss prob. < 0.01)

Fig. 12: Distribution of burst loss length.

6.2. comparison with the M/M/1(m) system

While for the comparison of individual values it is not representative, it was observed that binomial model gave better results in comparison with M/M/1(M) pattern. However, it is required of more intensive studies that support this statement.

Another interesting fact when comparing both models was the one that stops relatively big sizes of buffer (79 positions) and capacities of connection drops (50%) Probability of loss in the M/M/1(m) model it is inferior to the one obtained by simulation.

This way, theoretical model underestimates the capacities of the according connection it increases the size of the buffer and the load of the connection. A possible explanation is that due to the nature of the bursts generated by on-off sources, a buffer near to congestion is much easier that it is surpassed its capacity that in the case of Poission sources. A detailed analysis of this phenomenon will be reason of future studies.

7. Conclusion

Simulations and developed analytic models are relatively simple leaving aside certain aspects that appear in the reality. For example, supposition used by M/M/1(m) system of a FIFO queue, supposes that sources with long periods of burst can prevent that other sources can be enqueued, when it would be possibly more appropriate to use more equal politicians of having enqueued, as for example Round Robin.

Binomial model can be used for the estimation of the capacities of the connections and the control of admission of calls. However, as pure system of losses that is, it cannot be used for the estimation of the buffer, in spite of being fundamental parameter in the voice gateway. For the calculation of this value, it is possible to make use of M/M/1(m) model like complement to binomial model.

Bursty model that it neither follows the loss of packets can be dear by means of theoretical models. Only by means of simulations it can be carried out, what takes to think of new studies that allow to define those parameters that affect to this behavior, taking like base the simulations for the obtaining of appropriate theoretical models.

References: