A Queuing Delay Model for Dimensioning Interactive Service Traffic in HSDPA Networks

Ghassane Aniba and Sonia Aïssa
INRS-EMT
University of Quebec
Montreal, QC, Canada
{ghassane, aissa}@emt.inrs.ca

Abstract—**This paper considers interactive traffic queuing delay estimation in High Speed Downlink Packet Access (HSDPA) networks. The web browsing is the most widely used interactive application. Its traffic model is subdivided into three levels: session level, burst level and packet level. Based on this model, the statistical behavior of the queueing delay is studied. It is shown that the queueing delay for the web traffic follows an exponential distribution. Analytical modelling of the probability density function (PDF) of the queueing delay being untractable, we resort to simulated data and provide simple mathematical formulation of the different parameters which characterize the density function. Indeed, we present useful equations which could be used, directly, in HSDPA network dimensioning, and as a reference, to satisfy a certain quality of service.**

Keywords: Queuing delay modelling, HSDPA, Network Dimensioning.

I. Introduction

In 3G+ networks [1], packet-based data services having different requirements in terms of quality of service (QoS) and tolerance to delay, modelling the traffic and the queuing delay of such services is of extreme importance. Indeed, such models are needed in order to dimension practical networks and design efficient radio resource management strategies and communication protocols for these networks. In this context, the focus of this paper is to model the queuing delay of interactive services, and precisely the web-browsing traffic. For this purpose, the general traffic model provided in [2] is used to derive the parameters that specify the queuing delay probability density function (PDF) corresponding to the service under consideration.

The framework is applied to the High-Speed Downlink Packet Access (HSDPA). HSDPA is the new enhancement of the WCDMA downlink. Indeed, HSDPA introduces new adaptive link techniques, namely adaptive modulation and coding (AMC) and Hybrid Automatic Retransmission Query (HARQ). By means of these techniques, a HSDPA user equipment (UE) could reach an air throughput equal to 10Mbps. The AMC technique is applied by the use of a mapping between the channel quality, defined by the signal-to-interference-and-noise ratio (SINR), and the modulation-coding rate couple which could be used in such channel condition [3]. Each modulation-coding couple is equivalent to a number of simultaneous spreading codes and a transmission data rate value which could be used by the corresponding UE. There are thirty couple values, allowing thirty discrete data rate values. In addition, HSDPA UEs are classified upon their capabilities in terms of the maximum number of simultaneous codes and the modulation order a UE can support. Twelve categories are distinguished, eight of which (1–6, 11 and 12) are characterized by a processing capability of at most 5 simultaneous spreading codes. In practice, more than 70% of the UEs will belong to these categories, an issue that needs to be taken into consideration when modelling the delay [4]. Moreover, HSDPA allows previously transmitted bits from the original transmission to be combined with the retransmission ones. This combining strategy provides improved decoding efficiencies and diversity gains while minimizing the need for additional repeat requests. This operation is denoted as HARQ. The HARQ control is situated in the Node B, or base station, thus removing retransmission-related scheduling, an operation designed to reduce the delay and increase the efficiency of retransmitting data.

As aforementioned, the use of discrete data rate values and fast HARQ are used to improve the air throughput of the HSDPA network and reduce, at the same time, the delay introduced by the queuing and retransmission processes. Indeed, the air interface throughput in HSDPA networks is shared between different UEs. In order to increase the capacity of these networks, the capability of the network to carry packet-switched traffic is used along with multiplexing the traffic of different UEs. However, loading the network causes delay which, in turn, can cause degradation in the QoS of delay sensitive applications and in the overall performance of the network. In order to avoid degradation, it is essential to model the queuing delay of the studied service and use this model for the purpose of developing the communication strategies that ensure the increased capacity. However, because of the complexity of the packet traffic involved, as that of the web browsing, modelling the queuing delay through analytical computation is untractable. In [5], model fitting based on simulated data was used to model the packet data queuing delay PDF for video and voice services for a general wireless traffic. In this work, with the use of the web-browsing traffic model proposed herein, and based on the approach followed therein, we propose a general model for the web-browsing traffic applied to HSDPA, taking into consideration the use of discrete data rate values and ARQ retransmissions. In par-

0-7803-9305-8/05/$20.00 ©2005 IEEE

ticular, we provide the different parameters that characterize the queuing delay PDF, as a function of the air throughput T, network loading defined by the number of UEs N, error retransmission probability P_e, and finally the Reading Time parameter t_r associated to the web-browsing traffic. The latter is the time taken by users to read a downloaded web page.

The remainder of this paper is organized as follows. Section II describes the general traffic model for the web browsing with its associated parameters along with an overview of HSDPA. Our queuing delay modelling analysis and results are provided in section III, followed by conclusions drawn in section IV.

II. Web-Browsing Traffic and Characteristics of HSDPA

A. Web-browsing Traffic Model

In this paper, we use a general traffic model based on the definition provided in [2] for the web-browsing service. The model is divided into three levels:

- *Session level:*
  User sessions (web-browsing) are modelled at this level. We suppose the arrival of each session to follow a Poisson distribution with the assumption that each user has only one session during the busy hour.
- *Burst level:*
  Each packet session is formed by one/many *packet call(s)* and each packet represents a web page. The distribution of the inter-arrival of these packets, named *Reading Time*, follows a geometric distribution.
- *Packet level:*
  Each packet call is composed of a number of data packets. The distributions of the packet size and of the inter-arrival time are specified at this level.

The difference between the proposed traffic model and that provided in [2], is in the number of packet calls included in each session packet. Here, we suppose that each session remains active during the entire busy hour, which means that the number of packet calls is unlimited instead of being limited to 5 as in [2]. In Table I, we provide a summary of the traffic parameters considered for web-browsing sessions as represented in Figure 1. For each traffic parameter, the value given in Table I corresponds to the mean of each random variable according to which the parameter is modelled. In order to provide a general packet-call delay model, we consider different arrival data rate values, namely, 144kbps, 384kbps or 2048kbps, generated by means of choosing the value of packet data inter-arrival time within a packet call equal to $0.0277s$, $0.0104s$, or $0.00195s$ respectively. Moreover, different values are considered for the Reading Time $(2s, 4s, 8s, 16s, 32s)$. Higher values were studied $(64s, 128s, 256s)$, nevertheless, only the results corresponding to the above-mentioned ones are presented herein given that they were sufficient to provide a general queuing delay model. Finally, the quality of service is defined by the maximum packet call (web page) delay. Based on [1], an excellent quality of service is defined by a packet call delay no longer than $4s$. This limiting value was very useful in our delay modelling as will be shown later.

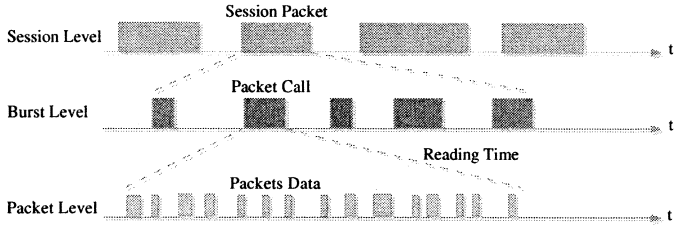

Fig. 1. WWW packet service session

Parameter	Value	Distribution
Session arrival process	-	Poisson
Reading time (t_r)	2s-32s	Geometric
Number of packets per call	25	Geometric
Packet inter-arrival time	0.0277s,...,0.00195s	Geometric
Arrival data rate	144kbps,...,2048kbps	-
Packet size	480 bytes	Pareto (1.2,81.5)

TABLE I
PARAMETERS OF THE WEB BROWSING TRAFFIC

B. High Speed Downlink Packet Access (HSDPA)

In HSDPA, a UE transmits its Channel Quality Indicator (CQI) parameter using the High Speed Dedicated Physical Control Channel (HS-DCCH). This parameter belongs to a set of thirty values, used to map the CQI into a SINR value that ranges over 30dB, in the interval [-5dB,25dB]. As for the mapping of a channel state, given by the SINR value, into a CQI, the following rule is used:

$$\text{CQI} \simeq \min(\max(0, \lfloor \text{SINR}_{hsdsch} + 6 \rfloor), 30) \qquad (1)$$

where SINR_{hsdsch} is expressed in dB and $\lfloor \cdot \rfloor$ denotes the integer floor operator. Based on the CQI value, the Node B assigns a maximum transmission rate r_i, or a transport block (TB) size that could be used in the next Transmission Time Interval (TTI) equal to 2ms, which could be used by the corresponding UE (Tab. II).

Besides that, there is a scheduling algorithm which chooses a UE or a set of UEs for which data will be transmitted in the next TTI. In order to make such decisions, some scheduling algorithms, such as the max CIR [6], use the SINR of each UE channel, when some others use the instantaneous r_i values and their average values R_i, for instance the Proportional Fairness method [7]. In this work, our focus being on dimensioning the network, we consider simple Round Robin (RR) scheduling. The RR algorithm allocates the channel to UEs in a circular way. In case all UEs have the same average channel quality, the RR scheduler provides a high efficiency in terms of fairness between users. Considering fairness as an important feature that should be maintained in servicing the different users, we implement the RR algorithm at Node B and assume an average transmission data rate of $R = 2Mbps$ for all UEs.

As previously mentioned, the HARQ technique is used in HSDPA in order to improve the performance of the network

TABLE II

CHANNEL QUALITY INDICATOR (CQI) TABLE MAPPING

CQI	0	1	2	3	4	5	6	7
TB size (bits)	0	137	173	233	317	377	461	650
CQI	8	9	10	11	12	13	14	15
TB size (bits)	792	931	1262	1483	1742	2279	2583	3319
CQI	16	17	18	19	20	21	22	23
TB size (bits)	3565	4189	4664	5287	5887	6554	7168	9719
CQI	24	25	26	27	28	29	30	
TB size (bits)	11418	14411	17237	21754	23370	24222	25558	

and reduce retransmission delays. Indeed, advanced combining techniques are used at the UE side, to take advantage from each retransmission of the same packet. In this paper, the HARQ protocol is taken into consideration by means of an error transmission probability P_e, defining the probability that a UE receives an erroneous packet.

A queuing delay model for dimensioning HSDPA networks is a useful and practical tool, especially because of the existence of a standardized mapping table that limits the values of allowed discrete data rates (Tab. II). Hence, by taking into consideration such limitation, the queuing delay model proposed herein will be general enough to be applied in realistic transmission scenarios so as to satisfy a required quality of service.

III. QUEUING DELAY MODELLING

Following the model-fitting approach used in [5], we provide results corresponding to the web-browsing service in HSDPA. Specifically, we provide the different parameters that characterize the queuing delay PDF as a function of the air throughput, the number N of UEs in the network and the Reading Time t_r.

In our study, a number of simulation runs, ranging between thirty and a hundred, were performed in order to provide precision to our delay modelling. In the same vein, unlike the data packet queuing delay measurements done in [5], measurement of the delay is performed here for each packet call (web page) in each session.

From our simulations for the web-browsing traffic, we found that the packet call queuing delay follows an exponential distribution (2). Indeed, Figure 2 shows the cumulative density function (3) of the packet call delay for different values of N. The Reading Time t_r in this figure is equal to $16s$, the arrival data rate equal to 144kbps, and P_e equals 0.5.

$$f_\tau(\tau) = \frac{1}{\mu}\exp(-\frac{\tau}{\mu}), \tag{2}$$

$$F_\tau(\tau) = 1 - \exp(-\frac{\tau}{\mu}), \tag{3}$$

where μ is the mean packet call queuing delay.

Our simulations show that the mean delay μ depends on the number of UEs in the network, the probability of a good packet transmission $P_c = (1 - P_e)$, and the Reading Time t_r. Besides, one important observation is that the mean delay

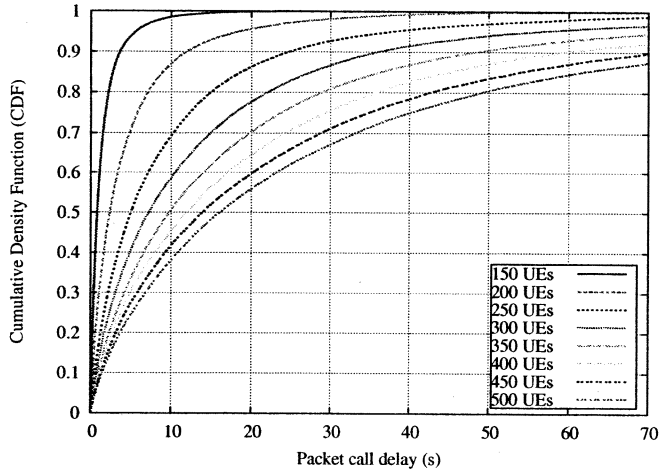

Fig. 2. Packet call delay cumulative density function for different numbers of UEs, N

μ is independent of the data arrival rate. Indeed, in Figure 3, we show the evolution of the mean delay μ, for different data rate values (144kbps, 384kbps and 2048kbps), different values of t_r (2s, 8s, 32s), and with $P_c = 0.5$. We observe that for all Reading Time values, we get a superposition of the three mean delay curves, corresponding to each data rate value. This result could be explained by the inherent specificity of an interactive service. In fact, even if we change the arrival data rate, the volume traffic of each packet data remains the same, only the inter-arrival of data packets changes. Hence, for such high values of arrival data rate, we can suppose that the packet call arrives at the buffer almost instantaneously, and as a consequence, the queuing delay will depend only on the volume of the packet call, which is the same in all cases. From the same results (Fig. 3), we can see that when the number of UEs present in the network exceeds a certain value, the mean delay μ becomes a linear function of N. This is one of the major differences between the mean packet data delay [5] and the mean packet call delay. Indeed, the mean packet data delay increases exponentially when the number of UEs N gets higher. Instead of that, and due to the interactive behavior of the web-browsing traffic, the mean packet call delay increases linearly with N.

As mentioned, a packet call delay value of 4s is considered as an excellent quality of service. Hence, we will consider that all delay values smaller than this threshold are negligible and considered equal to zero. As can be seen in Fig. 3, from a certain value of N_0, the mean μ is a linear function of N, and can be written as

$$\begin{cases} \mu = \frac{1}{a} \cdot N - b & \text{if } N > N_0 \\ \mu = 0 & \text{if } N < N_0 \end{cases} \tag{4}$$

with a and b parameters used to define the linear model.

The value N_0 defines the maximum number of allowed UEs in the network, when a delay value lower than 4s is needed. This limit can easily be calculated using:

$$N_0 = a \cdot b. \quad (5)$$

Hereafter, we present a formulation of the parameters a and b as a function of the air throughput of the network, T, and the Reading Time, t_r. Before that, two important observations must be made. Firstly, in Fig. 4, we can see that for all Reading Time values, the slope of the μ curve is the same, which means that the parameter a is independent of the Reading Time value t_r. Moreover, Fig. 5 shows that for different values of the probability P_c, the slope of the μ curve is different, which means that the parameter a depends on the probability P_c. Also, as can be seen in Fig. 6 at the value N_0, we have an intersection of all μ curves, which means that b is independent of P_c. These important observations simplify modelling of the mean delay μ. Indeed, Fig. 7 and Fig. 8 show that there is a linear relationship between the parameter a and the Reading Time t_r on one hand, and between the parameter b and the probability P_c on the other hand. These functions were found to be expressed as:

$$a = 22.357 \cdot P_c, \quad (6)$$

and

$$b = 0.88 \cdot t_r + 1.87. \quad (7)$$

Therefore, equations (4) and (5) can be expressed as:

$$\begin{cases} \mu = \dfrac{1}{22.357 \cdot P_c} \cdot N - (0.88 \cdot t_r + 1.87) & \text{if } N > N_0 \\ \mu = 0 & \text{if } N < N_0 \end{cases} \quad (8)$$

and,

$$N_0 = 22.357 \cdot P_c \cdot (0.88 \cdot t_r + 1.87). \quad (9)$$

As said before, air throughput is formulated as $T = P_c \times R$. The equations given above are all formulated in the case of $R = 2$Mbps, therefore, for a general formulation, we can write the mean delay μ, as function of the air throughput T according to

$$\begin{cases} \mu = \dfrac{1}{11.175 \cdot T} \cdot N - (0.88 \cdot t_r + 1.87) & \text{if } N > N_0 \\ \mu = 0 & \text{if } N < N_0 \end{cases} \quad (10)$$

with the air throughput T expressed in Mbps, and N_0 rewritten as

$$N_0 = 11.175 \cdot T \cdot (0.88 \cdot t_r + 1.87). \quad (11)$$

Equation (10) is the general analytic formulation of the mean delay μ. As such, for a given number N of UEs in the network, with an air throughput $T = (1 - P_e) \cdot R$, and a Reading Time value t_r, we can directly compute the density function of the packet call queuing delay τ (Eq. 2). By means of the latter, we can generate different statistics, for instance, the 95-percentile value τ_0, with

$$\tau_0 = -\mu \cdot \ln 0.05 = 3 \cdot \mu. \quad (12)$$

An important point to mention is that our general formulation is true for all *fair* scheduling algorithms. Indeed, the effect of the scheduling algorithm is included in the common air throughput $T = (1 - P_e)R$ over all UEs. For instance, current results and delay modelling can effectively be used in conjunction with a strict fair scheduling algorithm, such as the Adaptive Proportional Fairness (APF) method proposed in [8].

IV. Conclusions

This paper proposed a model for dimensioning the web-browsing service in HSDPA networks. The web browsing traffic was modelled through the layered structure: session level, burst level and packet level, in order to provide a general packet call queuing delay model for transmission in HSDPA. We showed that the packet call queuing delay is exponentially distributed, and presented a mathematical formulation of the different parameters which characterize the density function of the queuing delay. By means of this formulation or directly from our curves, for a given reading time value and a required mean queuing delay, one can deduce the maximum allowed UEs in the network, without the need of further simulations or computations. The proposed expressions can be used jointly with any fair scheduling algorithm which provides the same air throughput to all UEs. Under investigation is the use of the model presented in this paper for the design of a multi-class scheduling algorithm.

References

[1] 3rd *Generation Partnership Project; Technical Specification Group Services and System Aspects Service aspects; Services and service capabilities (Release 6)*, 3GPP Std. 22.105, Rev. 6.3.0, Mar. 2005.

[2] *Selection procedures for the choice of radio transmission technologies of the UMTS*, ETSI Std. 101.112, Rev. 3.2.0, Apr. 1998.

[3] 3rd *Generation Partnership Project; Technical Specification Group Radio Access Networks; Physical Layer Procedures (FDD) (Release 6)*, 3GPP Std. 25.214, Rev. 6.3.0, 2004.

[4] 3rd *Generation Partnership Project; Technical Specification Group Radio Access Network; Physical Layer Aspects of UTRA High Speed Downlink Packet Access (Release 4)*, 3GPP Std. 25.848, Rev. 4.0.0, Apr. 2001.

[5] G. Aniba and S. Aïssa, "A general traffic and queueing delay model for 3G wireless packet networks," in *Proc. 11th International Conference on Telecommunications, (ICT'2004)*, vol. 3, Fortaleza, Brazil, Aug. 2004, pp. 942–949.

[6] C. J. Ong, P. H. J. Chong, and R. Kwan, "Effects of various packet scheduling algorithms on the performance of high speed downlink shared channel in a WCDMA network," in *Proc. IEEE Pacific Rim Conference on Communications, Computers and signal Processing (PACRIM'2003)*, vol. 2, Aug. 28–30, 2003, pp. 935–938.

[7] A. Jalali, R. Padovani, and R. Pankaj, "Data throughput of CDMA-HDR a high efficiency-high data rate personal communication wireless system," in *Proc. IEEE Vehicular Technology Conference (VTC-S'2000)*, vol. 3, Tokyo, JP, May 2000, pp. 1854–1858.

[8] G. Aniba and S. Aïssa, "Adaptive proportional fairness for packet scheduling in HSDPA," in *Proc. IEEE Global Telecommunications Conference (Globecom'2004)*, Dallas, TX, Nov. 2004.

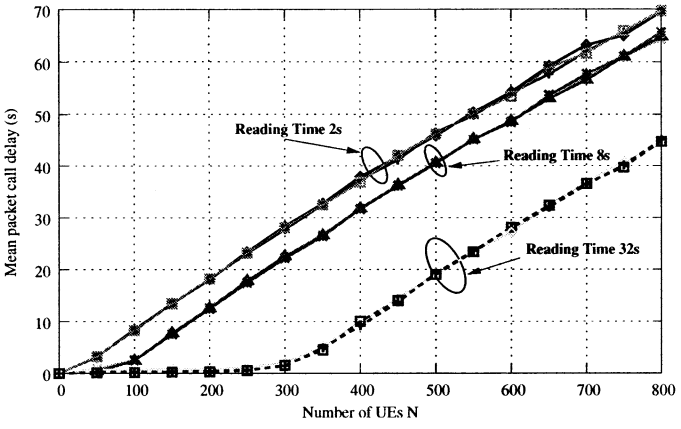

Fig. 3. Evolution of the mean packet call delay, μ, for different reading time values t_r and data arrival rates

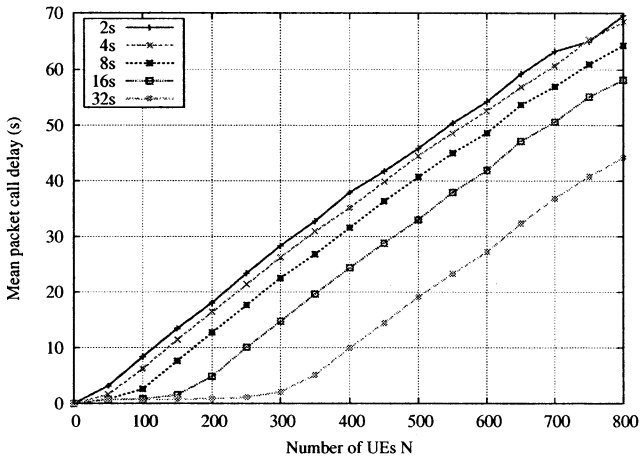

Fig. 4. Evolution of the mean packet call delay, μ, for different reading time values

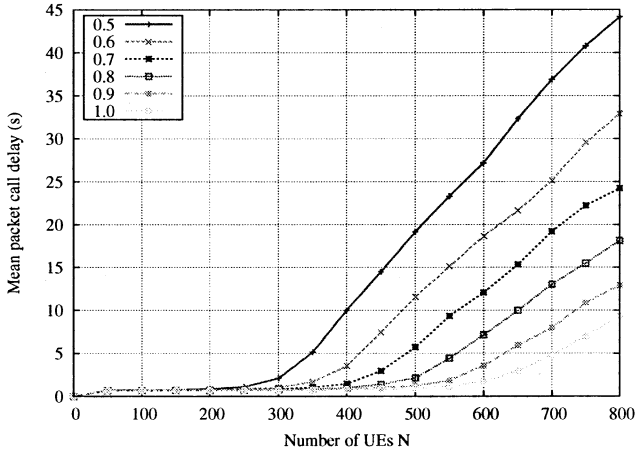

Fig. 5. Evolution of the mean packet call delay, μ, for different probability values P_c

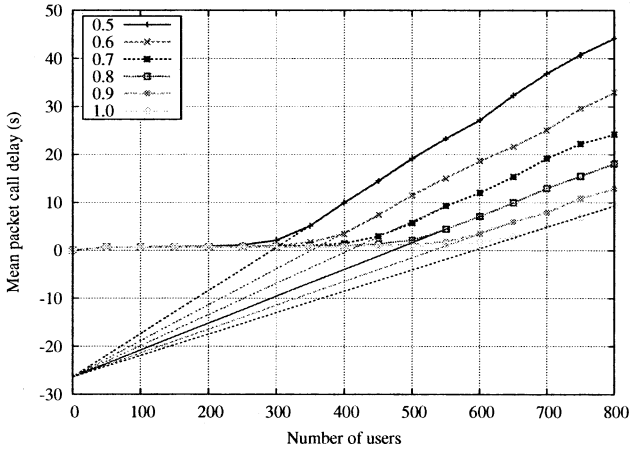

Fig. 6. Evolution of the mean packet call delay, μ, for different probability values P_c, including asymptotic linear curves

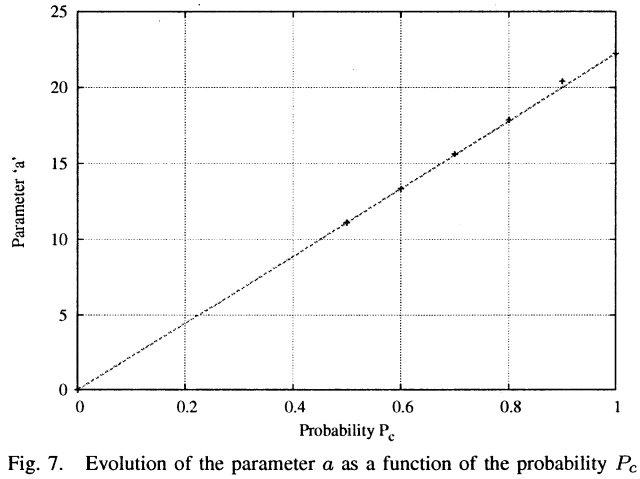

Fig. 7. Evolution of the parameter a as a function of the probability P_c

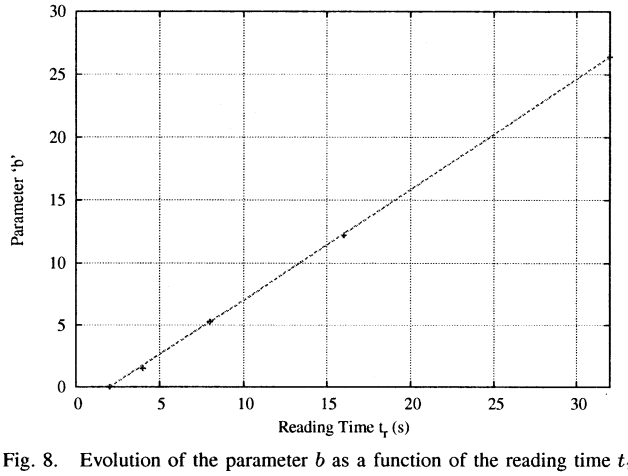

Fig. 8. Evolution of the parameter b as a function of the reading time t_r