Journal of Universal Computer Science, vol. 14, no. 5 (2008), 731-744
submitted: 1/11/07, accepted: 28/2/08, appeared: 1/3/08 © J.UCS

Approximation to a Behavioral Model for Estimating Traffic Aggregation Scenarios

Alberto E. Garcia
(University of Cantabria, Santander, Spain
agarcia@tlmat.unican.es)

Klaus D. Hackbarth
(University of Cantabria, Santander, Spain
klaus@tlmat.unican.es)

Abstract: This article provides a comparison among different methods for estimating the aggregation of Internet traffic resulting from different users, network-access types and corresponding services. Some approximate models usually used as isolated methods are combined with a temporally scaled ON-OFF model with binomial approximations. The aggregation problem is solved using a new form of parameterization based on the composition of the source traffic accordingly to the concrete characteristics of the users, the accesses and the services. This is a new concept, called CASUAL, included within an overall network planning methodology for the design and dimensioning of Next Generation Internet.

Keywords: network planning, traffic modeling, traffic aggregation, network scenario
Categories: C.2.1, C.4

1 Introduction

Access networks allocate their resources depending on the class of the users, the technology and the services offered. Currently, the estimation of fundamental traffic parameters uses methodologies based on specific approximations example, for example, some solutions are based on Markov models, mainly on so-called MMPP (Modulated Markov Poisson Processes), for modeling the main services offered by the Internet network [Dolzer 00], [Serbest 99]. On the other hand, there are other solutions which use the results of simulation studies, considering the traffic sources and the individual network elements [Clark 01]. Some of these studies obtain a statistical analysis of the network behavior, using real observations of the traffic in the current Internet network, such as in [Ferrari 99]. Following the same proposal different new techniques and approximations appeared, for example the Network Calculus Theory, see [LeBoudec 02]. These approaches obtain solutions based on estimations, normally calculated using pessimistic suppositions about the real behavior of the network. In all the cases, the fundamental problem is IP traffic characterization, with all its additional problems. For example, the mixture of different flows into specific points of the network generates several different source traffic patterns modifying the individual behavior of the sources.

The effect of the traffic aggregation has been broadly studied, usually under different approaches, to establish models and methods for carrying out estimations based on statistical parameters such as mean value and variance. Usually, the studies focus on specific Internet services, as for example the World Wide Web (WWW), as a reference of the evolution of user's behavior with Internet utilization [Choi 99]. Other cases consider only the global statistics related to a complete network such as [Xu 05], without considering the statistical behavior of specific services and even omitting some of them [Ortega 06].

All of these models are individually applied to specific scenarios with restricted parameters. In contrast, this article proposes the combined use of well-known traffic models with the objective of covering global scenarios. For this purpose, we develop an optimal combination of the different methods to approximate the behavior, both individually and globally, of the traffic flows from different Internet services. Thus, the network planning process can use these approximations, obtaining the basic dimensioning parameters associated with different network scenarios.

This paper makes a brief exposition of the problem, explaining some solutions and introducing corresponding applications. Chapter two describes the Internet traffic nature, and its model using different temporal scales. Chapter three makes some corrections based on the variation of the behavior of the aggregated traffic, including some simple solutions. Finally chapter four uses all the explained concepts and solutions describing a concrete application for the estimation of traffic figures depending of the specific behavior of the users, the type of services or/and the access technology.

2 Internet traffic model based on temporal scale

Traditional methods for IP traffic characterization consider the complete protocol stack as a unique M/G/∞ system [Pieda 99], where, for each session, the service requests distribution follows a classical Poisson model. However, the intrinsic characteristics of burst Internet traffic and the traditional telephonic traffic are quite different. Internet traffic traces generally show a very high correlation over large observation periods, so the current proposed solutions modify the classical distributions with the so-called heavy-tailed distributions. Following this correction of the Internet traffic behavior, the Markov-based models adapt their operation to the auto-similarity behavior, using distributions with hyperbolic decay autocorrelation, as [Leland 93] and [Mondragon 01] explain. A typical example is to consider Poisson arrival processes but with service time distributions near to the auto-similarity. This is the case of the M/G/∞ system, which uses a Pareto or Weibull distribution function to model the service duration [Paxson 95]. Depending on the type of the application or service, some classical distributions allow the simplification of the use of heavy-tailed queues, such as using exponential distribution as in the particular case of VoIP traffic modeling.

However, using M/G/∞ systems is useless when the arrival process itself shows auto-similar characteristics. A study in [Liu 03] proposes a variant of this model, called G/G/c. In addition, these models need to estimate the traffic along the full temporal scale while Markovian models only consider the call level.

This paper proposes considering different reference points along the temporal scale, in a similar way to studies for voice transmission service modeling do, [Liu 04], [Riedl 00]. Accordingly to this idea, our model considers three differentiated temporal scales, shown in Fig. 1:

- Connection level: It models the behavior of the line between two consecutive connection petitions where a connection comprises the call establishment with the Internet Service Provider (ISP) using any of the different network accesses.
- Session level: An Internet session considers, for example, downloading web pages, a VoIP dialogue or a videoconference. This level models the time between two consecutive sessions within the connection and the duration of each session.
- Burst level: This is the lowest reference level. Each session generates a specific traffic pattern modeled by the inter-arrival time between objects or bursts belonging to the service.

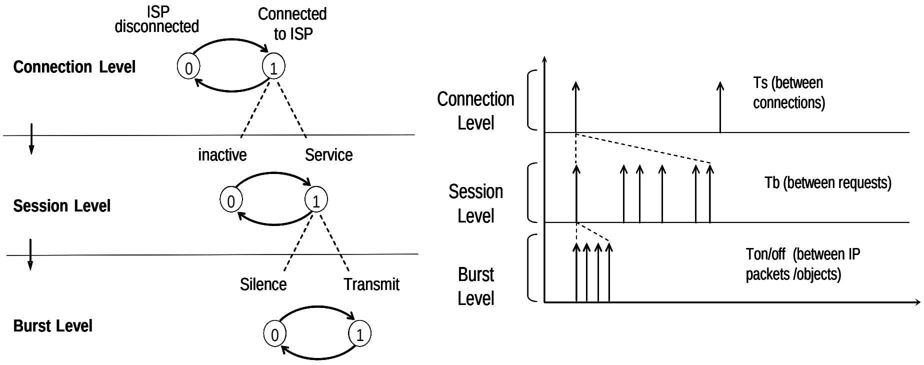


Figure 1: Time sequence showing the different requests inside an IP traffic flow and their correspondence with the multi-scale ON-OFF model.

The behavior of the traffic generated by a single source enables the consideration of each temporal level independently due to the great differences in the time scale of the different levels. Each level uses a two-state source directly related to the lower level. The relationship between each two-state source follows the temporal scale hierarchy.

Inside the connection level, the two states represent the connected / disconnected state respectively, *'0'* or *'1'*. This ON-OFF source models the time between two connections and its duration. In the same way, the ON-OFF source for the session level models the set of requests, transferences and waiting times (thinking times) during the download, for example, of a web page. The first state or *'0'* depicts the silence periods between consecutive sessions. The second state or *'1'* represents the set of requests and transfers associated with each session, voice dialogue, web page, FTP download, etc. The time between different sessions and the duration of each session are the basic parameters at this level. At the burst level, the ON-OFF source models each request individually. The *'0'* state represents the silence periods and the delays between requests, and the *'1'* state models the transfer of the associated objects for each request. In this case, the time between requests and the service time (for each request) are the fundamental parameters of the two-state model.

2.1 Mathematical approximations

Each level of the proposed model is solved recursively along its corresponding temporal scale, obtaining the associated traffic figures for each level (normally represented by the mean and variances of the bitrate). The connection level bitrates and their typical deviance (σ_{con}) are the results of applying the activation/deactivation probabilities of the associated source (P_1 and P_0 corresponding to the ON-OFF states). The activation probability for each state allows the calculation of the associated bitrates, identifying the permanence times in the active /deactivate state, and the data transmission rate associated with the silence level (V_{min}) and the session level (V_s). The calculation of the first two moments for each statistic is pondered using the correction factor γ, with 2 or 3 as typical values. Table 1 shows the mathematical model used for the traffic approximation under the proposed multi-scale model.

Connection Level	$V_{mc} = P_{0c}V_{\min} + P_{1c}V_s$ $\sigma_{con} = \sqrt{m_{con} - V_{mc}{}^2}$ $m_{con} = P_{0c}(V_{\min})^2 + P_{1c}(V_s)^2$ $V_c = V_{mc} + \sigma_{con}\gamma_c$
Session Level	$V_{ms} = P_{0s}V_{\min} + P_{1s}V_r$ $\sigma_{session} = \sqrt{m_{session} - V_{ms}{}^2}$ $m_{session} = P_{0s}(V_{\min})^2 + P_{1s}(V_r)^2$ $V_s = V_{ms} + \sigma_{session}\gamma_s$
Burst Level	$V_{mr} = P_{0r}V_{\min} + P_{1r}V_{\max}$ $\sigma_r = \sqrt{m - V_{mr}{}^2}$ $m = P_{0r}(V_{\min})^2 + P_{1r}(V_{\max})^2$ $V_r = V_{mr} + \sigma_r\gamma_r$

Table 1: Multi-scale ON-OFF model formulation.

2.2 Source traffic aggregation

The previous model only describes the behavior of single sources with a unique application and so does not express the multi-source traffic which really generates a traffic aggregate. Actually, the reference model for the aggregated traffic calculation follows three approximations: statistical multiplexing, models based on Modulated Markov Models and binomial approximations.

The statistical approximation, see [Wang 98], calculates the equivalent capacity associated with N multiplexed sources as:

$$C = \min\left\{\sum_{i=1}^{N}\rho_i + a\sqrt{\sum_{i=1}^{N}\sigma_i^2}, \sum_{i=1}^{N}R_i\right\} = \min\left\{\sum_{i=1}^{N}\rho_i + a\sqrt{\sum_{i=1}^{N}\rho_i\left(R_i - \rho_i\right)}, \sum_{i=1}^{N}R_i\right\} \qquad (1)$$

where ρ_i expresses the mean bit rate, σ_i the variance of the i^{th} source, R_i the maximum bit rate and α a normalization factor depending on the error ε:

$$a = \sqrt{-2\ln(\varepsilon) - \ln(2\pi)} \qquad (2)$$

The Modulated Markov sources consider the superposition of N sources, with an upper bound:

$$C = \sum_{i=1}^{N} C_i \qquad (3)$$

with C_i as the equivalent capacity of the i^{th} source.

An example of this type of solutions appears in [Yang 96], called D-MMDP (Discrete Time Markov Modulated Deterministic Process). This solution models the aggregation of ON-OFF sources using a discrete Markov chain as modulator process.

In agreement with this idea, a D-MMDP system with (M+1) states defines an arrival process with a bitrate controlled by the probability that two sources are in the active state. This probability is modeled using a binomial distribution function. However, assuming the Gaussian approximation (individual and independent normal sources) the required bandwidth is calculated using the first two moments of the aggregated traffic as:

$$C = \mu + \sigma\sqrt{\left(-\ln(2\pi) - 2\ln P_l\right)} \qquad (4)$$

where P_l is the loss probability. This result is only an approximation near to the upper bound, based on maximum values and obtaining an error free estimation of the bandwidth. The D-MMDP model uses an approximation based on the effective bandwidth with very similar results, near to the upper bound, but the temporally scaled ON-OFF model provides the lowest cost (computational cost) solution..

3 Estimation of the aggregation: Binomial approximation

Traditionally the traffic mixture is estimated using binomial distribution functions, providing the corresponding aggregation model, see [Valadas 2004], [Parkinson 02] and [Garcia 02]. Each source generates v_p bits/sec (active state) and the aggregated data bitrate is Nv_p in the case of CBR (Constant Bit Rate) sources. Using silence compression, the statistical multiplexing gain allows the calculation of the capacity of the server C as $C = \varepsilon v_p$, with ε as the number of equivalent CBR sources. If k independent sources are active with a probability p_{on}, (N-k) sources are deactivated with probability p_{off} = (1-p_{on}). On the other hand, there are N over k possibilities to make the selection of k elements distinct to *N*, as shown in [Garcia 02]. Therefore, the probability of *k* active sources follows a binomial distribution:

$$P(k)=\binom{N}{k}\cdot p_{ON}{}^{k}\cdot(1-p_{ON})^{N-k}=\binom{N}{k}\cdot\left(\frac{\alpha}{\alpha+\beta}\right)^{k}\cdot\left(\frac{\beta}{\alpha+\beta}\right)^{N-k} \quad (5)$$

Considering α and β as the corresponding transition probabilities between the active and inactive states for each source), the mean value gives $E(k)=N\cdot p_{ON}$ sources are active, and the mean data bitrate generated by *N* sources gives $E(v)=v_p\cdot N\cdot p_{ON}$. Consequently ε complies with $N>\varepsilon>E(k)$.

The overloaded condition appears when the server capacity is lower than the maximum data bitrate of the source. Under this condition the server buffer is full, producing packet loss.

This same situation could be reproduced using a Markov chain obtained using the M(N)/M/N binomial model. If the server provides the capacity for *N* sources with $s_{-1}<\varepsilon<s_0$, all the status from s_0 to *N* will produce overload, and the overload probability is calculated as the probability that the active sources surpass the available capacity:

$$P_{ol}=P(k>C)=\sum_{i=C+1}^{N}\binom{N}{i}\cdot p_{ON}{}^{i}\cdot(1-p_{ON})^{N-i} \quad (6)$$

C represents the server capacity (the backbone) in number of sources, *k* is the number of active sources and *N* is the total number of sources. In the special case of pure loss systems without buffers, the overload probability coincides with the loss probability of the system. The loss probability in these cases is calculated by the Engset loss formula and for a large number *N*, simply by the well-known Erlang loss formula.

The capacity of the server is modified using a corrector factor $\gamma(P_B,p_{ON},N)$ to consider the loss probability as a function of p_{ON} and the source number *N*. The value γ is a multiple of the standard deviation of the aggregated data flow:

The total value of the capacity for *N* binomial aggregated sources in bps is:

$$C=E(v)+\gamma(P_B,N,p_{ON})\cdot\sigma(v)=(N\cdot p_{ON}+\gamma(P_B,N,p_{ON})\cdot\sqrt{N\cdot p_{ON}\cdot(1-p_{ON})})\cdot v_p \quad (7)$$

The number of equivalent circuits for *N* sources in v_p units is:

$$N_{eq}=\frac{C}{v_p}=E(N)+\gamma(P_B,N,p_{ON})\cdot\sigma(N)=N\cdot p_{ON}+\gamma(P_B,N,p_{ON})\cdot\sqrt{N\cdot p_{ON}\cdot(1-p_{ON})} \quad (8)$$

The equivalent capacity of a binomial source in v_p units is:

$$v_{eq}=\frac{C}{C_{CBR}}=\frac{C}{v_p\cdot N}=\frac{N_{eq}}{N}=p_{ON}+\gamma(P_B,N,p_{ON})\cdot\frac{\sqrt{p_{ON}\cdot(1-p_{ON})}}{\sqrt{N}} \tag{9}$$

The limitation of the behavior of the equivalent capacity v_{eq} is:

$$\lim_{N\to\infty} v_{eq}=\lim_{N\to\infty}(p_{ON}+\frac{const}{\sqrt{N}})=p_{ON} \tag{10}$$

Binomial distributions provide the number of concurrent occurrences within the group of independent statistical processes. However, the characteristics of bursty traffic make it recommendable to use negative binomial functions to model the aggregation of several bursty sources. This happens in the case of the aggregation between consecutive temporal levels and hence the lower levels (burst and session) display bursty behavior, and the aggregation is solved using negative binomials. The connection level, depending on the service, can show the same bursty behavior, but in other cases, the aggregation suffers a smoothing and hence positive binomial functions should be used, this is mainly the case when the connection level is independent of the session and burst levels. For these reasons we propose the three-level structure using a mixed model, applying negative binomial functions in the session and burst levels, and positive binomial functions for the connection level. Following this concept the mean and variance of the number of active users are calculated as:

$$\begin{aligned} n_{ms}&=pop*\frac{P_1}{1-P_1}\\ v_s&=\mathrm{var}[n_{ms}]=pop*\frac{P_1}{(1-P_1)^2}\end{aligned} \tag{11}$$

where *pop* expresses the number of potential users and P_1 the probability of the active state in the connection level. The number of users producing active bursts is calculated as:

$$n_{mr}=(n_{ms}+\gamma*v_s)*\frac{P_2}{1-P_2} \tag{12}$$

with n_{mr} is the mean number of active users in the burst level, n_{ms} is the mean number of users with active sessions, v_s the corresponding variance, and P_2 the probability of the active state of the burst level. The variance of the active users in the burst level is calculated as:

$$v_r=\mathrm{var}\left[n_{mr}\right]=(n_{ms}+\gamma*v_s)*\frac{P_2}{(1-P_2)^2} \tag{13}$$

with γ between 1 and 3.

Using negative binomials, the calculation of the users within the session level follows the same method, obtaining the probabilities of the active state. The mean and variance pass towards the connection level directly. When the two lower levels are completed, the number of active connections is calculated following the same mechanism but using positive binomials. The bit-rate calculation uses the mean and the variance of the number of users:

$$\begin{aligned} V_{avg} &= n_{mr} * v_r \\ \mathrm{var}\left[V_{avg}\right] &= \mathrm{var}[n_{mr}] * v_r^{\,2} \end{aligned} \tag{14}$$

This value is the required bit-rate to offer a specific service to a group of users.

3.1 Estimation of the aggregation: Network Calculus

The analytic approximations shown in the last section provide results with a clear tendency toward linearity, and the simplifications use simple traffic figures. Specifically the Network Calculus theory [LeBoudec 02], uses these types of simple approximations, but provides general solutions. The corresponding models use values from real traffic observations and define bounded traffic figures based on these values, which represent arrival and service curves. The application of arrival curves allows a generic approximation towards the real behavior of the individual traffic sources over different types of network elements, see [Liebeher 01] and [Altman 02]. Network Calculus is composed of two types of methods. The first one is deterministic Network Calculus which considers only the bounded figures corresponding to the most pessimistic cases, and it omits the benefits associated with typical effects resulting from traffic aggregation such as statistical multiplexing of several sources over a single link. The second method refers to Statistical Network Calculus which considers the specific characteristics of the GoS (Grade of Service) over the traffic aggregation in the form of deterministic service curves and each individual flow in form of effective service curves, where an effective service curve shows the most probable bound towards which a specific traffic flow/service tends.

This type of curves establishes three possible approximations applied to estimate the GoS requirements:

- Maximal bit-rate estimation: If the service curve j shows a $S_j(t) = P_j t$ form, then the estimation obtains the maximal bound associated with the resources used (for each flow j).
- Average bit-rate estimation: If the service curve j shows a $S_j(t) = \rho_j t$ form, then the minimal bound associated with the resources used (for each flow j) is obtained. For example, the LBAP model (Linear Bounded Arrival Processes) considers each traffic source as a token bucket (b, ρ), with capacity b, bitrate ρ and service curve $A(t) \le b + \rho t \quad \forall t > 0$, see [Garroppo 01].
- Deterministic estimation: this method considers the best service curve that complies with the resources for each flow j while assuring the end-to-end delay conditions.

The literature shows multiple related examples and solutions. In [Lombardo 04] the estimation of service curves uses Markovian traffic models, called SBBP (Switched Batch Bernoulli Process). In [Riedl 03] the same calculation uses the token bucket concept and this idea was followed by the IETF. The Guaranteed Service

Class definition is the result and the calculation obtains an assured loss-free bandwidth and limited maximum delay, see [Schmitt 99] and [Schmitt 02]. This type of services generates a traffic pattern called TSpec (Traffic Specification). Two cascaded token buckets model traffic flow with the following parameters: a service bitrate *r* (bytes/sec), a bucket capacity *b* (bytes), a maximal bitrate *p* (bytes/sec), the maximal packet length *M* (bytes) and the minimal data unit *m* (bytes). Specifically the traffic figure for a Guaranteed Service follows an arrival curve *Tspec(r, b, p, M)* with the following form:

$$a(t) = \min(M + pt, b + rt) \tag{15}$$

The corresponding service curve follows the form:

$$c(t) = \begin{cases} 0 & t \le V \\ R(t-V) & t > V \end{cases} \tag{16}$$

with

$$V = \frac{C}{R} + D \tag{17}$$

R is the service bit-rate, and *C* and *D* depend on the type of the Server (for example for a *PGPS* server: *C=M* y *D=M'/c*, with MTU *M'* and link capacity *c*).

The required capacity for a determined link (assuring a maximal delay d_{max}) follows the expression:

$$R = \begin{cases} \dfrac{p\dfrac{b-M}{p-r} + M + C}{d_{\max} + \dfrac{b-M}{p-r} - D} & p \ge R \ge r \\ \dfrac{M+C}{d_{\max} - D} & R \ge p \ge r \end{cases} \tag{18}$$

Following RFC 2212 and RFC 2216, this methodology calculates the aggregation of several TSpec sources into the aggregation points of the network. *N* TSpec flows generate a new TSpec flow with the following form:

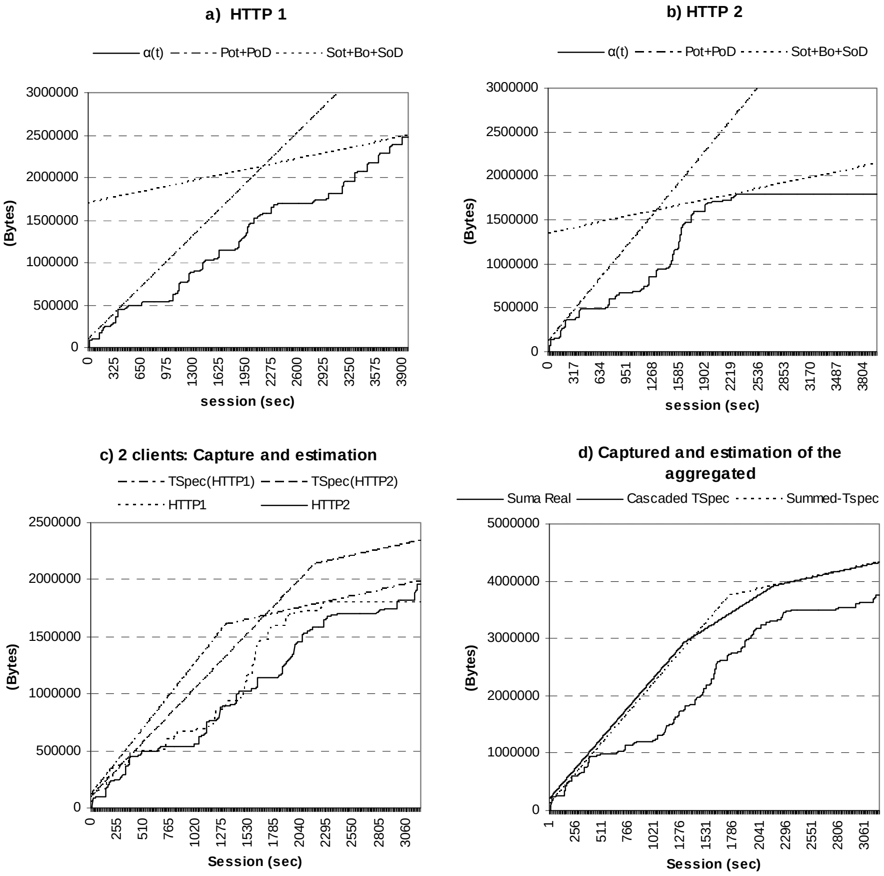
$$\sum_{i=1}^{n} \text{TSpec}(r_i, b_i, p_i, M_i) = \text{TSpec}\left(\sum_{i=1}^{n} r_i, \sum_{i=1}^{n} b_i, \sum_{i=1}^{n} p_i, \max(M_i)\right) \tag{19}$$

If we consider the aggregation of *N* input flows, the curve solves the dimensioning of shared resources systems, and the resulting traffic value is higher than the typical sum.

Network Calculus makes an exact calculation of the sum of TSpec based flows, and the resulting curve is lower or equal to the sum of individual TSpecs. The calculation considers the concatenation of (*N*+1) token buckets obtaining an arrival curve for each group of *N* flows. This calculation appears within the Network Calculus theory as a specific operation called "*min-plus convolution*" and represented by ⊗, see [LeBoudec 02]. The resulting curve is another TSpec curve called Cascaded-TSpec with the form:

$$\text{TSpec}\left(\sum_{j=k+1}^{n} p_j + \sum_{l=1}^{k} r_l, \sum_{l=1}^{k}(b_l - M_l) + M, \sum_{j=k}^{n} p_j + \sum_{l=1}^{k-1} r_l, \sum_{l=1}^{k-1}(b_l - M_l) + M\right) \tag{20}$$

Figure 2 shows an example of the application of arrival curves. The curves use real traffic traces corresponding to two different web clients but both located within the same access point. The methodology explained estimates the nearest TSpec arrival curve accordingly to the observations. The Cascaded-TSpec estimates the aggregated traffic arrival curve using the ideal individual TSpecs. The resulting empirical curve provides a clear reference for validating the results obtained by the analytical model based on the binomial approximation and the multi-scaling ON-OFF model. Figs. 2 a) and b) show the approximation to the flows HTTP1 and HTTP2 using two TSpecs, with slopes of 8 and 2 Kbps for each segment. The difference between each curve lies only in the duration of the session. If we want to compare these results with the analytical approximation, a 128 Kbps cable user generates a 9 Kbps web traffic flow (downloading a mean of 2 MBytes per session) aggregated within an Internet access point with 10000 users.



a) HTTP 1

b) HTTP 2

c) 2 clients: Capture and estimation

d) Captured and estimation of the aggregated

Figure 2: Example of arrival curves and aggregated traffic estimation

4 Application

The application of the proposed approximations and methodologies simplifies some of the procedures related to network planning and dimensioning. Following the development of these applications, the GIT (Telematic Engineering Group) of Cantabria University (Spain) developed a specific methodology for a network planning scenario generator, called CASUAL (Cube of Accesses / Services / Users for Free Assignment). The CASUAL methodology simplifies the aggregated traffic parameterization over the Internet access network. It considers each access network as a set of services with related characteristics from the point of view of the types of services, users and/or access network architecture. A specific network scenario is represented by a group of traffic flows classified along the three orthogonal axes of a cube (type of access, type of users, type of service). Accordingly to the characteristics of the users and the type of access, each cube cell models the traffic behavior of a specific service.

This conceptualization simplifies the traffic aggregation problem defining clear steps of the calculations, as Figure 3 shows. Each cell of the cube calculates the homogeneous flow aggregation individually. Increasing the complexity, an axis of the cube estimates the global aggregated traffic for a specific service, access or type of users. This is the case of heterogeneous flow aggregation. The combination of two or the three axes completes the variety of possible estimations (e.g. the entire Internet traffic corresponding to residential and SOHO DSL users).

Depending on the specific characteristics of each axis, the aggregation traffic estimation mechanism combines the explained approximations of the multi-scale ON-OFF model and the binomial approximation model from individual services. Their results allow the associated arrival and service curves to be obtained. Applying the CASUAL model, each cell inside the network scenario cube shows a specific arrival/service curve. There are two possibilities to calculate each curve, using analytical approximations or empirical data on real traffic traces. After that the Network Calculus solves the aggregation between cells. This requires an iterative process along each cube axis using the TSpec curve approximations or directly applying the associated arithmetic.

5 Conclusions

There are studies about the effect of traffic aggregation which show that establishing pure analytical methods for obtaining the corresponding estimation results is quite difficult. This paper explains some simple well-known solutions for this problem, and develops a new model which is based on a combination of some simple methods. The proposal provides a new model called CASUAL applying the methodology within a 3D conceptual representation of generic access network scenarios where a multi-scale ON-OFF model estimates the traffic associated with each individual service. The CASUAL model provides both the fundamental parameters of the different traffic flows and an aggregation criterion which depends on the related aggregation point. The classification applied in the CASUAL model allows the determination of the arrival curves and their probable relationship with specific “scheduling” mechanisms for the dimensioning of network access elements (e.g. a multiplexer or an edge router).

Additionally, the proposed methodology opens new directions in studying the adaptation of the simple ON-OFF model to a more specific behavior for new IP services. As currently the estimation of the aggregated traffic is based on pessimistic scenarios (associated with the binomial and Network Calculus considerations), new studies might provide decreased bound limits to model the corresponding situation more realistically and hence provide better performance in the corresponding network dimensioning.

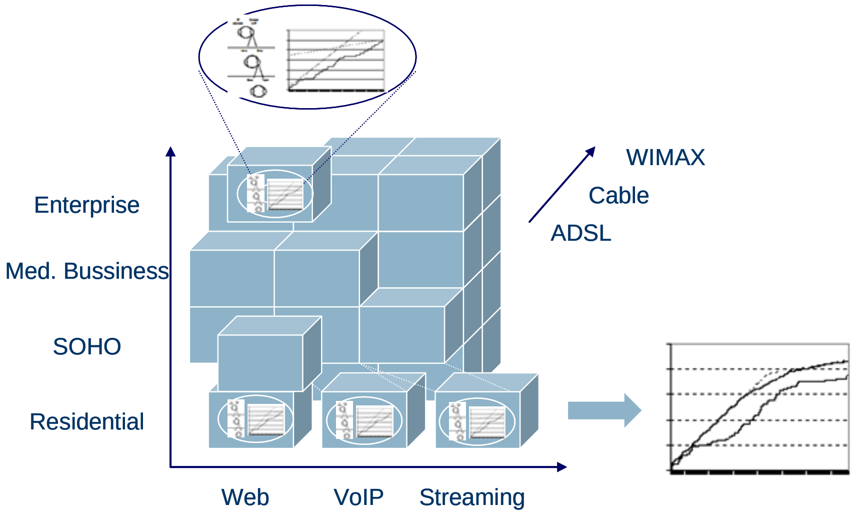


Figure 3: The CASUAL model and an application example

References

[Altman 02] Altman, E., Avrachenkov, K., Barakat, C.: "TCP Network Calculus: The case of large delay-bandwidth product". IEEE Infocom 2002, ISBN 0-7803-7476-2. New York 2002

[Choi 99] Choi, H.K., Limb, J. O.: "A Behavioral Model of Web Traffic" Conference of Networking Protocol 99' (ICNP 99'), Sep 1999.

[Clark 01] Clark, D., Lehr, W.: "Provisioning for Bursty Internet Traffic: Implications for Industry and Internet Structure", MIT Press, 2001

[Dolzer 00] Dolzer, Payer: "A simulation study on traffic aggregation in Multiservice Network" Proceedings of the IEEE Conference on High Performance Switching and Routing (ATM 2000), pp. 157-165, Heidelberg, 2000

[Ferrari 99] Ferrari, T.: "End to end performance analysis with traffic aggregation", Computer Networks Journal, Vol. 34, nº6, pp. 905-914, Amsterdam, 1999

[Garcia 02] Garcia, A. E., Hackbarth, K. D., Brand, A., Lehnert, R.: "Analytical Model for Voice over IP traffic characterization", WSEAS Transactions on Communications, vol. 1, pp. 59-65, 2002.

[Garroppo 01] Garroppo, R. G., Giordano, S., Niccolini, S., Russo, F.:"DiffServ Aggregation Strategies of Real Time Services in a WF2Q+ Schedulers Network," Lecture Notes in Computer Sciences, vol. 2170, pp. 481-491, 2001.

[LeBoudec 02] LeBoudec, J. Y., Thiran, P.: “Network Calculus: A theory of deterministic Queuing Systems for the Internet”. Ed. Springer Verlag LNCS 2050. July 2002

[Leland 93] Leland, W. E., Taqq, M. S., Willinger, W., Wilson, D. V.:"On the self-similar nature of {Ethernet} traffic," ACM SIGCOMM Conference on Communications Architectures, San Francisco, California, 1993.

[Liebeherr 01] Liebeherr, J., Patek, S., Burchard, A.:"A Calculus for End-to-End Statistical Service Guarantees," University of Virginia, Charlottesville, USA CS-2001-19, June 2001 2001.

[Liu 03] X. Liu, “Network Capacity Allocation for Traffic with Time Priorities”, International Journal of Network Management, Ed. Willey & Sons Ltd., pp. 411-417, 2003

[Liu 04] N. X. Liu and J. S. Baras, "Long-Run Performance Analysis of a Multi-Scale TCP Traffic Model", IEE Proceedings Communications, vol. 151, pp. 251-257, 2004.

[Lombardo 04] A. Lombardo, G. Morabito, and G. Schembra, "A Novel Analytical Framework Compounding Statistical Traffic Modeling and Aggregate-Level Service Curve Disciplines: Network Performance and Efficiency Implications," IEEE/ACM Transaction on Networking, vol. 12, pp. 443-455, 2004.

[Mondragon 01] R. J. Mondragon, D. K. Arrowsmith, J. M. Griffiths, and J. M. Pitts, "Chaotic Maps for Network Control: Traffic Modeling and Queuing Performance Analysis," Performance Evaluation, vol. 43, pp. 223-240, 2001.

[Ortega 06] J. M. Ortega, M. R. Menéndez, M. V. Román: “Diffusion and usage patterns of Internet services in the European Union”, ISIC 2006: the 6th Information Seeking in Context Conference, Sydney, Australia, 19-21 July, 2006

[Parkinson 02] R. Parkinson, "Traffic Engineering Techniques in Telecommunications," vol. 2005: Infotel System Corporation, 2002.

[Paxson 95] V. Paxson, S. Floyd. “Wide area traffic: the failure of Poisson Modeling”. IEEE/ACM Transactions on Networking, pp.226-244, June 1995

[Pieda 99] P. Pieda. “The dynamics of TCP and UDP interconnection in IP-QoS differentiated services networks”. Proceedings of the 3rd Canadian Conference on Broadband Research (CCBR), Ottawa, November 1999.

[Riedl 00] A. Riedl, M. Perske, T. Bauschert, and A. Probst, "Dimensioning of IP Access Networks with Elastic Traffic", Networks 2000, Toronto, Canada, 2000.

[Riedl 03] S. Sharafeddine, A. Riedl, J. Glasmann, and J. Totzke, "On Traffic Characteristics and Bandwidth Requirements of Voice over IP Applications," presented at 8th IEEE International Symposium on Computers and Communications, Kemer-Antalya, Turkey, 2003.

[Schmitt 99] J. Schmitt, M. Karsten, and R. Steinmetz, "Aggregation of Guaranteed Service Flows," 7th IEEE/IFIP International Workshop on Quality of Service (IWQoS'99), London, UK, 1999.

[Schmitt 01] J. Schmitt, M. Karsten, and R. Steinmetz, "On the Aggregation of Deterministic Service Flows," Computer Communications, vol. 24, pp. 2-18, 2001.

[Serbest 99] Y. Serbest, San-qi Li: “Unified Measurement Functions for Traffic Aggregation and Link Capacity Assessment”, IEEE Infocom '99:The Conference on Computer Communications, Volume 3, pp. 1522-1531, 1999.

[Valadas 04] R. Valadas, A. Pacheco, P. Salvador, A. Nogueira: “Markovian Modelling of Internet Traffic”, Proceeding s of HET-NET’04, Ilkley, West Yorkshire, UK. July 2004

[Wang 98] S. Wang, H. Zheng, J.A. Copeland,: “Video Multiplexing with QoS Constraints” in IEEE SPIE Conference on Internet Routing and QoS, (1998), 81-91.

[Xu 05] K. Xu, Z. Zhang, S. Bhattacharyya: “Profiling Internet BackboneTraffic: Behavior Models and Applications” SIGCOMM 2005

[Yang 96] J. Ni, ,T. Yang, D.H.K.Tsang: “Source Modelling, Queuing Analysis and Bandwidth Allocation for VBR MPEG-2 Video Traffic” in ATM Networks. IEE Proceedings on Communications, 143 (4). 197-205.