Statistical Distribution of Traffic Sources in Network Simulation Tools

F. Camps, E. Thibault and S. Harasse

***Abstract*—In this paper, we present an approach to execute more realistic traffic sources in network simulators (like Network Simulator 2) for fixed and mobile networks. Traffic source models can be defined from network measures to give an accurate statistical distribution of inter-arrival time and packet size at packet level. These traffic sources models can be integrated into a network simulator to perform simulations. Furthermore, this approach can be applied to classical or emerging data applications.**

***Index Terms*—statistical distribution parameters, inter-arrival time, packet, network simulator**

I. Introduction

Network simulator tools are helpful to understand many problems that occur in large networks like Internet or internal networks. They are used to define complex networks with very high densities of routing equipment and traffic. These same tools also offer traffic source libraries that are necessary to simulate the dynamic behavior of a network. The problem with these libraries lies in the fact that the parameters of distribution of inter-arrival time and sizes of packets or the distribution itself are not always adapted to new applications that one wants to simulate [9,10].

Characterizing a source of traffic is to find a model that can generate inter-arrivals times (usually in milliseconds) and packets size (usually in bytes) [11]. The traffic models are defined by the laws of probability. In this mathematical approach, three levels are considered: Session, Application and Packet. The behavior of application level depends on the application in question (http, ftp, P2P, and so on…).
The applications are very different in their needs. They can impose “hard” or "flexible" time. Some real-time applications are associated with UDP/IP, the unreliable transport protocol. A second category includes non-real-time applications that are usually associated with TCP/IP, the reliable transport protocol.

In section II, we present the classical model for traffic sources in a network system. The section III presents the E/M algorithm applied to the statistical modeling of traffic sources at the packet level. In section IV, an example of application is detailed. Finally, section V presents the use of the method in a network simulator.

Manuscript received July 1, 2008.
F. Camps is with the French Scientific Research Center CNRS/LAAS as a software engineer. He works in the field of networks and more particularly on the modeling of the sources of multimedia traffic (email:fcamps@laas.fr).
E. Thibault was with the French Scientific Research Center CNRS/LAAS as a postdoctoral fellow. He is now in the Statistical Department of Gaz de France (email:ericthib2000@hotmail.fr).
S. Harasse is with the French Scientific Research Center CNRS/LAAS as a postdoctoral fellow, he researched fields are image and signal processing (email:sebastien.harasse@laas.fr).

II. Application Behavior in Three Levels

Usually a system that generates traffic sources is represented by three levels [5]:
• **Session level**: to model the arrival of all clients who connect the system to use a certain type of application (Poisson model is usually used).
• **Application level**: the density of information is defined according to the required application.
• **Packet level**: is the most basic level which generates packets that correspond to the information issued by the application.

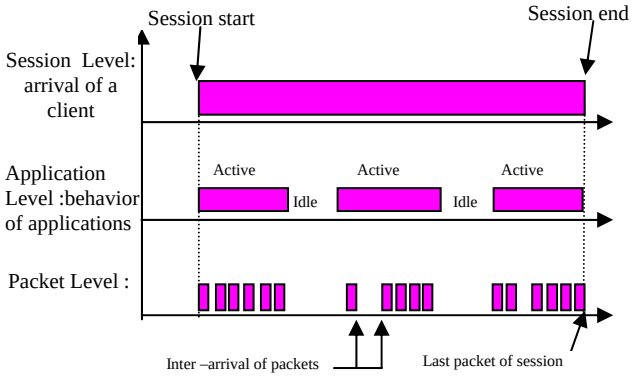

Fig. 1 - Three level representation

Below is a typical example that you can find in the literature: the distribution of session volume in mobile networks.

Table 1 - Distribution of HTTP session [8]

	Distribution	**144 kbps**	**384 kbps**
HTTP	Lognormal(μ,σ^2)	(5.2943 ;14.1839)	(5.4197 ;14.4979)

The volume of data at application level can be estimated by knowing the distribution of objects in web pages and simulating a web session.

Table 2 - Distribution of web object [2]

Variable	**Distribution**	**Mean**	**Std dev**
Web size page	Lognormal(8.43, 2.31)	19 Kbyte	124 Kbyte
HTML Document	Lognormal (7.97, 0.76)	4.6 Kbyte	11 Kbyte
Images	Lognormal(7.45, 4.04)	12 Kbyte	35 Kbyte

How can these distributions be estimated ?

III. TRAFFIC SOURCE ESTIMATION

In the literature [4, 5, 6, 7], we observe that traffic sources are often characterized by simple statistical distributions (exponential, normal, Pareto, …). These distributions don't always fit the real behavior of the application (at application or packet level). In order to improve the modeling of our application, a mixture of analytical models can be used [2]. The parameters for such mixtures are estimated by methods like E/M [1] (Expectation/Maximizing) and Levenberg-Marquardt (non linear least squares) from the observed samples in networks. The obtained mixed analytical model can be replayed in a network simulator. We briefly present in this section the E/M method to characterize a traffic source.

Let $f_1,...f_k$ be k parametric functions representing probability densities with parameters $\theta_1,...,\theta_k$. These functions are typically: exponential, normal, lognormal, Weibull, Pareto, gamma, Gaussian. We consider a mixture of statistical distribution [2,3]:

$p(y) = \pi_1 f_1(y) + ... + \pi_k f_k(y)$ with the coefficients (or weights) π_i positives and $\pi_1 + ... + \pi_k = 1$.

Let $y = (y_1,...y_n)$ be n samples of the observed traffic source and $\Psi = (\pi, \theta)$ the parameter vector to estimate where $\pi = (\pi_1,...,\pi_k)$ and $\theta = (\theta_1,...,\theta_k)$. The component f_i from which the sample y_i arises is unknown, and labeled by z_i . The complete sample is $x = (x_1,...,x_n)$, with $x_i = (y_i, z_i)$. The EM algorithm generates, from an initial approximation $\Psi^{(0)}$, a sequence of estimation $\{\Psi^{(m)}\}_m$. In each iteration, there are two steps:

Step E : compute $E\left[\log p(x \mid \Psi) \mid y, \Psi^{(m)}\right] = Q(\Psi, \Psi^m)$

Step M : find $\Psi = \Psi^{m+1}$ to maximize $Q(\Psi, \Psi^m)$

E/M can easily be applied to well known statistical distributions. However, for some distributions, it is difficult to obtain the parameters directly. This is the case of the **gamma distribution**, for which the steps of E/M algorithm are presented here.

Density function :

$$f(x) = \frac{b^{-a}}{\Gamma(a)} x^{a-1} \exp(\frac{-x}{b})$$

for all $x \geq 0, b > 0$

Parameters of the law according to the mean and variance:

$$mean = ab \rightarrow a = \frac{mean^2}{\text{var}}$$

$$\text{var} = ab^2 \rightarrow b = \frac{\text{var}}{mean}$$

$$(1): \begin{cases} \ln b + \dfrac{\Gamma(a)'}{\Gamma(a)} = \dfrac{\sum_{i=1}^{n} \omega_{ij}^m \ln x_i}{\sum_{i=1}^{n} \omega_{ij}^m} \\ ab = \dfrac{\sum_{i=1}^{n} \omega_{ij}^m x_i}{\sum_{i=1}^{n} \omega_{ij}} \end{cases}$$

$$(2): \begin{cases} b = \dfrac{1}{a} \dfrac{\sum_{i=1}^{n} \omega_{ij}^m x_i}{\sum_{i=1}^{n} \omega_{ij}^m} \end{cases}$$

Parameter a is obtained by dichotomy:

$$\left\{ g(a) = \frac{\Gamma(a)'}{\Gamma(a)} - \ln a = \frac{\sum_{i=1}^{n} \omega_{ij}^m \ln x_i}{\sum_{i=1}^{n} \omega_{ij}^m} - \ln\left(\frac{\sum_{i=1}^{n} \omega_{ij}^m x_i}{\sum_{i=1}^{n} \omega_{ij}} \right) \right.$$

By iteration, the right side of the equation is known, whereas a is unknown. So we must to solve $g(a) = \text{constant}$. It is necessary to multiply each term by -1 to get a gamma decreasing function in way to a dichotomy resolution.

The decreasing function is:

$$\left\{ \ln a - \frac{\Gamma(a)'}{\Gamma(a)} = \ln\left(\frac{\sum_{i=1}^{n} \omega_{ij}^m x_i}{\sum_{i=1}^{n} \omega_{ij}} \right) - \frac{\sum_{i=1}^{n} \omega_{ij}^m \ln x_i}{\sum_{i=1}^{n} \omega_{ij}^m} \right.$$

with : $\dfrac{\Gamma(x)'}{\Gamma(x)} = (\ln(\Gamma(x))' = \dfrac{\Gamma(x+h) - \Gamma(x)}{h}$

Now, we compute parameter b : $b = \dfrac{1}{a} \dfrac{\sum_{i=1}^{n} \omega_{ij}^m x_i}{\sum_{i=1}^{n} \omega_{ij}^m}$

1) Initial conditions

The initialization of the E/M algorithm must follow a few rules:

- The observation of the sample may provide information to initialize θ criteria starter. The mean and variance of the sample y are computed and used for the initialization of parameters for each analytical function. It leads to a better convergence of the E/M estimation.

- In the case of a mixture of two identical analytical models f_1 and f_2 (for example, a mixture of two Gaussian distributions), the weights must not be identical.

- The sample must respect the range of validity of the distribution function. In our case, the discrete values are strictly positive (sizes data, inter-arrival time). Before launching a simulation on a mixture, it is necessary to check the content of the sample. This operation is carried out with the following criteria:

$$\forall i \in [0,n], y_i > 0$$

2) Stop condition

The test of stop is to calculate the distance between $\Psi^{(m)}$ and $\Psi^{(m+1)}$:

Step (m) : $\Psi^{(m)} = (\pi_1^{(m)}, \pi_2^{(m)}, m_1^{(m)}, \sigma_1^{(m)}, \delta_2^{(m)})$

Step (m+1) : $\Psi^{(m)} = (\pi_1^{(m+1)}, \pi_2^{(m+1)}, m_1^{(m+1)}, \sigma_1^{(m+1)}, \delta_2^{(m+1)})$

The distance (accuracy) is represented by $\varepsilon = dist(\Psi^{(m+1)}, \Psi^{(m)})$:

$$\varepsilon = \sum_{i=1}^{n} (\Psi_i^{(m+1)} - \Psi_i^{(m)})^2 =$$

$$(\pi_1^{(m+1)} - \pi_1^{(m)})^2 + (\pi_2^{(m+1)} - \pi_2^{(m)})^2 + (m_1^{(m+1)} - m_1^{(m)}) + ...$$

If the distance is more than ε (constant) then we continue to compute with: "parameters step E = parameters step M".

3) Mixed distribution approval

E/M estimates the best distribution parameters without ensuring a perfect fit with the empirical distribution. The statistical tests (like Kolmogorov-smirnov, Cramer-von Mises, χ^2) compute the adequacy between the empirical distribution and the obtained theoretical model. If the statistical tests reject the theoretical distribution, we choose the mixture of analytical functions that minimizes the least square distance.

IV. Application

1) Obtaining a network sample

Many network tools can be used to get a sample of network packets with an efficient filtering. *Tcpdump* is proposed on Linux operating system. Some other tools such as network probes with *Netflow* will retrieve samples at the session level by detecting arrivals of new customers in the system.

2) Example of application with network packets inter-arrival times

Test results for an unknown sample of inter-arrival times are:

- *Sample size n=10000*
- Required accuracy (noted as distance) 10^{-4}
- Weight uniformly distributed *(0.33, 0.33, 0.33)*
- *Mean : 4.882E-4 Variance = 3.382E-7*

Table 3 - Distribution computed with E/M

E/M	Weight	Mean	Variance
LogNormal	*0.2069*	*8.9536E-4*	*7.1608E-7*
Normal	*0.00841*	*0.001778*	*1.6459E-7*
Exponential	*0.7846*	*3.6571E-4*	-

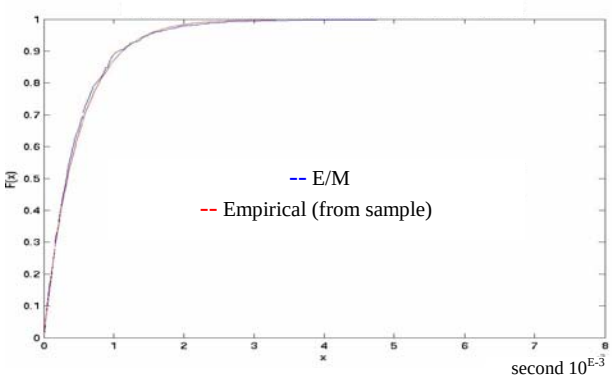

Fig. 2 - CDF of packet inter-arrival

3) Example of application with web objects size

Many network simulation studies deals with the size of web objects on a website. As an example, we study here the web objects size distribution of the Herald Tribune website http://www.heraldtribune.com/ on 08/01/2008. We use HTTrack (with default configuration) to get online objects of this site. Objects have been classified in different category: pictures, html and multimedia.

Table 4 - Online web objects of the Herald Tribune

Web objects	Sample size	Mean	Std dev
Pictures (gif, jpg, png …)	*2657*	*36.13E3 byte*	*47.22E 3*
Html, Xml	*8477*	*40.21E3 byte*	*29.76E 3*
Multimedia (mp3, wav)	*13*	*7.74E6 byte*	*6.65E6*

In more details, we now apply the E/M method to estimate the distribution of picture objects. First of all, we observe below an almost regular CDF with an exponential behavior.

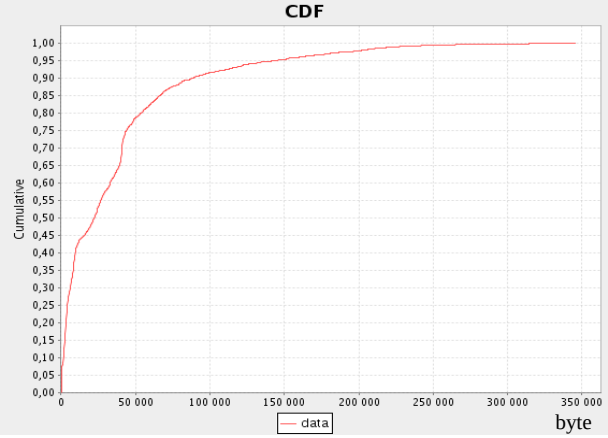

Fig. 3 – CDF Pictures objects

Below, the corresponding PDF of pictures object shows an exponential behavior with some irregularities above 40E3 and 45E3. This indicates it will take several distributions to define the distribution of image objects.

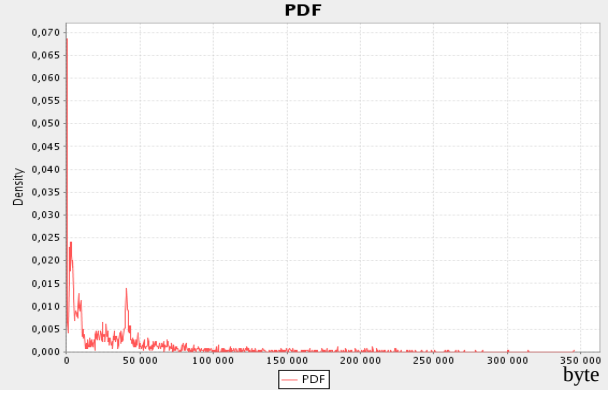

Fig. 4 – PDF Pictures objects

After several iterations with different distributions one of the solutions proposed by E/M is as follows, with an accuracy of 10E-5 (the calculation time is about 3 minutes with a standard PC).

Table 5 - Distribution computed with E/M

E/M	Weight	Mean	Variance
Exponential	*0.36367*	*4.56E3*	-
Exponential	*0.43928*	*3.8E9*	-
Normal	*0.06487*	*40.76E3*	*997.78E3*
Normal	*0.13216*	*35.89E3*	*148.9E6*

The mean square error is 479 with a Cramer von Mises test approval with a significance level of 1% and 5%. Fig. 5 presents the calculated CDF with E/M (red) well fits the CDF of web images (blue).

Then, this approach can be applied to define the html, xml, and multimedia size distribution.

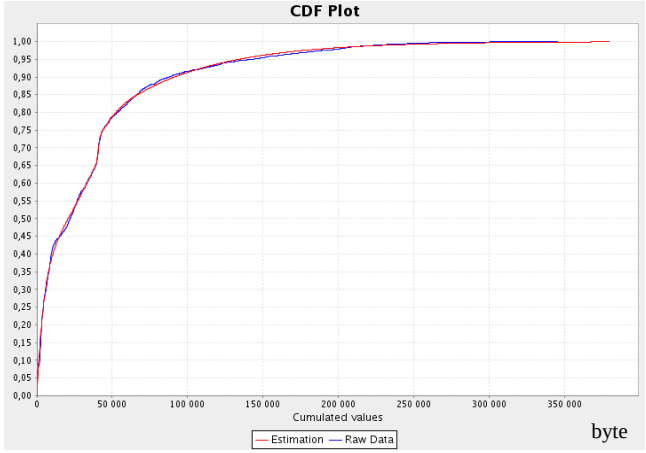

Fig. 5 – E/M Pictures objects

V. USING MIXED FUNCTION IN NETWORK SIMULATOR

In NS2, traffic generators are characterized by:
- a period of activity: ON
- a waiting period: OFF
- the size of packets
- the flow rate during the period ON

Traffic generators are derived from a C++ class (class *TrafficGenerator*). To incorporate a distribution mixed with the parameters calculated by E/M, *TrafficGenerator* can be extended and used in *TclObject* object [ ].

New multimedia formats and the increasing network flow rate (ADSL) require significant changes in the size of web online objects. Another use of mixed functions is to define the size of objects from a website: html page size, image and multimedia objects. In case of NS2, http server and client can use an appropriate mixed distribution to fit the actual behavior of web content.

A. References

[1] A. P. Dempster; N. M. Laird; D. B. Rubin, Journal of the Royal Statistical Society. Series B , Vol. 39, No. 1. (1977), pp. 1-38
[2] Z. Liu —N. Niclausse —C. J.-Villanueva —S. Barbier, Traffic Model and Performance Evaluation of Web servers, INRIA December 1999
[3] Frank Dellaert - The Expectation Maximization Algorithm - College of Computing, Georgia Institute of Technology Technical Report number GIT-GVU-02-20 - February 2002 February 2002
[4] Mah, B. An Empirical Model of HTTP Network Traffic, Proceedings of the IEEE INFOCOM 97, Kobe, April 1997, vol. 2, pages 592-600.
[5] D. Staehle, K. Leibnitz and P. Tran-Gian, Source Traffic Wireless Applications, Report N°261, June 2000, University of Würzburg
[6] A multimedia traffic modeling framework for simulation-based performance evaluation studies - Assen Golaup and Hamid Aghvamia, University of London - Computer Networks Volume 50, Issue 12, 24 August 2006, Pages 2071-2087
[7] Frost V & Melamed B, Traffic modelling for telecommunication network, IEEE Communication Magazine, 32(3), 70-81, 1994
[8] A. Klemm, C. Linderman, M. Lohmann- University of Dortmund -Traffic Modelling and Characterization for UMTS Networks
[9] Dinil Mon Divakaran, Hema A. Murthy, and Timothy A. Gonsalves - Traffic Modeling and Classification Using Packet Train Length and Packet Train Size - Springer-Verlag Berlin Heidelberg 2006
[10] J. Potemans1, B. Van den Broeck, Y. Guan, J. Theunis, E. Van Lil, A. Van de Capelle - Implementation of an Advanced Traffic Model in OPNET Modeler - Department of Electrical Engineering – ESAT-TELEMIC division - IEEE INFOCOM 2004
[11] Jin Cao, William S. Cleveland, Yuan Gao, Kevin Jeffay, F. Donelson Smith, Michele Weigle - Stochastic Models for Generating Synthetic HTTP Source Traffic - IEEE INFOCOM 2004