Published in *Proceedings of the Third International Workshop on Services in Distributed and Networked Environments (SDNE'96),* IEEE Computer Society, June 3-4, Macau, pp. 52-60, 1996.

A Scheme for the Quantification of Congestion in Communication Services and Systems

Edmundo Monteiro Gonçalo Quadros Fernando Boavida

Affiliation and Address:
Universidade de Coimbra
Departamento de Engenharia Informática
Pólo II - Pinhal de Marrocos
3030 Coimbra
PORTUGAL
Tel: +351 39 7000000
Fax: +351 39 701266
e-mail: edmundo@dei.uc.pt

Abstract

The characterization of congestion in communication systems is a pre-requisite for the definition of mechanisms that can be used in congestion control, with the goal to minimize its effects on the performance of distributed applications. In order for those mechanisms to work properly, there is the need to quantify congestion of services and systems.

This paper proposes a scheme that enables the quantification of the congestion of a particular communication service, the comparison of the congestion degree of two or more services, and the quantification of the global system congestion, based on the definition of a set of system QoS parameters, of their normal variation limits and of their degradation thresholds. The paper begins with the characterization of the congestion problem. Following, a congestion definition is presented and a congestion metric is proposed. The paper ends with an analysis of the quantification of congestion in communication systems when looked at as a set of several consecutive modules.

1. Causes and effects of congestion

The congestion phenomenon is associated with all communication systems of variable geometry, that is communication systems where there is a dynamic number of communication players (acting as information sources and/or destinations), where traffic characteristics can change, or where the available communication resources are not constant. This is the case of the majority of modern communication systems, for which congestion control has become a major issue due to the need to support a variety of applications with and without stringent communication requirements, over a variety of transmission technologies.

From a global communication system perspective, and as a first step to the problem characterization, it can be said that congestion occurs whenever the total amount of traffic that enters a communication system in a fixed time interval is greater than the communication system outgoing flow capacity in the direction of traffic destinations, in the same time interval. Congestion affects the quality of communicating applications - leading, in extreme cases, to their termination - and results in resource wasting, as communication resources must deal with the original traffic overload as well as with the traffic that results from information retransmissions generated by losses due to congestion.

Under working conditions, one can identify three distinct operating zones for a communication system: the *linear zone*, the *congestion zone* and the *collapse zone* .

In the linear zone the throughput increases linearly with the system load, and the transit delay remains practically constant. In this zone, the communication system is being used well below its maximum capacity and, thus, it can rapidly respond to the load variation without significant impact on the transit delay. From the applications point of view, communication systems should be engineered and tuned to work in this zone. Nevertheless, cost/benefit considerations prohibit this, leading to a dimensioning that implies a non-zero saturation probability, which corresponds to a non-zero probability of entering the congestion zone.

In the congestion zone, the saturation of the transmission resources - due to the load increase - leads to the intensive use of the communication system buffering capacity, which results in an increase of the transit delay and in a practically constant liquid throughput. It is in this zone that resource utilization is maximized, at the cost of a significant increase of the transit delay.

In case there are no mechanisms to prevent the load increase when the system is already in the congestion zone, the system will enter in the collapse zone, which is characterized by the blocking of the communication resources due to intensive retransmissions generated by the overflow of the system queues. In this zone the throughput decreases (approaching zero), and the transit delay can increase up to extremely high values. In this case, all of the communication system bandwidth is being used for retransmissions.

From a macroscopic point of view, the optimum system working point is at the threshold between the linear and the congestion zones. At this point the ratio between the liquid throughput and the transit delay - also known as the "power" of the communication system [1] [2] - is maximized. This situation corresponds to a compromise between the maximization of resource usage (which is desirable from the communication system perspective) and the minimization of the transit delay (desirable from the communication services perspective). The congestion control functions should lead the system to this optimum working point, controlling the system in such a way that avoids the entrance in the congestion zone, and preventing it from entering the collapse zone.

But keeping the optimum system working point does not guarantee the fulfillment of individual requirements of the services supported by the communication system. Each service may have its own requirements in terms of the compromise between the liquid throughput and the transit delay, which may be different from the compromise adopted at system level. In addition, each communication service may have specific needs determined by other communication parameters, that can hardly be met by the traffic characteristics that result from the system optimum working point. Finally, a communication system can present a globally congestion-free behavior and, when analyzing each of the active applications, show great asymmetries in resource usage.

Thus, in addition to search and maintain the global optimum working point, congestion control functions should contribute to the fulfillment of individual communication services needs, and avoid excessive use of resources by some services when compared to others.

2. Congestion definition

As mentioned above, a globally non-congested communication system may behave as a congested one for one or more of its users. Thus, there is the need for a congestion definition that is able to cope with congestion from a global as well as from an individual point of view.

The following types of congestion definitions are commonly found in the literature:

- *“a communication system is congested if the transit delay is greater than X”* [2] [3];
- “*a communication system is congested if the effective throughput is less than Y”* [4];
- a combination of the above definitions [5] [6].

These types of definitions are not precise, as they evaluate congestion on the basis of one or more of its effects (increase of transit delay and/or decrease of throughput) without taking into account the main cause of congestion: the *load increase*. In addition, the determination of the congestion thresholds X and Y is subjective and/or difficult, which results in the inability to quantify the congestion degree of the communication system. Another limitation lies in the fact that congestion is evaluated from a global perspective, without concern with individual applications. Thus, it is perfectly possible that in a congested system (according to one or more of the above presented definitions) certain applications may be able to perform within the throughput and delay limits that suite their needs.

In [7] a congestion definition is proposed that overcomes some of the above mentioned limitations. According to this author, a communication system is under congestion (from a user point of view) if the system usefulness decreases due to an increase of system load. The concept of usefulness expresses the user preference for the communication resources, through a usefulness function. This definition evaluates congestion from the users point of view, and characterizes congestion on a cause basis (as opposed to an evaluation based on the effects of congestion). Nevertheless, it does not support the quantification of the congestion situation, as the usefulness functions are user-specific and, thus, usefulness values cannot be compared - in order to measure relative congestion among users - nor combined - in order to measure the global congestion of the communication system.

Normally, communication systems and applications are dimensioned in such a way that the traffic alterations caused by the physical and technical limitations are within the limits tolerated by the applications and do not prevent their normal functioning. When the traffic characteristics are modified in a degree that decreases the performance of a given distributed application, the application is said to be affected by congestion. Thus, the congestion phenomenon can be characterized, from a communication service user point of view, by the following definition:

> Definition 1 - *congestion of a communication system: a communication system is congested whenever the functioning of communication services is affected in a way that is perceptible to their users.*

This definition emphasizes the communication services users perspective and accommodates all the factors that may cause the refusal, interruption, or degradation of the communication services, in addition to the load increase factor. In fact, this factor independence is consistent with the user point of view, to whom service degradation is the only effect he/she is concerned with, regardless of the factors that cause it.

The proposed definition has - when compared to the previous definitions based on throughput variation and/or transit delay - the advantage of characterizing congestion from a *microscopic* point of view - for each of the supported services, and at each instant of time - as opposed to a *macroscopic* characterization based on the global throughput and transit delay. Nevertheless, the proposed definition is subjective and needs to be complemented by a *metric*, in order to enable the quantification of the congestion of communication systems.

3. Congestion metric

From a user perspective, a communication system without congestion is characterized by the ability to provide the *quality of service* (QoS) required by the active applications. Quality of service can be objectively defined by a set of operating parameters or, implicitly, by a set of values that are considered "normal" when the application is active. On networks that are based on the resource reservation paradigm - normally operating in the connection-mode - it is always possible to objectively establish a set of quality of service parameters, because these are the parameters that are necessary for the resource reservation at the time of connection establishment. On networks that are based on the best effort paradigm - normally operating in the connectionless-mode - there is no need to explicitly define the QoS parameters; nevertheless, they can be implicitly deducted from the evaluation of the behavior of applications.

In either case - best effort or resource reservation - it is always possible to determine a set of parameters - the QoS parameters - that are responsible for the characterization of the quality of service of the supported applications. Let P_{QoS} , defined in Expression 1, be the set of all of the QoS parameters supported by a given communication system. The description of the physical significance of each parameter and the identification of its respective units is considered to be associated with the definition of the P_{QoS} set.

$$P_{QoS} = \{q_1, q_2, q_3, ..., q_n\} \qquad (1)$$

For each of the supported services s_i , and at discrete time instants t_k (or continuously in time), it is possible to measure or compute a set of values (one for each parameter belonging to the P_{QoS} set) that can be stored in a vector, as shown in Expression 2.

$$V_{QoS}(s_i)_{t_k} = [v_1 \quad v_2 \quad v_3 \quad \ldots \quad v_n] \qquad (2)$$

The specification of the quality of service necessary for each user application may be done, for each parameter, by the specification of an interval in which the parameter values imply no QoS degradation, and the specification of one lower threshold and one upper threshold beyond which the quality of service is unacceptable. The set of interval limits and degradation thresholds, specified for a given service s_i, may take the form of the matrix presented in Expression 3 - the QoS matrix, M_{QoS} - in which m_j and M_j are, respectively, the minimum and maximum values that parameter q_j may take without QoS degradation, and l_{mj} and l_{Mj} are the thresholds that subtracted from m_j and added to M_j, respectively, define two operating zones with degraded - but still acceptable - quality of service.

$$M_{QoS}(s_i) = \begin{bmatrix} m_1 & l_{m_1} & M_1 & l_{M_1} \\ m_2 & l_{m_2} & M_2 & l_{M_2} \\ m_3 & l_{m_3} & M_3 & l_{M_3} \\ : & : & : & : \\ m_n & l_{m_n} & M_n & l_{M_n} \end{bmatrix} \qquad (3)$$

The M_{QoS} elements m_j, M_j, l_{mj} and l_{Mj} may define static intervals, in the case of services with deterministic QoS needs, or probability intervals, in the case of services with probabilistic QoS needs. At the limit, M_{QoS} elements may be random variables, with associated probability distribution functions, in order to deal with QoS parameters of stochastic nature. In this work we will only consider QoS parameters defined in an absolute way or QoS parameters that are associated with fixed probability values. The study of the needs and implications of stochastic QoS parameters is the subject of further work.

The QoS matrix expresses the *service contract* - whether explicitly negotiated or implicitly assumed - between the user and the service provider, for a given communication service. In the QoS matrix of a specific service only the values (limits and thresholds) for the QoS parameters that are of concern for that service are specified. This parameter set is, normally, a subset of the system QoS parameters. For the parameters that are of no concern for a given service, the QoS matrix will hold the values $-\infty$, $+\infty$ and/or ∞. This is also the case for the parameters for which it is only required to specify the upper or the lower limit and threshold.

The congestion state of a communication service can be qualitatively evaluated at each instant - in light of the Definition 1 presented above - by the deviation of the parameter values with respect to the limits defined in the respective service QoS matrix. Consider the following definition:

Definition 2 - *deviation index (Id): being q_j a QoS parameter for the communication service s_i , $v_j(t_k)$ its value in instant t_k , m_j and M_j its normal variation limits, and l_{mj} and l_{Mj} its QoS degradation thresholds, then the QoS parameter deviation index, (Id), is given by:*

$$Id_{s_i,q_j}(t_k) = \begin{cases} 0 & , for\ m_j \leq v_j \leq M_j \\ 1-10^{-\left(\frac{m_j - v_j(t_k)}{l_{mj}}\right)} & , for\ v_j < m_j \\ 1-10^{-\left(\frac{v_j(t_k) - M_j}{l_{Mj}}\right)} & , for\ v_j > M_j \end{cases}$$

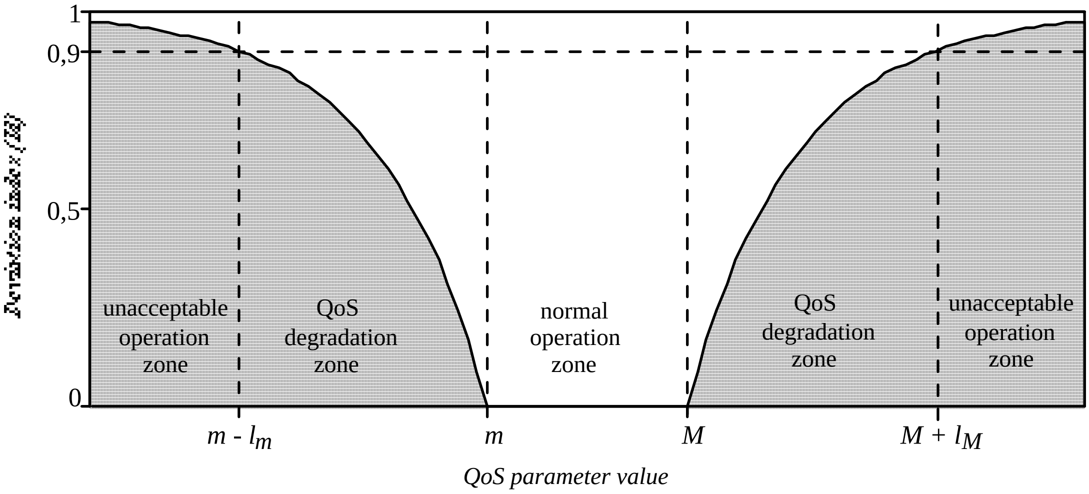

Figure 1 QoS parameter deviation index

The *deviation index (Id)*, proposed in the above definition, enables the quantification of the divergence of the value of a QoS parameter in relation to the limits and thresholds defined in M_{QoS}, in any point of the communication system. It can take all the values between 0 (null deviation) and 1 (maximum deviation), being 0 for all the parameter values inside the interval defined by m_j and M_j (normal variation limits). Outside this interval, the value of *Id* is a function of the divergence in relation to the normal variation interval and of the QoS degradation thresholds.

The deviation index identifies five different operating zones, as illustrated in Fig. 1. If the QoS parameter value is between *m* and *M*, the *Id* index is 0, and this identifies the normal operating zone. The intervals $[m - l_m, m]$ and $[M, M_j + l_M]$ correspond to two operating zones with QoS degradation, in which the *Id* index rises up to 90% of its maximum value. For *Id* values greater than 90% the parameter values are such that the quality of service is unacceptable.

In the QoS degradation zones, *Id* follows a base 10 exponential law - which dictates the so called quality degradation curve - as established by definition 2. This was chosen mainly by the fact that the majority of the QoS parameters have a direct or indirect effect on human senses (and patience), which have response curves that normally obey decimal logarithmic functions.

Figure 1 shows a particular situation in which the QoS parameter has symmetric, and finite, variation intervals. In Figure 2 other possible situations are illustrated.

The worsening of the deviation index for a QoS parameter - which corresponds to a quality of service degradation - is a sign of congestion, in light of definition 1. Congestion of a given communication service may be quantified using the respective QoS parameter deviation indexes, evaluated at the output of the communication system, immediately before the receiving communication services. Nevertheless, we must take into account the possibility of the incoming traffic having an *Id* not equal to zero (traffic not conformant with the service contract expressed by the QoS matrix). In this case, a non null *Id* at the system output is not exclusively due to congestion in this communication system.

In order to correctly evaluate the congestion introduced in a service by a communication system, it is necessary to relate the QoS parameter deviation indexes at the input and at the output of that communication system. Consider the following definition:

Definition 3 - *congestion index (Ic): being q_j a QoS parameter for the communication service s_i and, $Id^{in}_{s_i,q_j}$ and $Id^{out}_{s_i,q_j}$ its deviation indexes at the communication system input and output, respectively, in instant t_k, then the QoS parameter congestion index, (Ic), is given by:*

$$Ic_{s_i,q_j}(t_k) = Id^{out}_{s_i,q_j}(t_k) - Id^{in}_{s_i,q_j}(t_k)$$

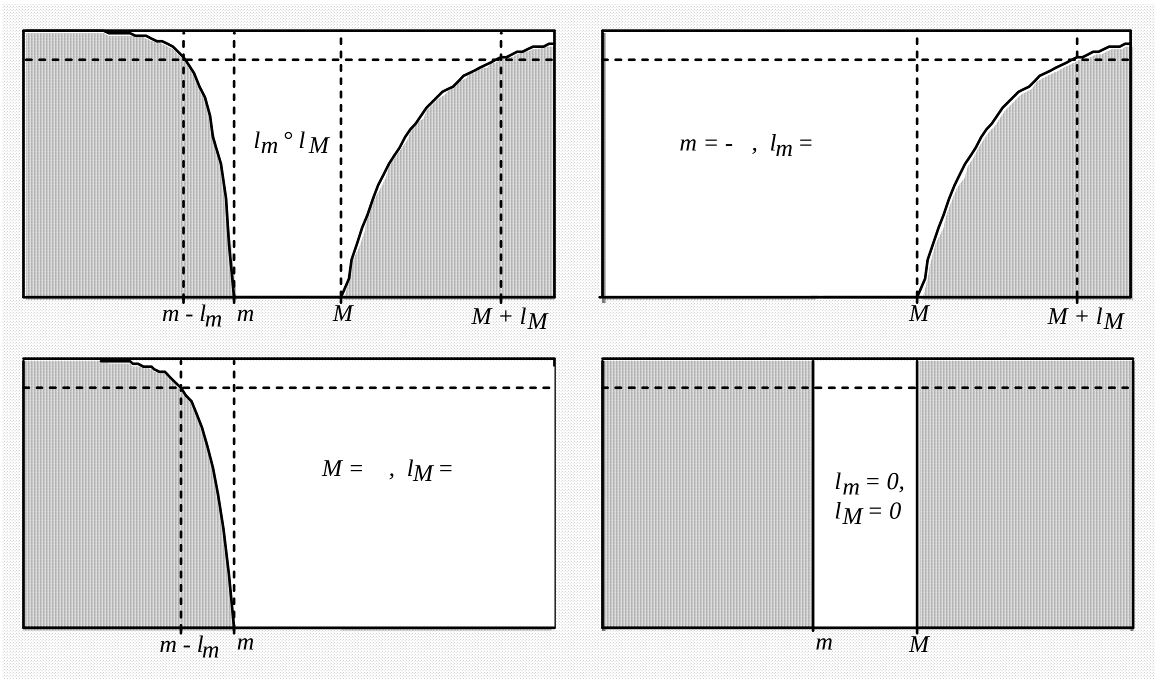

Figure 2 Some examples of quality degradation curves

The above defined *congestion index* measures the congestion introduced by the communication system in a given QoS parameter of a given service, and it can take values between -1 and 1. When the input traffic is in conformance with the normal variation limits, congestion is simply the value of the QoS parameter *Id* index at the system output. When the input traffic violates the service contract and the parameter deviation index is worsened in the communication system, congestion is evaluated by the difference between the output and input deviation indexes. Finally, whenever the deviation index is reduced by the communication system, congestion is negative. Negative congestion situations correspond to an active participation of the communication system in the improvement of the QoS characteristics.

Using the congestion indexes *(Ic)* of the set of QoS parameters of a given communication service s_i, congestion can be evaluated, at instant t_k, and from the user perspective, by the *service congestion index* — $C_{s_i}(t_k)$, — which is given by the weighted average of all the QoS parameter congestion indexes, as shown in Expression 4. The c_j constants are defined by the system management, for each service, and are used to weight the QoS parameters according to their relative importance for the communication service.

$$C_{s_i}(t_k) = \frac{\sum_{j=1}^{n} Ic_{s_i,q_j} \cdot \; c_j}{\sum_{j=1}^{n} c_j} \qquad (4)$$

The *global system congestion* at instant t_k, $Cg(t_k)$, can be obtained by the average of the service congestion indexes of all of the active services, affected by the *service success rate*, as indicated by Expression 5, where N is the number of active services at instant t_k. The *service success rate* measures the probability of success of a given service, taking into account the number of times the service was aborted - not established by lack of resources or abruptly terminated - (Na), and the number of times it was successfully accomplished, (Ns), since the system activation until instant t_k.

$$Cg(t_k) = \frac{Ns}{Ns + Na} \cdot \frac{1}{N} \sum_{i=1}^{N} C_{s_i}(t_k) \tag{5}$$

The *service success rate* is included in the expression that evaluates the global system congestion in order to reflect the effects of resource shortages in service establishment as well as the effects of faults of any kind that lead to the abrupt termination of communication services. Expression 5 uses a plain average, instead of a weighted average, because the relative importance of each service can be reflected in its QoS matrix values, making the service weighting in Expression 5 unnecessary.

Using the expression presented above, one can obtain global system congestion averages by integration of the instantaneous values of Cg over a period of time (Expression 6) or by averaging the Cg samples at discrete - and regular - time instants $t_1, t_2, \ldots, t_m$ (Expression 7).

$$\overline{Cg} = \frac{1}{t} \int_0^t Cg(t) \tag{6}$$

$$\overline{Cg} = \frac{1}{m} \sum_{k=1}^{m} Cg(t_k) \tag{7}$$

In the case there is the need to evaluate the congestion caused by a specific communication system component (e.g., an intermediate system) the proposed expressions can still be applied. For that, it is sufficient to consider the deviation indexes at the input (Id^{in}) and output (Id^{out}) of the particular component, as long as the normal variation limits (m and M) and the degradation thresholds (l_m and l_M) are established for that component (instead of being established for the whole system). This fact brings out an interesting issue - that will be addressed in the next section - concerning the congestion variation along the path constituted by the various system components, and the determination of the limits and thresholds for the parameter values in each component that guarantee a given global congestion degree.

4. Congestion variation along the communication path

Consider the communication system model presented in Figure 3, in which the communication system is subdivided in N consecutive modules. The modules can correspond to a functional block (e.g., a protocol layer), to a whole or part of a physical system (e.g., physical medium, switch, bridge, router, end-system), or to the concatenation of multiple physical systems forming a communication subsystem.

In order to relate the end-to-end congestion with the congestion in each of the communication system modules, it is necessary to identify the influence of each module in the QoS parameters of each communication service.

In connection-mode networks, the communication services QoS matrix must be verified during the negotiation/reservation process, at the time of connection establishment. In each module *p* of the communication system, it is necessary to reserve the resources required by the service and to establish the normal variation limits (m_j^p and M_j^p) and the degradation thresholds ($l_{m_j}^p$ and $l_{M_j}^p$) with respect to the module in question, and for each of the q_j QoS parameters. Using the information collected for each module, a verification must be made in order to determine if it is possible to guarantee the QoS values contained in the system QoS matrix. In the case it is not possible, the connection should be aborted, by lack of operating conditions.

In connectionless-mode networks, normally there isn't a resource reservation process, and the QoS matrix must be checked using previous knowledge relating to the typical characteristics of the service to be supported, or in the previous information - if it exists - on the normal variation limits (m_j^p and M_j^p) and on the degradation thresholds ($l_{m_j}^p$ and $l_{M_j}^p$) that correspond to each module, *p* , and for each of the q_j QoS parameters.

Knowing the normal variation limits m_j^p and M_j^p, and the degradation thresholds $l_{m_j}^p$ and $l_{M_j}^p$, that characterize each module *p*, for each of the q_j QoS parameters, it is possible to determine - by the use of definition 2 - the *deviation indexes* at the input and output of the module ($Id^{p_{in}}$ and $Id^{p_{out}}$, respectively). Using these values in Definition 3, one can obtain the *congestion index* of the q_j QoS parameter, concerning service s_i and the module *p*, for instant t_k (Expression 8).

$$Ic_{s_i,q_j}^p(t_k) = Id_{s_i,q_j}^{p_{out}}(t_k) - Id_{s_i,q_j}^{p_{in}}(t_k) \tag{8}$$

It is important to note that, in Expression 8, the $Id^{p_{in}}$ and $Id^{p_{out}}$ indexes are obtained by applying Definition 2, using the *local* QoS parameter limits and thresholds (m_j^p, M_j^p, $l_{m_j}^p$ and $l_{M_j}^p$), and *exclusively* taking into account the values of each of the QoS parameters at the input and at the output of module p , with no concern for any eventual deviation introduced by the upstream modules.

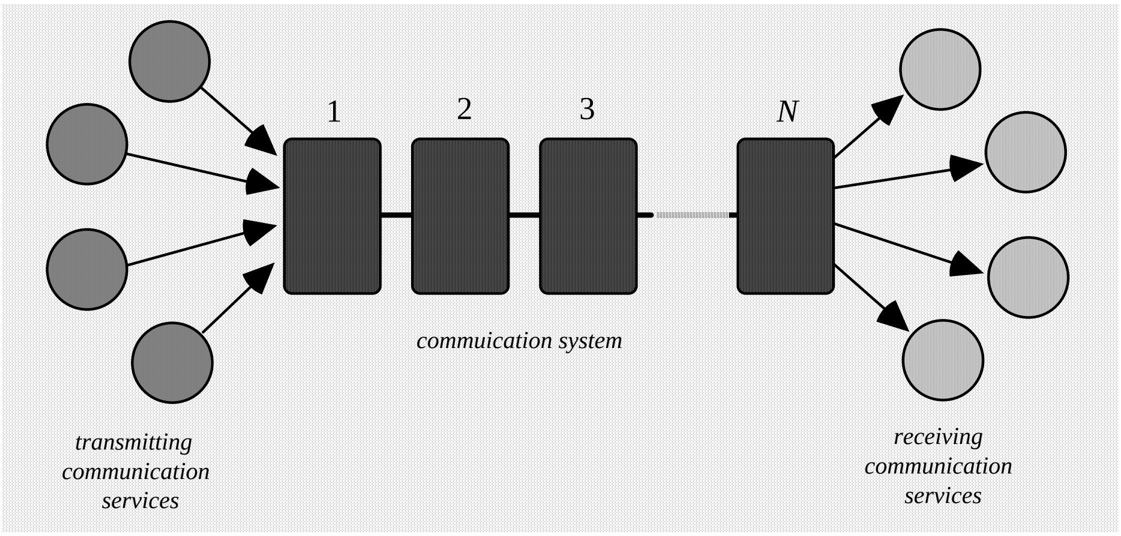
1
2
3
N
commuication system
transmitting communication services
receiving communication services

Figure 3 Model of an N-module communication system

The influence of the congestion indexes of each communication system module on the *global system congestion* mainly depends on the nature of the QoS parameters. One can identify two types of parameters, with regard to the way in which their variation influences the end-to-end degradation. Some parameters - that we call *cumulative* QoS parameters - suffer a continuous and progressive degradation, in a way that the resulting end-to-end degradation is the sum of the partial degradations along the communication path. Others - that we call *non-cumulative* QoS parameters - suffer a discontinuous and abrupt degradation, with the total (end-to-end) degradation being determined by the greater partial degradation suffered in the modules along the path.

The majority of the QoS parameters is composed of cumulative parameters. As examples, consider the parameters related to the transit delay (e.g., average delay, delay variation) and the parameters related to reliability and availability (e.g., error probability, mean time between failures).

Non-cumulative QoS parameters are less frequent. Some examples are the parameters related to the throughput (e.g., average throughput, peak throughput), as the end-to-end throughput is determined by the module where the throughput suffers the greater degradation. The parameters related to the relative priority of the communication services are also non-cumulative. Specifically, the parameters that, in some systems, determine the admission of new service requests - sometimes leading to the preemption of other services - are non-cumulative parameters: the end-to-end service priority is determined by the module where the service priority is the lowest.

The classification of QoS parameters into cumulative and non-cumulative parameters is one fundamental step to the characterization of the QoS parameter set, P_{QoS} (Expression 1), supported by the communication system. The definition of this set and the characterization of its members must be known to all the communication system modules, and must be made taking into account the particular properties of each module (for instance, whenever a module does not support a particular QoS parameter, the support of this parameter on a system-wide basis is compromised).

As already mentioned, for cumulative QoS parameters the end-to-end degradation is the result of the accumulation of the partial degradations in each module of the communication system. Thus, the normal variation limits and the degradation thresholds of a service QoS matrix must comprise the normal variation limits and the degradation thresholds of all the communication system modules that are used by the communication service.

Table 1 shows the relationships between the global limits and thresholds (m_j, l_{m_j}, M_j and l_{M_j}) and the partial ones (m_j^p, M_j^p, $l_{m_j}^p$ and $l_{M_j}^p$), for each QoS parameter, q_j, belonging to the QoS matrix of a given service, s_i for the cases of cumulative and non-cumulative parameters (Expressions 9 and 11). This table also presents the relationship between the end-to-end congestion index and the partial congestion indexes introduced by the set of the communication system modules, at instant t_k, and for both types of QoS parameters (Expressions 10 and 12).

The results contained in Expressions 10 and 12 enable the decomposition of the end-to-end congestion control problem in a set of more restricted congestion problems, each pertaining to a communication system module.

There are two reasons for the existence of an inequality in Expression 10. The first one has to do with the relation between the end-to-end limits and thresholds and their partial counterparts. The fact that the end-to-end limits and thresholds comprise - by excess, according to Expression 9 - the set of the partial limits and thresholds, leads to the fact that end-to-end congestion index may be less than or equal to the sum of the partial congestion indexes. The second reason has to do with the fact that the traffic deviation introduced by a module in a QoS parameter may be compensated by the behavior of another module.

For non-cumulative QoS parameters, the end-to-end congestion index is the greater of the partial congestion indexes introduced by each module.

Table 1 - Relationship between global and partial congestion

Cumulative QoS parameters	Non-cumulative QoS parameters
$M_{QoS}(s_i)_{q_j} = \begin{cases} m_j \leq \sum_{p=1}^{N} m_j^p \\ l_{m_j} \geq \sum_{p=1}^{N} l_{m_j}^p \\ M_j \geq \sum_{p=1}^{N} M_j^p \\ l_{M_j} \geq \sum_{p=1}^{N} l_{M_j}^p \end{cases}$ (9)	$M_{QoS}(s_i)_{q_j} = \begin{cases} m = \min \{m_p\} \\ l_m = \max \{l_{m_p}\} \\ M = \max \{M_p\} \\ l_M = \max \{l_{M_p}\} \end{cases}$ (11) $p = 1, 2, 3, \ldots, N$
$Ic_{s_i,q_j}(t_k) \leq \sum_{p=1}^{N} Ic_{s_i,q_j}^p(t_k)$ (10)	$Ic_{s_i,q_j}(t_k) = \max \{Ic_{s_i,q_j}^p(t_k)\}$ (12)

5. Conclusion

Present and emerging quality of service frameworks [8] [9] [10] do not address the quantification of the congestion degree of communication services and systems.

Congestion control is a key element of communication systems and services, as it is the basis for the provision and guarantee of quality of service to the applications. Any congestion control scheme must be able to detect and quantify not only the global system congestion but also the congestion degree of individual services, due to the fact that a globally non-congested communication system may behave as a congested one for one or more of its users.

The definition of a set of system quality of service parameters, of their normal variation limits and of their degradation thresholds is an indispensable step for congestion control, as it can assist in the identification of detailed service requirements and in the estimation of congestion.

This paper presented a proposal for the quantification of the congestion of a particular communication service, for the comparison of the congestion degree of two or more services, and for the quantification of the global system congestion, based on the definition of a set of system QoS parameters.

The proposed quantification scheme is applicable to *black box* communication systems, to individual communication modules, or to communication systems composed of N consecutive modules, which makes it possible to determine the influence of a given individual module in the global system congestion, as well as to control the congestion in the system modules knowing the end-to-end constraints imposed by the service. The presented scheme is used to support a congestion control framework for the provision of quality of service to applications which is presented and discussed in [11].

Acknowledgement

The work presented in this paper was partly supported by JNICT (Junta Nacional de Investigação Científica e Tecnológica) under contract PBIC/C/TIT/2460/95.

References

[1] - Raj Jain, K. Ramakrishnan e D. Chiu,*Congestion Avoidance in Computer Networks With a Connectionless Network Layer* , DEC-TR-506, Digital Equipment Corporation, Littleton, USA, August 1987.

[2] - Raj Jain, "A Delay-Based Approach for Congestion Avoidance in Interconnected Heterogeneous Computer Networks", *ACM Computer Communication Review,* Volume 19 (5), pp. 56-71, October 1989.

[3] - K. Ramakrishnan e Raj Jain, "A Binary Feedback Scheme for Congestion Avoidance in Computer Networks with a Connectionless Network Layer", in *Proceedings of the SigComm'88 Conference,* ACM, pp. 303-313, Stanford, EUA, August 1988.

[4] - Van Jacobson, "Congestion Avoidance and Control", in *Proceedings of the SigComm'88 Conference,* ACM, pp. 314-329, Stanford, EUA, August1988.

[5] - Raj Jain, D. Chiu e W. Hawe, *A Quantitative Measure of Fairness and Discrimination for Resource Allocation in Shared Computer Systems.*, DEC-TR-301, Digital Equipment Corporation, Littleton, USA, Setptember 1984.

[6] - Raj Jain, "Myths About Congestion Management in High-speed Networks", *Internetworking: Research and Experience,* Volume 3, pp. 101-113, 1992.

[7] Srinivasan Keshav, *Congestion Control in Computer Networks,* PhD thesis, Technical Report UCB:CSD-91-649, University of California, Berkeley, USA, December 1991.

[8] - Bob Braden, Dave Clark, Scott Shenker, "Integrated Services in the Internet Architecture: an Overview", RFC 1633, IETF - Networking Group, May 1994.

[9] - Andrew Campbell, Geoff Coulson, David Hutchison, "Quality of Service Framework", *Computer Communication Review*, vol. 24(2), pp. 6-27, April 1994.

[10] - ISO/IEC JTC1/SC21 N9680, Open Systems Interconnection, Data Management and Open Distributed Processing, "Information Technology - Quality of Service - Framework", Final CD, July 1995.

[11] - Edmundo Monteiro, *Controlo da Congestão em Sistemas Intermediários da Camada de Rede (Congestion Control in Network Layer Intermediate Systems)*, PhD. thesis (portuguese), University of Coimbra, Coimbra, Portugal, 1995.