Internet Erlang formula: survey on relation between demand, capacity and performance in IP network

Seferin T. Mirtchev

The internet traffic is the result of interaction among millions of users, hundreds of heterogeneous applications, and dozens of sophisticated protocols. The internet traffic is much more complex than the telephone traffic and the mix of applications that produce it continues to vary widely over time. Selecting the appropriate traffic model can lead to successful design of computer networks and accurate capacity planning. In the last years, increasing interest in developing models and methods of classical queuing systems (especially Erlang formulae) for studying Internet network has led to many extensions of previously existing results. For this new models and methods it is used phrase "Internet Erlang Formula". In this article, it is presented a survey of relation between demand, capacity and performance in the Internet network by the Erlang formulae. It is shown that Erlang C formula has in fact much more general application in the Internet.

Формула на Ерланг за Интернет: Обзор на връзката между натоварването, ресурсите и характеристиките в IP мрежи (Сеферин Т. Мирчев). *Трафикът в интернет е резултат от взаимодействието между милиони абонати, стотици хетерогенни приложения, както и десетки сложни протоколи. Трафикът в интернет е много по-сложен от телефонния трафик и множеството от приложения, които го генерират, продължават да се променят в широки граници в течение на времето. Избирането на подходящ модел на трафика може да доведе до успешно проектиране на телекомуникационните мрежи и до точно планиране на ресурсите. През последните години нарастващият интерес към разработването на модели и методи на класическите телетрафични системи (особено формулите на Ерланг) за изучаване на интернет мрежата доведе до много разширения на вече съществуващи резултати. За тези нови модели и методи се използва израза "Формула на Ерланг за интернет". В тази статия е представен обзор на връзка между натоварването, ресурсите и характеристиките в интернет мрежата чрез формулите на Ерланг. Показано е, че С формулата на Ерланг има в действителност много по-общо приложение в интернет.*

The future Internet will be a network of data centres - Jim Roberts [24].

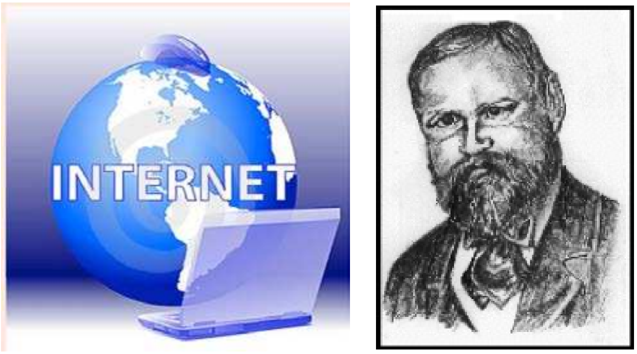
1. Introduction

It becomes essential to design the network appropriately, facing the steadily growing services that an IP-based network may support. A future IP-based network is expected to allow service differentiation to be efficiently managed. This depends on the cost of introducing functionality allowing differentiation compared with the gains that can be achieved.

A report on the National Science Foundation (NSF) program Future Internet Design (FIND) concluded that an important open issue for future research is the identification of "Erlang formulae" for the Internet (Fig.1) [4].

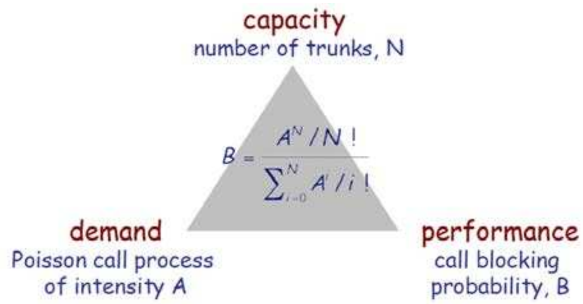
The Erlang loss formula is used in engineering the telephone network. It gives the probability of call blocking on a trunk group as a function of the number of trunks and the offered traffic. It is the prototype of the relation between demand, capacity and performance whose understanding is essential for cost effective network engineering (Fig.2).

Identifying Internet Erlang formulae by the Future Internet Design report is important. Currently, the Internet is engineered more by the use of pragmatic rules of thumb than by applying mathematical models like that which led to the Erlang formula. This leads not only to inefficiencies through inappropriate sizing but also to misconceptions about the effectiveness of traffic controls and their ability to support differentiated services. Much of the necessary research has already been performed and the main problem is a lack of awareness of known results and their implications. Crucially, this research considers Internet traffic in terms of stochastic processes of packet, flow and session arrivals.



Source: Roberts, An Erlang formula for the Internet. Slides [23].

Fig.1. Current research challenge: Internet Erlang Formula - lack of up-to-date applicable tools.



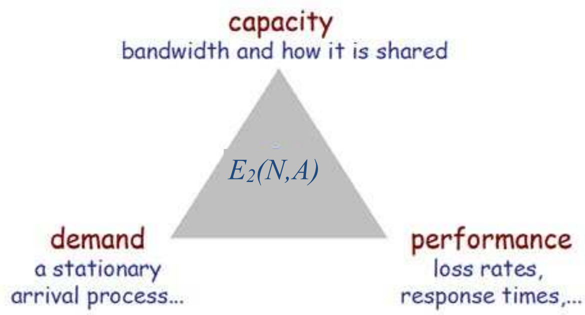
Source: Roberts, The Cloud is the future Internet: How do we engineer a cloud? Keynote, Slides. [24].

Fig.2. Three-way relation between demand, capacity and performance in the telephone network: the Erlang loss formula.

The assuring of QoS relies on understanding the traffic performance relationship. Surprisingly, the essential demand – capacity - performance relation in IP network turns out to be none other than the so-called Erlang delay formula (Fig.3)! This is clearly a surprising result since it is commonly believed that Internet traffic is so complex that it is practically impossible to characterize performance in a simple way. A very large number of different applications use the Internet and its traffic characteristics change continually as new applications gain popularity. Furthermore, while it is well established that telephone calls arrive as a Poisson process, the arrival process of IP datagrams has been shown to exhibit much more complex, self-similar or fractal-like behaviour [20].

In the last years, increasing interest in developing models and methods of classical queuing systems (especially Erlang C formula) for studying Internet network has led to many extensions of previously existing results. For this new models and methods it is used phrase “Internet Erlang Formula”. Vint Cerf listed in 2007 seven research problems concerning the Internet. The second problem is: an Internet Erlang Formula.

In [14] is presented a robust and efficient algorithm for evaluating multi-service multi-rate queuing systems, including finite buffer systems and loss systems, based on Erlang formulae.



Source: Roberts, The Cloud is the future Internet: How do we engineer a cloud? Keynote, Slides. [24].

Fig.3. Three-way relation between demand, capacity and performance in the IP network: the Erlang delay formula.

In [5] is addressed the idea of utilization of Erlang formulas in asynchronous ATM and IP networks. Through the Erlang C formula, they can estimate the probability of delay, which usually occurs in IP networks. Erlang Models in IP Network gives the opportunities to monitor the Quality of Service parameters. The simplicity of Erlang formulas can be their strong advantage against other methods for traffic description in asynchronous networks, but the future research is necessary in this field.

In [8] is dealt with the possibility of the Erlang B and Erlang C formula utilization in Next Generation Networks. Based on the common properties of synchronous and asynchronous networks it is possible the utilization of Erlang formulas also for asynchronous networks.

A model based on queuing theory for service performance in cloud computing is presented in [27]. The model is based on event-driven simulation and possesses scheduling capabilities for heterogeneous and non-dedicated clouds. It is based on a classical open network M/M/m. The results demonstrate the usefulness of the presented simulation models for the design of cloud computing systems with guarantees of QoS.

In [6] is concerned of Erlang formulas practical approval and their usage with video flow transfer in IP networks.

The important issues of the network planning process for multi-service IP networks are discussed in [21]. The presented ideas and concepts provide a first framework for develop a uniform view of the overall planning process. In order to do so, IP QoS mechanisms are categorized and a systematic approach for classification and modelling of Internet traffic is suggested.

In this article, it is presented a survey of relation between demand, capacity and performance in the Internet network by the Erlang formulae. It is shown that Erlang C formula has in fact application in the Internet that is much more general.

2. Internet traffic

The internet traffic is the result of interaction among millions of users, hundreds of heterogeneous applications, and dozens of sophisticated protocols. The technical components of the Internet are complex in themselves, and they are augmented by a general unpredictability and diversity of the human components.

The traffic, in its turn, is a combination of application mechanisms and users' behaviour, including attitude towards technology, life habits, and other intangible cultural phenomena. Such a mix of heterogeneous components is made even more difficult to understand by the fast evolution of technologies and the rapid rise and fall of new stars among applications [18], [26].

The internet traffic is clearly much more complex than telephone traffic and the mix of applications that produces it continues to vary widely over time [23].

Selecting the appropriate traffic model can lead to successful design of computer networks and accurate capacity planning. The more accurate the traffic model is, the better system quantified in terms of its performance. Successful design leads to enhancement the overall performance of the whole of network. In literature, there are plenty of traffic models proposed for understanding and analyzing the traffic characteristics of computer networks.

The term "teletraffic theory" originally encompassed all mathematics applicable to the design, control and management of the public switched telephone networks (PSTN): statistical inference, mathematical modelling, optimization, queuing and performance analysis. Later, its practitioners would extend this to include data networks such as the Internet, too. Internet engineering, an activity that includes the design, management, control and operations of the global Internet, would thus become part of teletraffic theory, relying on the mathematical sciences for new insights into and a basic understanding of modern data communications.

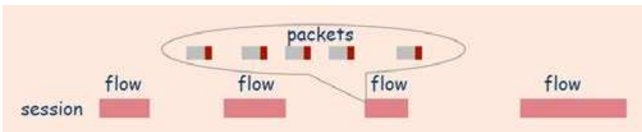
The voice traffic has the property that it is relatively homogeneous and predictable, and, from a signalling perspective, spans long time scales. In contrast to the voice traffic, the data traffic is much more variable, with individual connections ranging from extremely short to extremely long and from extremely low-rate to extremely high-rate. These properties have led to a design for data networks in which each individual data „packet" or „datagram" transmitted over the network is forwarded through the network independently of previous packets that may have been transmitted by the same connection.

However, as the voice traffic turns out to differ drastically from the data traffic, so too do the underlying mathematical ideas and concepts. The relevant mathematics for PSTN is one of limited variability in both times: traffic processes are either independent or have temporal correlations that decay exponentially fast and in space, i.e., the distributions of traffic-related quantities have exponentially decaying tails. However, for the data networks, the mathematics is one of high or extreme variability. Statistically, temporal high variability in traffic processes is captured by long-range dependence, i.e., autocorrelations that exhibit power-law decay. On the other hand, extreme forms of spatial variability can be described parsimoniously using heavy-tailed distributions with infinite variance.

It turns out that power-law behaviour in time or space of some of their statistical descriptors often cause the corresponding traffic processes to exhibit fractal characteristics.

2.1 Packets, flows, sessions

Though the Internet protocols only deal with packets (datagrams), it is important for network engineering to recognize that these belong to "flows" which in turn are components of "sessions" (Fig.4). For present purposes, a flow is defined as the succession of packets handled by a given link that relate to one instance of some application [23].



Source: Roberts, An Erlang formula for the Internet. Slides [23].

Fig.4. Packets, flows and sessions.

Flows are basically of two types:

- elastic flows, download documents as fast as possible by adjusting their packet emission rate (e.g., through TCP) to use all available capacity,
- streaming flows, typically based on UDP, send packets as and when they are generated by the audio or video codec.

Each flow is characterized by some peak rate. For elastic flows, this peak rate is typically due to the access network, the server capacity or some other bottleneck on its path. For streaming flows, it is the peak rate of the codec.

A session is loosely defined as a set of flows that are related in some way. The session is in fact better defined by the requirement that any two sessions relate to independent activities, usually by distinct users. In general, sessions cannot be identified as such but, for the same reason a large population generates telephone calls as a Poisson process. It is natural that the arrival epochs of sessions using a considered network link are Poissonian.

Poisson session arrivals have been observed experimentally by Paxson and Floyd for some kinds of session that can be identified in Internet trace data [20]. The same authors conformed that flow and packet arrivals are anything but Poisson, on the other hand, and even exhibit the extreme correlation of self-similar processes. These characteristics are in fact only of secondary importance.

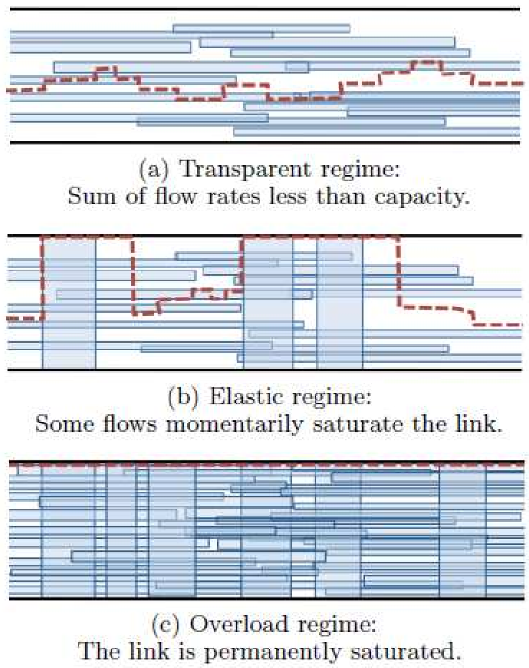
2.2 Traffic regimes

Three traffic regimes that help to understand the scope for meeting performance requirements are illustrated in fig.5. Assuming flows have a constant peak rate, they can be represented simply in the figure as rectangles where height is peak rate and area is size. They share the bandwidth of a link represented by the parallel black lines. The dashed line represents the instantaneous overall input rate.

In the transparent regime, all flows have a relatively low peak rate and demand is such that, with very high probability, the sum of rates is less than link capacity. In this regime, packet loss is negligible and delays are tiny. The elastic regime occurs when some flows have a peak rate that momentarily saturates the link. The buffer is then sure to overflow and flows suffer loss and delays that can be significant. Periods of transparency alternate with periods of saturation. Performance may be considered satisfactory if degradation can be confined to the high rate elastic flows. The overload regime occurs when demand (flow arrival rate × mean flow size) exceeds link capacity. Performance is then typically very bad for all flows so that this regime needs to be avoided by appropriate traffic engineering.

2.3 Bandwidth sharing

The flows that are concurrently active on a given link may be said to share its bandwidth. In the transparent regime, there is capacity to spare and every flow realizes its peak rate. A simple FIFO buffer is then sufficient to resolve contention between packets from distinct flows. In the elastic regime, some flows must reduce their rate. TCP normally realizes the necessary adjustment resulting in each flow receiving an allocation that is approximately max-min fair [15]. With max-min fairness, only the high peak rate flows are constrained to reduce their rate. The others maintain their rate and suffer negligible loss, as if they experienced the transparent regime.



Source: Bonald, Roberts. Internet and the Erlang Formula [1].

Fig.5. Link occupancy regimes: the rectangles represent flows (minimum duration - peak rate), the dashed line traces the sum of realized rates.

The bandwidth sharing in the Internet can be controlled to some extent by QoS mechanisms like Diffserv. It is possible, for example, to consider certain classes of traffic with priority, ensuring they experience a transparent or elastic regime even when other classes are in overload. It remains difficult to control performance, however, since this depends critically on the amount of traffic in each class and on the peak rates of its flows, parameters that are largely uncontrollable.

3 The Erlang formulae

3.1 The Erlang loss formula

The Erlang loss formula (The Erlang B-formula) gives the probability of call blocking B when N trunks serve offered traffic A:

$$B = E_1(N, A) = \frac{A^n}{N!} \Bigg/ \sum_{j=0}^{N} \frac{A^j}{j!}, \quad (1)$$

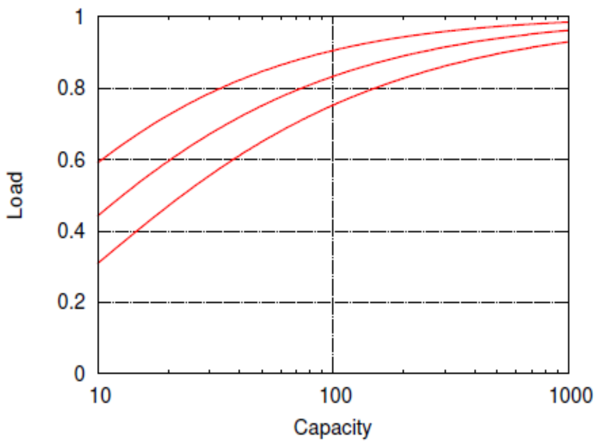
where A is the product, call arrival rate × mean call holding time. The formula is derived from a mathematical model that makes a number of assumptions about telephone switching and the nature of traffic, including the following: call arrivals constitute a stationary Poisson process; calls are blocked if and only if all trunks are busy; blocked calls are cleared.

Some assumptions are realistic, being based on observable reality, others are merely convenient, enabling a simple formula when a more accurate model would be intractable. For instance, to assume Poisson arrivals is realistic over relatively short timescales (less than 1 hour, say) when a large user population generates calls. The second assumption is realistic for present day switches though in Erlang's days, all trunks could not usually be tested by every call. It is convenient to suppose blocked calls are cleared though in practice callers usually make repeat attempts. This means the B-formula is good for dimensioning (for low blocking) but not for analysing performance in overload. It is well known now that Erlang's B-formula is insensitive to the distribution of the call holding time. The blocking probability depends only on the simple average measure of offered traffic A. This insensitivity explains why the formula remains a precious dimensioning tool today despite significant changes in usage during the past decades. It also underlies much of network engineering practice since it informs us that the essential measure of demand to be monitored and forecast is offered traffic A, even when the above assumptions may not be perfectly reasonable.

The Erlang formula reveals the important scale economies phenomenon of networking (Fig.6): achievable load $A(1\text{-}B)/N$, for a given target blocking probability B, increases towards 100% with trunk group size N.

The Erlang-B model uses traffic intensity and Grade of Service (GoS) to determine the number of trunks in circuit-switched networks. In [28] is proposed a new measurement scheme to translate VoIP packet-switched network bandwidth into the maximum call load. With this new metric, the Erlang-B model is applicable to VoIP.

The Erlang formula is not a universal tool that solves all telephone network-engineering problems. It have a very important emblematic role through the lessons it provides about traffic modelling and the confidence it inspires that we are indeed able to relate demand, capacity and performance for a construction as complex as the telephone network.



Source: Bonald, Roberts. Internet and the Erlang Formula [1].

Fig.6. Admissible load A/N as a function of trunk group capacity N for blocking probabilities B = 5%; 1% and 0.1% (from top to bottom).

3.2 Generalizations

The following two generalizations are actually more significant for the Internet than for the telephone network. First, the packets do not occur as a Poisson process but they are components of "sessions". We assume sessions occur as a Poisson process. Under very general assumptions about the distributions of the number of packets per session, individual packet holding times and silence intervals and their correlation, the packet blocking probability is still given by Erlang B [3]. Under this Poisson session model the packet arrival process would even be self-similar if the number of packets per session had a so-called heavy-tailed distribution.

The second generalization is the Erlang Multirate Loss Model and the well known as Kaufman-Roberts recursion. This model relates to heterogeneous packet types distinguished by the number of trunks each packet requires throughout its holding time. A packet requiring c trunks is blocked and cleared if the number of free trunks on its arrival is less than c. Traffic for this type of packet is defined by the product, call arrival rate × mean call holding time × c. Consider m types of packets, class-i calls requiring c_i trunks and offering traffic a_i. The probability that N trunks are occupied is proportional to $f(n)$, given by the following simple recurrence relation (Kaufman-Roberts) [16], [25]:

$$f(n) = \frac{1}{n}\sum_{i=1}^{m} a_i\, f(n-c_i), \tag{2}$$

for $n = 1; : : : ;N$, with $f(0) = 1$ and $f(n) = 0$ if $n < 0$.

The blocking rate of class-i packets then follows as the probability that more than $N - c_i$ trunks are occupied. The formulas are valid under the same general assumptions as Erlang B, including the Poisson session model.

Kaufman-Roberts recursion becomes the springboard of teletraffic modelling for QoS assessment in the multidimensional traffic environment of contemporary communication networks, because of the accurate and efficient call blocking probability calculation. This recurrence is the most desirable computational feature of a teletraffic model in order to cope with the high bandwidth capacities of network links.

In [19] is used the Erlang Multirate Loss Model for elastic traffic. They are considered a single-link loss system of fixed bandwidth capacity, which accommodates K service-classes of Poisson traffic with different bandwidth-per-call requirements and provided a recurrent formula for the calculation of the link occupancy distribution. Based on it they are determined call blocking probabilities, link utilization and average number of calls in the system. The accuracy of the proposed formula is verified by simulation and is found to be quite accurate.

In [17] relying on the Erlang Multirate Loss Model is studied and proposed many efficient teletraffic loss models as the so called Connection-Dependent Threshold Model that comprises call retries, and of the Batched Poisson Multirate Loss Model, in which the input process is Batched Poisson. The considered call admission policy is the complete sharing policy, as well as the bandwidth reservation policy, suitable for QoS guarantee.

3.3 The Erlang delay formula

The Erlang's C-formula is derived initially to dimension the number of operators managing a call centres. The formula gives the probability $E_2(A;N)$ that a caller must wait when N operators receive offered traffic A, assuming $A < N$.

The Erlang delay formula is valid under the following assumptions: Poisson arrival process, exponential service times, identical servers, full accessibility, calls are served in arrival order and unlimited queue length. When all servers are busy, an arriving customer joins a queue and waits until a server becomes idle. The probability that an arbitrary arriving customer has to wait in the queue is equal to the proportion of time all servers are occupied. The waiting time is a random variable denoted by t_w. For an arbitrary arriving customer we have [13]:

$$P(t_w > 0) = E_2(N,A) = 1 - \sum_{j=0}^{N-1} P_j =$$

$$= \frac{\frac{A^N}{N!}\,\frac{N}{N-A}}{\sum_{j=0}^{N-1}\frac{A^j}{j!} + \frac{A^N}{N!}\,\frac{N}{N-A}} \quad when \quad A < N, \tag{3}$$

where A is offered traffic (Erl),
N *is* number of servers,
$P(t_w>o)$ *is* probability of waiting for service.

The formula has several names: Erlang's C–formula, Erlang's second formula, or Erlang's formula for waiting time systems. In [7] is dealt with calculation of important parameters of the Call Centre using the Erlang C formula and the results have been verified through simulations. In [10] is modelled the cloud system using Erlang C formula and four different cloud utility models is presented. The presented models are simulated in order to characterize the performance.

Erlang C formula has in fact much more general application in the Internet.

4. PERFORMANCE OF FAIR SHARING

4.1 Assumptions

Max-min fair sharing between concurrent flows is a realistic assumption if routers impose per flow fairness [12]. It is just a convenient assumption if we must rely on end-system compliance in implementing congestion control.

The further convenient assumption that even streaming flows adjust their rate as necessary to respect fairness and that they preserve their volume (i.e., like elastic flows, if their rate is reduced below their peak rate as they last longer). This simplifies modelling and the resulting approximation is accurate as long as the probability a streaming flow would suffer loss is small. As this should be an objective of dimensioning, this is similar in effect to Erlang's "blocked calls cleared" assumption.

Another realistic assumption is that the sessions occur as a Poisson process and that on any given link, their flows occur singly in an alternating sequence with think times. Farther assumption is that the concurrent flows of the same session are considered as one for bandwidth sharing. The flow sizes and the think time durations are generally distributed and can be correlated. The number of flows in the same session has a general distribution.

4.2 An equal peak rate

Consider a link of capacity C offered traffic A, in bit/s, and suppose each flow has the same peak rate c. Under the above traffic and sharing assumptions, the number of flows concurrently active x behaves like the number of customers in a multi-server processor-sharing queue [11]. In particular, when C/c is an integer and under the stability condition $A < C$, the proportion of time a flow suffers congestion (in the sense that $xc \geq C$) is given by the Erlang C-formula, $E_2(C/c, A/c)$.

Thus, under the assumption of equal peak rates, the Internet Erlang formula is precisely the Erlang delay formula (3). Note that, due to the insensitivity property of the processor sharing discipline, this formula is valid for the considered Poisson session traffic model with general flow size distribution while, for the FIFO queue envisaged by Erlang, it is only valid for Poisson flow arrivals and exponentially distributed call-holding times.

4.3 A mixture of peak rates

In practice, flows in the Internet have a wide range of peak rates. Assume for convenience that the number of possible peak rates is limited to m, that these rates are $c_1, c_2, \ldots, c_m$ in increasing order and that flows of rate c_i offer traffic a_i. Let the number of active flows of class i be x_i and consider a time interval of congestion, where

(4) $$\sum_{i=1}^{m} x_i a_i \geq C.$$

With max-min fair sharing, the congestion is confined to flows of high rates. Specifically, there is an index j such that flows of classes j to m reduce their rate to a "fair rate" r satisfying $c_{j-1} < r \leq c_j$, while flows of classes 1 to $j-1$ maintain their peak rate. The precise value of r is such that the sum of realized flow rates is equal to C. It turns out that performance evaluation under max-min fairness is now analytically intractable.

To make progress it is useful to make a further convenient assumption [2]. They assume that sharing is "balanced fair". In the present context, this means that when the system is in congestion all flows see a rate reduction, the reduction of flows of class i being approximately proportional to c_i. It has been proved that balanced fairness is the only policy for which it is possible to derive explicit performance results for general rate and demand vectors, $\{c_i\}$ and $\{a_i\}$, and that these results do not depend on any more detailed traffic characteristics.

Under balanced fairness, the probability that the total flow rate $\Sigma\, x_i c_i$ is equal to n, assuming C and the $\{c_i\}$ are integers, is proportional to a function $f(n)$ that satisfies the recurrence relations:

(5) $$f(n) = \begin{cases} 1/n \sum_{i=1}^{m} a_i f(n-c_i) & if \quad n < C, \\ 1/C \sum_{i=1}^{m} a_i f(n-c_i) & if \quad n < C. \end{cases}$$

Comparison of (5) and (2) reveals a quite remarkable parallel between loss systems on one hand and balanced fair systems on the other that in fact extends well beyond the results we are able to summarize here.

The function $f(n)$ can be used to derive a number of performance parameters like the congestion rate, the probability input rate $\Sigma\, x_i c_i$ exceeds capacity C. Importantly, it has been verified by simulation that many performance results derived under the convenient balanced fairness assumption closely approximate those obtained for max-min as well as other fairness criteria. The balanced fairness assumption is then reasonable as well as convenient.

4.4 Throughput and congestion

It has been shown in particular that the expected throughput of a flow of peak rate c_i is approximately equal to the minimum of c_i and C-A where $A = \Sigma\, a_i$ is overall demand. The fact that C is large and utilization A/C is typically not more than 80% explains why Internet backbone links rarely impact perceived performance. They are in the transparent regime since no flows are able to saturate the residual free capacity, i.e., $c_i << C - A$ for all i.

If the $\{c_i\}$ and $\{a_i\}$ are known, it is possible to use the recurrence relations (5) to dimension links to ensure a low congestion probability. This means the link stays with high probability in the transparent regime. This is only satisfactory however if the c_i are all guaranteed to be relatively small. Otherwise, the variance of the offered traffic is high so that congestion can only be avoided by limiting mean load to a small fraction of capacity.

While the transparent regime prevails in the Internet, justifying the Gaussian traffic models [9], we anticipate that the elastic regime will become more common as flow rates increase, notably between well-connected servers and data centres.

5. Internet Erlang formula

It is identified an explicit performance relation that, like the Erlang loss formula, involves only link capacity and expected demand [1]. For all practical purposes, this relation is an upper bound on the probability of congestion and can be used to dimension Internet links.

5.1 Performance criteria

It is considered the demand-capacity-performance relation with the following choice of performance criterion. It is supposed a network provider seeks to limit the degradation suffered by streaming flows of peak rate no greater than c. Specifically, assuming max-min fair bandwidth sharing, the proportion of time P_c any currently active flow of rate c would suffer loss or have to reduce its rate should be less than some target ε. It is referred to P_c as the rate-c congestion probability. The dimensioning objective, given demand A, is to provide sufficient capacity C such that $P_c < \varepsilon$.

Under max-min fair bandwidth sharing, P_c is simply the probability that the fair rate is less than c. According to the model of the previous section, with some traffic mix defined by $\{c_i\}$ and $\{a_i\}$, we have:

(6) $$P_c = \Pr\left(\sum_{i=1}^{m} x_i \ \min\ (c_i, c) \geq c\right).$$

5.2 A congestion probability bound

Unfortunately, max-min sharing is intractable: it is not possible to calculate the probability distribution of the fair rate. Moreover, any formula that depends on precise knowledge of $\{c_i\}$ and $\{a_i\}$ is hardly useful in practice since these data are not usually available. On the other hand, as shown below, there is a formula that, for all practical purposes, constitutes an upper bound on the congestion probability P_c and is valid for any traffic mix [1].

The test rate c divides the flows into two categories: "low rate-flows" that have a peak rate less than c and "high-rate flows" that have a peak rate greater than or equal to c. For given overall traffic A, P_c tends to increase either as the peak rate of low-rate-flows increases to c or as the peak rate of high-rate-flows decreases to c. In particular, the rate-c congestion for any traffic mix is upper bounded by that for traffic concentrated on rate c alone. This statement is not strictly true and precisely qualifying the sets of parameters and conditions for which it is remained an open research challenge. But, intuitive arguments, some mathematical demonstrations and the results of simulations lead to the conviction to calculate P_c assuming uniform rate-c flows constitutes a valid, conservative approach for link dimensioning.

Reducing the rate of high-rate-flows while maintaining the same demand tends to increase the number of flows in progress. This naturally reduces the fair rate thus increasing P_c. Consider now a set of classes such that $c_i < c$ for all i (i.e., after reducing the rate of high-rate-flows). The dimensioning objective is to ensure the link leaves the transparent regime with probability less than ε. There is a folk theorem that such congestion increases with the "burstiness" of the arrival process. Assimilating burstiness to the variance of the input rate, this indeed increases with flow rates. It is explicit in the Gaussian approach to dimensioning where the congestion increases with input rate variance.

For all practical purposes, the worst case traffic mix for rate c congestion thus corresponds to all flows having the same peak rate, c. This is precisely the assumption of Section 4.2 where P_c was shown to be given by the Erlang C-formula. We conclude that a dimensioning rule to ensure $P_c < \varepsilon$ for any traffic mix with overall demand A is to determine C such that $E_2(C/c, A/c) < \varepsilon$. The Internet Erlang formula is none other than the Erlang delay formula [1]!

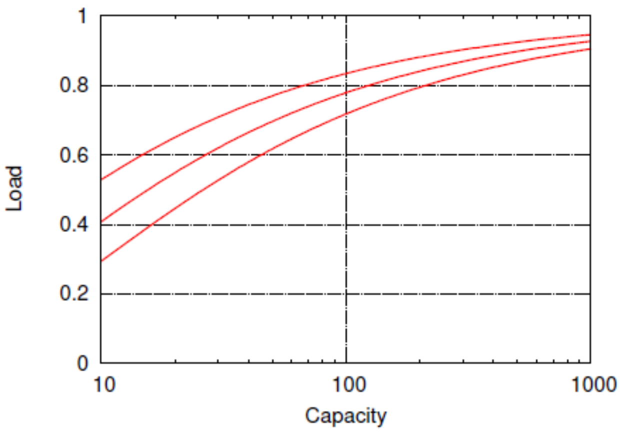
5.3 Significance

Note that the Erlang C-formula provides a bound that is not necessarily tight. The importance of the bound, even when it is not tight, is that it has precisely the same robustness as the Erlang loss formula. Performance depends only on overall expected demand A for whatever mix of flow rates and for the very general Poisson session traffic model. Recall that under this traffic model, packet and flow arrival processes are self-similar whenever the distributions of the flow size and the number of flows per session, respectively, have a heavy tail. The above results demonstrate that these characteristics have no significant impact on performance: the Erlang C bound is insensitive [1].

While the bound may not always be tight, it is nevertheless a very useful dimensioning tool. In particular, it exhibits scale economies similar to those of the telephone network, as depicted in Fig.7. Achievable utilization increases towards 100% as the ratio C/c grows. For example, utilization greater than

80% is compatible with 5 Mb/s streaming flows suffering congestion of less than 0.1% on any link of capacity greater than 1 Gb/s.

The Erlang C bound is a solid result that is independent of any assumption about traffic demand other than its overall average. If an operator knows more about the actual traffic mix, notably the traffic proportion due to high-rate-flows and their rates, a more precise demand-capacity-performance relation could be derived, using the mathematical models developed in [2] for instance.



Source: Bonald, Roberts. Internet and the Erlang Formula [1].

Fig.7. Admissible load A/C as a function of relative link capacity C=c for rate-c congestion probabilities 5%; 1%; 0.1% (from top to bottom).

6. Summary

This article is a survey of the related works in the literature on teletraffic engineering in IP network based on Erlang formulae. It is surveyed various techniques proposed for evaluation of the three-way relation between demand, capacity and performance that are available in the Internet.

It is demonstrated that, under a realistic flow-level model of Internet traffic, a simple performance parameter useful as a dimensioning criterion for network links depends only on link capacity and overall demand. Explicitly, the probability a flow of given peak rate must reduce it is bounded by the Erlang delay formula.

The Internet performance relation is analogous to the Erlang loss formula of the telephone network in several ways. It depends only on link capacity and expected load and not on more detailed traffic characteristics [1]. It is therefore robust to changing usage. Both formulae reveal scale economies that validate simple, maximum-load dimensioning criteria for large capacity links. Its adoption as a dimensioning criterion facilitates network management since we only need to monitor and estimate average overall demand in representative busy periods.

Like the Erlang loss formula, the bound is not sufficient for all purposes and more precise performance measures are sometimes required. The presented analysis is based on an extensive body of work, summarized in the paper [2], where many more results relevant to both wired and wireless networks are presented. It is possible to account for finite source traffic in the access network or to derive end-to-end performance measures for a network path.

The validity of the Internet Erlang formula relies on the assumption of max-min fair sharing. In practice, fairness does not need to be perfectly precise but one must question current reliance on end-systems voluntarily implementing TCP, or TCP-friendly, congestion control. J. Roberts and T. Bonald consider that the future Internet should impose per-flow fairness. This is technically feasible and is arguably the only traffic control needed to satisfy performance requirements. Fair sharing makes the network manageable precisely and then it is possible to apply the Erlang delay formula. This cannot be said for other traffic control architectures, like Diffserv for instance.

We hope that present survey can give a new direction to the research in the IP network teletraffic engineering.

REFERENCES

[1] Bonald, T., J. Roberts. Internet and the Erlang Formula. ACM SIGCOMM Computer Communication Review, Vol. 42, Num. 1, 2012, pp. 23-30.

[2] Bonald, T., L. Massoulie, A. Proutiere and J. Virtamo. A queueing analysis of max-min fairness, proportional fairness and balanced fairness. Queueing Systems, No. 53, 2006, pp. 65–84.

[3] Bonald, T. The Erlang model with non-Poisson call arrivals. In SIGMETRICS / Performance, No. 34, 2006, pp. 276-286.

[4] Cerf, V., B. Davie, A. Greenberg, S. Landau, D. Sincoskie. FIND (Future Internet Design) Observer Panel Report. US National Science Foundation, 2009.

[5] Chromý E., M. Kavacký. Asynchronous Networks and Erlang Formulas. International Journal of Communication Networks and Information Security (IJCNIS) Vol. 2, No. 2, 2010, pp. 85-89.

[6] Chromy, E., J. Suran, M. Kovacik and M. Kavacky. Usage of Erlang Formula in IP Networks. Communications and Network, Vol. 3 No. 3, 2011, pp. 161-167.

[7] Chromy, E., T. Misuth, M. Kavacky. Erlang C Formula and Its Use in the Call Centers. Information and Communication Technologies and Services, Vol. 9, No. 1, 2011, pp. 7-13.

[8] Erik, C., T. Misuth, A. Weber. Application of Erlang Formulae in Next Generation Networks. I. J. Computer Network and Information Security, N: 1, 2012, pp. 59-66.

[9] Faleigh, C., F. Tobagi, C. Diot, Provisioning IP Backbone Networks to Support Latency Sensitive Traffc. In Infocom, Vol. 1, 2003, pp. 375 - 385.

[10] Firdhous, M., O. Ghazali, S. Hassan. Modeling of Cloud System using Erlang Formulas. Proc. of 17th Asia-Pacific Conference on Communications (APCC), Sabah, Malaysia, 2011, pp. 411-416.

[11] Ben Fredj, S.B., T. Bonald, A. Proutiere, G. Regnie and J. Roberts. Statistical bandwidth sharing: a study of congestion at flow level. In SIGCOMM, 2001, pp. 111-122.

[12] Hahne, E. Round-Robin Scheduling for Max-Min Fairness in Data Networks. IEEE JSAC, Vol. 9, No 7, 1991, pp. 1024-1039.

[13] Iversen, V. Teletraffic Engineering and Network Planning, ITU-D SG 2/16 & ITC, Revised 2010, ftp://ftp.dei.polimi.it/users/Flaminio.Borgonovo/Teoria/teletraffic_Iversen.pdf.

[14] Iversen, V. The Internet Erlang Formula. 12th International Conference, NEW2AN 2012: Internet of Things, Smart Spaces, and Next Generation Networking, St. Petersburg, Russia, 2012 Proceedings: Lecture Notes in Computer Science, Vol. 7469, Springer, 2012, p. 328-337.

[15] Kelly, F. Mathematical modelling of the Internet. In "Mathematics Unlimited - 2001 and Beyond" (Editors B. Engquist and W. Schmid). Springer-Verlag, Berlin, 2001, pp.

[16] Kaufman, J. Blocking in a shared resource environment. IEEE Trans. Commun., 29:500, 1981, pp. 1474-1481.

[17] Logothetis, M, I. Moscholios. Teletraffic Models beyond Erlang. Proc. of the 10th International ELEKTRO 2014 – Zilina, Slovakia, 2014, pp. 10-15.

[18] Mirtchev, S. Internet Evolution, Teletraffic and QoS: a Survey of Network Traffic. Electrotechnica & Electronica (E+E), Vol. 50, No.1-2, 2015, pp. 2-17.

[19] Moscholios, I., V. Vassilakis, M. Logothetis and A. Boucouvalas. Blocking Equalization in the Erlang Multirate Loss Model for Elastic Traffic. Proc. of 2nd Int. Conference on Emerging Network Intelligence, EMERGING 2010, Florence, Italy, 25-30 October 2010.

[20] Paxson, V., S. Floyd. Wide Area Traffic: The Failure of Poisson Modeling. IEEE/ACM Transactions on Networking, Vol. 3, 1995, pp. 226-244.

[21] Riedl, A., T. Bauschert, J.Frings. A Framework for Multi-Service IP Network Planning. 10’th International Network Planning Symposium (NETWORKS’02), München, 2002, pp. 183-190.

[22] Roberts, J. Internet traffic, QoS, and pricing. Proceedings of the IEEE, vol. 92, no. 9, 2004, pp. 1389-1399.

[23] Roberts, J. An Erlang formula for the Internet. Slides, 2011, last accessible 2016:

http://ets2011.et.put.poznan.pl/RobertsSlides.pdf

[24] Roberts, J. The Cloud is the future Internet: How do we engineer a cloud? Keynote, Slides. Infocom 2013, 32nd IEEE International Conference on Computer Communications, Torino, Italy, 2013.

[25] Roberts, J. A service system with heterogeneous user requirement. In G. Pujolle, ed.: Performance of Data Communications Systems and Their Applications, 1981, pp. 423-431.

[26] Sandvine’s Global Internet Phenomena Reports: 2H 2014, www.sandvine.com.

[27] Vilaplana, J., F. Solsona1 and I. Teixid. A performance model for scalable cloud computing. Proceedings of the 13th Australasian Symposium on Parallel and Distributed Computing (AusPDC 2015), Sydney, Australia, 2015, pp. 51-60.

[28] Yu, J., I. Ajarmeh. Design and Traffic Engineering of VoIP for Enterprise and Carrier Networks. International Journal on Advances in Telecommunications, Vol. 1, No 1, 2008, pp.27-39.

Assoc. Prof. DSc Seferin T. Mirtchev *has graduated telecommunications at Technical University of Sofia in 1981. He is with Department of Communication Networks, vice president of the Union of Electronics, Electrical Engineering and Telecommunications (CEEC), member of IEEE and has research interest in teletraffic engineering, quality of service, switching systems in telecommunication networks.*

tel.: +359 2 965 2254 *e-mail: stm@tu-sofia.bg*

Received on: 08.02.2016