
Teletraffic
Engineering
Marylee Mcduffie

First Edition, 2012

ISBN 978-81-323-4037-9

Published by:
White Word Publications
4735/22 Prakashdeep Bldg,
Ansari Road, Darya Ganj,
Delhi - 110002
Email: info@wtbooks.com

Table of Contents

Introduction

Telecommunications traffic engineering, **Teletraffic engineering** or **traffic engineering** is the application of traffic engineering theory to telecommunications. Teletraffic engineers use their basic knowledge of statistics including queuing theory, the nature of traffic, their practical models, their measurements and simulations to make predictions and to plan telecommunication networks such as a telephone network or the Internet. These tools and basic knowledge help provide reliable service at lower cost.

The field was created by the work of A. K. Erlang for circuit-switched networks but is applicable to packet-switched networks. The most notable difference between these sub-fields is that packet-switched data traffic is self-similar. This is a consequence of the calls being between computers, and not people.

The crucial observation in traffic engineering is that in large systems the law of large numbers can be used to make the aggregate properties of a system over a long period of time much more predictable than the behaviour of individual parts of the system.

Teletraffic in PSTN architectures

The measurement of traffic in a public switched telephone network (PSTN) allows network operators to determine and maintain the quality of service (QoS) and in particular the grade of service (GoS) that they promise their subscribers. The performance of a network depends on whether all origin-destination pairs are receiving a satisfactory service.

Networks are handled as:

- **loss systems** where calls that cannot be handled are given equipment busy tone or
- **queuing systems** where calls that cannot be handled immediately are queued.

Congestion is defined as the situation when exchanges or circuit groups are inundated with calls and are unable to serve all the subscribers. Special attention must be given to ensure that such high loss situations do not arise. To help determine the probability of congestion occurring, operators should use the Erlang formulas or the Engset calculation.

Exchanges in the PSTN make use of trunking concepts to help minimize the cost of the equipment to the operator. Modern switches generally have full availability and do not make use of grading concepts.

Overflow systems make use of alternative routing circuit groups or paths to transfer excess traffic and thereby reduce the possibility of congestion.

A very important component in PSTNs is the SS7 network used to route signalling traffic. As a supporting network, it carries all the signalling messages necessary to set up, break down or provide extra services. The signalling enables the PSTN control the manner in which traffic is routed from one location to another.

Transmission and switching of calls is performed using the principle of time-division multiplexing (TDM). TDM allows multiple calls to be transmitted along the same physical path, reducing the cost of infrastructure.

Teletraffic engineering in call centers

A good example of the use of teletraffic theory in practice is in the design and management of a call center. Call centers use teletraffic theory to increase the efficiency of their services and overall profitability through calculating how many operators are really needed at each time of the day.

Queueing systems used in call centers have been studied as a science. For example completed calls are put on hold and queued until they can be served by an operator. If callers are made to wait too long, they may lose patience and default from the queue (hang up), resulting in no service being provided.

Teletraffic engineering in broadband networks

Teletraffic Engineering is a well-understood discipline in the traditional voice network, where traffic patterns are established, growth rates can be predicted, and vast amounts of detailed historical data are available for analysis. However, in modern broadband networks, the teletraffic engineering methodologies used for voice networks are inappropriate. Various aspects relating to teletraffic engineering in broadband networks are discussed here.

Long-tail traffic

Of great importance is the possibility that extremely infrequent occurrences are more likely than anticipated. This situation is known as long-tail traffic. In some designs, the network might be required to withstand the unanticipated traffic.

Teletraffic economics and forecasting

As mentioned in the introduction, the purpose of teletraffic theory is to reduce cost in telecommunications networks. An important tool in achieving this goal is forecasting. Forecasting allows network operators to calculate the potential cost of a new network / service for a given GoS during the planning and design stage, thereby ensuring that costs are kept to a minimum.

An important method used in forecasting is simulation, which is described as the most common quantitative modelling technique in use today. An important reason for this is that computing power has become far more accessible, making simulation the preferred analytical method for problems that are not easily solved mathematically.

As in any business environment, network operators must charge tariffs for their services. These charges must be balanced with the supplied QoS. When operators supply services internationally, this is described as trade in services and is governed by the General Agreement on Trade in Services (GATS).

Chapter-1

Erlang

The **erlang** (symbol **E**) is a dimensionless unit that is used in telephony as a statistical measure of offered load or carried load on service-providing elements such as telephone circuits or telephone switching equipment. It is named after the Danish telephone engineer A. K. Erlang, the originator of traffic engineering and queueing theory.

Traffic measurements of a telephone circuit

When used to represent **carried traffic**, a value (which can be a non-integer such as 43.5) followed by "erlangs" represents the average number of concurrent calls carried by the circuits (or other service-providing elements), where that average is calculated over some reasonable period of time. The period over which the average is calculated is often one hour, but shorter periods (e.g., 15 minutes) may be used where it is known that there are short spurts of demand and a traffic measurement is desired that does not mask these spurts. One erlang of carried traffic refers to a single resource being in continuous use, or two channels being in use fifty percent of the time, and so on. For example, if an office has two telephone operators who are both busy all the time, that would represent two erlangs (2 E) of traffic; or a radio channel that is occupied for one hour continuously is said to have a load of 1 Erlang.

When used to describe **offered traffic**, a value followed by "erlangs" represents the average number of concurrent calls that would have been carried if there were an unlimited number of circuits (that is, if the call-attempts that were made when all circuits were in use had not been rejected). The relationship between offered traffic and carried traffic depends on the design of the system and user behavior. Three common models are (a) callers whose call-attempts are rejected go away and never come back, (b) callers whose call-attempts are rejected try again within a fairly short space of time, and (c) the system allows users to wait in queue until a circuit becomes available.

A third measurement of traffic is **instantaneous traffic**, expressed as a certain number of erlangs, meaning the exact number of calls taking place at a point in time. In this case the number is an integer. Traffic-level-recording devices, such as moving-pen recorders, plot instantaneous traffic.

The concepts and mathematics introduced by Agner Krarup Erlang have broad applicability beyond telephony. They apply wherever users arrive more or less at random to receive exclusive service from any one of a group of service-providing elements without prior reservation, for example, where the service-providing elements are ticket-sales windows, toilets on an airplane, or motel rooms. (Erlang's models do not apply where the server-providing elements are shared between several concurrent users or different amounts of service are consumed by different users, for instance, on circuits carrying data traffic.)

Offered traffic (in erlangs) is related to the **call arrival rate**, λ, and the **average call-holding time**, h, by:

$$E = \lambda h$$

provided that h and λ are expressed using the same units of time (seconds and calls per second, or minutes and calls per minute).

The practical measurement of traffic is typically based on continuous observations over several days or weeks, during which the instantaneous traffic is recorded at regular, short intervals (such as every few seconds). These measurements are then used to calculate a single result, most commonly the **busy hour traffic** (in erlangs). This is the average, over several days, of the average number of concurrent calls during the same one-hour period each day, where that period is selected to give the highest result. (This result is called the time-consistent busy hour traffic). An alternative is to calculate a busy hour traffic value separately for each day (which may correspond to slightly different times each day) and take the average of these values. This generally gives a slightly higher value than the time-consistent busy hour value.

The goal of Erlang's traffic theory is to determine exactly how many service-providing elements should be provided in order to satisfy users, without wasteful over-provisioning. To do this, a target is set for the grade of service (GoS) or quality of service (QoS). For example, in a system where there is no queuing, the GoS may be that no more than 1 call in 100 is blocked (i.e., rejected) due to all circuits being in use (a GoS of 0.01), which becomes the target probability of call blocking, P_b, when using the Erlang B formula.

There are several Erlang formulae, including Erlang B, Erlang C and the related Engset formula, based on different models of user behavior and system operation. These are discussed below, and may each be derived by means of a special case of continuous-time Markov processes known as a birth-death process.

Where the existing busy-hour carried traffic, E_c, is measured on an already-overloaded system, with a significant level of blocking, it is necessary to take account of the blocked calls in estimating the busy-hour offered traffic E_o (which is the traffic value to be used in the Erlang formula). The offered traffic can be estimated by $E_o = E_c/(1 - P_b)$. For this purpose, where the system includes a means of counting blocked calls and successful calls, P_b can be estimated directly from the proportion of calls that are blocked. Failing

that, P_b can be estimated by using E_c in place of E_o in the Erlang formula and the resulting estimate of P_b can then be used in $E_o = E_c/(1 - P_b)$ to estimate E_o. Another method of estimating E_o in an overloaded system is to measure the busy-hour call arrival rate, λ (counting successful calls and blocked calls), and the average call-holding time (for successful calls), h, and then estimate E_o using the formula $E = \lambda h$.

For a situation where the traffic to be handled is completely new traffic, the only choice is to try to model expected user behavior, estimating active user population, N, expected level of use, U (number of calls/transactions per user per day), busy-hour concentration factor, C (proportion of daily activity that will fall in the busy hour), and average holding time/service time, h (expressed in minutes). A projection of busy-hour offered traffic would then be $E_o = (NUC/60)h$ erlangs. (The division by 60 translates the busy-hour call/transaction arrival rate into a per-minute value, to match the units in which h is expressed.)

Erlang B formula

Erlang-B (sometimes also written without the hyphen **Erlang B**), also known as the **Erlang loss formula**, is a formula for the **blocking probability** derived from the Erlang distribution to describe the probability of call loss on a group of circuits (in a circuit switched network, or equivalent). It is, for example, used in planning telephone networks. The formula was derived by Agner Krarup Erlang and is not limited to telephone networks, since it describes a probability in a queuing system (albeit a special case with a number of servers but no buffer spaces for incoming calls to wait for a free server). Hence, the formula is also used in certain inventory systems with lost sales.

The formula applies under the condition that an unsuccessful call, because the line is busy, is not queued or retried, but instead really lost forever. It is assumed that call attempts arrive following a Poisson process, so call arrivals are independent. Further it is assumed that message length (holding times) are exponentially distributed (Markovian system) although the formula turns out to apply under general holding time distributions.

Erlangs are a dimensionless quantity calculated as the average arrival rate, λ, multiplied by the average call length, h. The Erlang B formula assumes an infinite population of sources (such as telephone subscribers), which jointly offer traffic to N servers (such as links in a trunk group). The rate of arrival of new calls (birth rate) is equal to λ and is constant, *not* depending on the number of active sources, because the total number of sources is assumed to be infinite. The rate of call departure (death rate) is equal to the number of calls in progress divided by h, the mean call holding time. The formula calculates blocking probability in a loss system, where if a request is not served immediately when it tries to use a resource, it is aborted. Requests are therefore not queued. Blocking occurs when there is a new request from a source, but all the servers are already busy. The formula assumes that blocked traffic is immediately cleared.

The formula provides the GoS (grade of service) which is the probability P_b that a new call arriving at the circuit group is rejected because all servers (circuits) are busy: $B(E, m)$ when E Erlang of traffic are offered to m trunks (communication channels).

$$P_b = B(E,m) = \frac{\frac{E^m}{m!}}{\sum_{i=0}^{m} \frac{E^i}{i!}}$$

where:

- P_b is the probability of blocking
- m is the number of resources such as servers or circuits in a group
- $E = \lambda h$ is the total amount of traffic offered in erlangs

This may be expressed recursively as follows, in a form that is used to simplify the calculation of tables of the Erlang B formula:

$$B(E,0) = 1$$
$$B(E,j) = \frac{EB(E,j-1)}{EB(E,j-1)+j} \ \forall j = 1,2,...,m$$

Typically, instead of $B(E, m)$ the inverse $1/B(E, m)$ is calculated in numerical computation in order to ensure numerical stability:

$$\frac{1}{B(E,0)} = 1$$
$$\frac{1}{B(E,j)} = 1 + \frac{j}{E}\frac{1}{B(E,j-1)} \ \forall j = 1,2,...,m$$

```
Function ErlangB (E as Double, m As Integer) As Double
Dim InvB As Double
Dim j As Integer

  InvB = 1.0
  For j = 1 To m
    InvB = 1.0 + j / E * InvB
  Next j
  ErlangB = 1.0 / InvB
End Function
```

The Erlang B formula applies to loss systems, such as telephone systems on both fixed and mobile networks, which do not provide traffic buffering, and are not intended to do so. It assumes that the call arrivals may be modeled by a Poisson process, but is valid for any statistical distribution of call holding times with finite mean. Erlang B is a trunk sizing tool for voice switch to voice switch traffic. The Erlang B formula is decreasing and convex in m.

Extended Erlang B

Extended Erlang B is an iterative calculation, rather than a formula, that adds an extra parameter, the Recall Factor, which defines the recall attempts.

The steps in the process are as follows:

1. Calculate

$$P_b = B(E, m)$$

as above for Erlang B.

2. Calculate the probable number of blocked calls

$$B_e = EP_b$$

3. Calculate the number of recalls, R assuming a Recall Factor, R_f:

$$R = B_e R_f$$

4. Calculate the new offered traffic

$$E_{i+1} = E_0 + R$$

where E_0 is the initial (baseline) level of traffic.

5. Return to step 1 and iterate until a stable value of E is obtained.

Erlang C formula

The **Erlang C formula** expresses the waiting probability in a queuing system. Just as the Erlang B formula, Erlang C assumes an infinite population of sources, which jointly offer traffic of A erlangs to N servers. However, if all the servers are busy when a request arrives from a source, the request is queued. An unlimited number of requests may be held in the queue in this way simultaneously. This formula calculates the probability of queuing offered traffic, assuming that blocked calls stay in the system until they can be handled. This formula is used to determine the number of agents or customer service representatives needed to staff a call centre, for a specified desired probability of queuing.

$$P_W = \frac{\frac{A^N}{N!}\frac{N}{N-A}}{\sum_{i=0}^{N-1}\frac{A^i}{i!} + \frac{A^N}{N!}\frac{N}{N-A}}$$

where:

- A is the total traffic offered in units of erlangs
- N is the number of servers
- P_W is the probability that a customer has to wait for service

It is assumed that the call arrivals can be modeled by a Poisson process and that call holding times are described by a negative exponential distribution. A common use for Erlang C is modeling and dimensioning call center agents in a call center environment.

Engset formula

The **Engset calculation** is a related formula, named after its developer, T. O. Engset, used to determine the probability of congestion occurring within a telephony circuit group. It deals with a finite population of S sources rather than the infinite population of sources that Erlang assumes. The formula requires that the user knows the expected peak traffic, the number of sources (callers) and the number of circuits in the network.

Example application

A business installing a PABX needs to know the minimum number of voice circuits it needs to have to and from the telephone network. An approximate approach is to use the Erlang-B formula. However, if the business has a small number of extensions, then it should instead use the more exact Engset calculation, which reflects the fact that extensions already in use will not make additional simultaneous calls. (For a large user population, the Engset and the Erlang-B calculations give the same result.)

Technical details

Engset's equation is similar to the Erlang-B formula; however it contains one major difference: Erlang's equation assumes an infinite source of calls, yielding a Poisson arrival process, while Engset specifies a finite number of callers . Thus Engset's equation should be used when the source population is small (say less than 200 users, extensions or customers).

$$P_b(N, A, S) = \frac{A^N \binom{S}{N}}{\sum_{i=0}^{N} A^i \binom{S}{i}}$$

where

A = offered traffic intensity in erlangs, from all sources
S = number of sources of traffic
N = number of circuits in group
P(b) = probability of blocking or congestion

In practice, like Erlang's equations, Engset's formula requires recursion to solve for the blocking or congestion probability. There are several recursions that could be used. One way to determine this probability, one first determines an initial estimate. This initial estimate is substituted into the equation and the equation then is solved. The answer to this initial calculation is then substituted back into the equation, resulting in a new answer which is again substituted. This iterative process continues until the equation converges to a stable result..

Engset's equation follows:

$$P(b) = \frac{\left[\frac{(S-1)!}{N!\cdot(S-1-N)!}\right] \cdot M^N}{\sum_{X=1}^{N} \left[\frac{(S-1)!}{X!\cdot(S-1-X)!}\right] \cdot M^X}$$

$$M = \frac{A}{S - A \cdot (1 - P(b))}$$

Chapter-2

Call Management and Cellular Traffic

Call management

In telecommunications, **call management** is the process of designing and implementing rules and parameters governing the routing of inbound telephone calls through a network. These rules can specify how calls are distributed according to the time and/or date of the call as well as the location of the caller (usually defined by the outbound Caller ID). Call Management also involves the use of Calling Features such as Call Queues, IVR Menus, Hunt Groups and Recorded Announcements to provide a customised experience for the caller and to maximize the efficiency of inbound call handling. Call management is most effective when a call logging software tool is used.

Network types

Call Management is performed on varying degrees of scale, from an individual screening unwanted calls from a residential landline to an international call carrier routing calls to different worldwide locations by percentage. Systems for governing Call Management can be in the form of hardware, such as a PBX Telephone System attached to an ISDN30 or a hosted software-based system.

Calling features

Calls are routed according to the setting up of calling features within the given system. Common examples of Calling Features include:

- **Translation** – The automatic routing of inbound calls from one telephone number to another.
- **Hunt Group** – A directory containing one or many destination numbers which, on receiving an incoming call, is programmed to ring them in a particular order, simultaneously or simply in the order in which they have most recently answered before being sent to a final destination if still unanswered.
- **Call Queue** – A directory similar to a Hunt Group that keeps the caller on hold until one of the destination numbers becomes available.

- **Auto Attendant** – A large directory of extension numbers which can be chosen by the caller, each with its own specific routing behaviour.
- **Location-Based Routing** – Rules programmed in at particular points in a system to route the call on to different destinations depending on the location of the caller.
- **Time and Date-Based Routing** – Rules programmed in at particular points in a system to route the call on to different destinations depending the time or date of the call.
- **Call Whisper** – A message played to an agent after answering a call that can give them information about the call in advance based on the Caller ID, number dialled or route taken through the system.
- **Interactive voice response** – A sound recording device to allow a caller to give information to the system verbally about what services or support they require.
- **Fax to Email** – A Device for routing inbound fax calls to one or more email addresses, usually as attachments.

Call records

Systems often retain information about received calls which can be stored, analysed and interpreted by the system administrator.

- **Call Detail Records** (CDRs) – Records of all received calls, usually including time, date, duration, calling number and called number. Hosted services can also show pricing information.
- **Call Recording** – Many systems have the ability to record and store calls for future reference.
- **Voice and Fax Mailboxes** – Inbound faxes and voicemail messages can be stored on systems also.

Cellular traffic

Here we, discusses the **mobile cellular network** aspect of **teletraffic measurements**. Mobile radio networks have traffic issues that do not arise in connection with the fixed line PSTN. Important aspects of cellular traffic include: quality of service targets, traffic capacity and cell size, spectral efficiency and sectorization, traffic capacity versus coverage, and channel holding time analysis.

Teletraffic engineering in telecommunications network planning ensures that network costs are minimised without compromising the quality of service delivered to the user of the network. This field of engineering is based on probability theory and can be used to analyse mobile radio networks, as well as other telecommunications networks.

A mobile handset which is moving in a cell will record a signal strength that varies. Signal strength is subject to slow fading, fast fading and interference from other signals,

resulting in degradation of the carrier-to-interference (C/I) ratio. A high C/I ratio yields quality communication. A good C/I ratio is achieved in cellular systems by using optimum power levels through the power control of most links. When carrier power is too high, excessive interference is created, degrading the C/I ratio for other traffic and reducing the traffic capacity of the radio subsystem. When carrier power is too low, C/I is too low and QoS targets are not met.

Quality of Service targets

At the time that the cells of a radio subsystem are designed, Quality of Service (QoS) targets are set, for: traffic congestion and blocking, dominant coverage area, C/I, dropped call rate, handover failure rate, overall call success rate,

Traffic load and cell size

The more traffic generated, the more base stations will be needed to service the customers. The number of base stations for a simple cellular network is equal to the number of cells. The traffic engineer can achieve the goal of satisfying the increasing population of customers by increasing the number of cells in the area concerned, so this will also increases the number of base stations. This method is called cell splitting (and combined with sectorization) is the only way of providing services to a burgeoning population. This simply works by dividing the cells already present into smaller sizes hence increasing the traffic capacity. Reduction of the cell radius enables the cell to accommodate extra traffic. The cost of equipment can also be cut down by reducing the number of base stations through setting up three neighbouring cells, with the cells serving three 120° sectors with different channel groups.

Mobile radio networks are operated with finite, limited resources (the spectrum of frequencies available). These resources have to be used effectively to ensure that all users receive service, that is, the quality of service is consistently maintained. This need to carefully use the limited spectrum, brought about the development of cells in mobile networks, enabling frequency re-use by successive clusters of cells. Systems that efficiently use the available spectrum have been developed e.g. the GSM system. Walke defines spectral efficiency as the traffic capacity unit divided by the product of bandwidth and surface area element, and is dependent on the number of radio channels per cell and the cluster size (number of cells in a group of cells):

$$\text{efficiency} = \frac{N_c}{\text{BW} \cdot A_c},$$

where N_c is the number of channels per cell, BW is the system bandwidth, and A_c is Area of cell.

Sectorization is briefly described in **traffic load and cell size** as a way to cut down equipment costs in a cellular network. When applied to clusters of cells sectorization also

reduces co-channel interference, according to Walke. This is because the power radiated backward from a directional base station antenna is minimal and interfering with adjacent cells is reduced. (The number of channels is directly proportional to the number of cells.) The maximum traffic capacity of sectored antennas (directional) is greater than that of omnidirectional antennas by a factor which is the number of sectors per cell (or cell cluster).

Traffic capacity versus coverage

Cellular systems use one or more of four different techniques of access (TDMA, FDMA, CDMA, SDMA). Let a case of Code Division Multiple Access be considered for the relationship between traffic capacity and coverage (area covered by cells). CDMA cellular systems can allow an increase in traffic capacity at the expense of the quality of service.

In TDMA/FDMA cellular radio systems, Fixed Channel Allocation (FCA) is used to allocate channels to customers. In FCA the number of channels in the cell remains constant irrespective of the number of customers in that cell. This results in traffic congestion and some calls being lost when traffic gets heavy.

A better way of channel allocation in cellular systems is Dynamic Channel Allocation (DCA) which is supported by the GSM, DCS and other systems. DCA is a better way not only for handling bursty cell traffic but also in efficiently utilising the cellular radio resources. DCA allows the number of channels in a cell to vary with the traffic load, hence increasing channel capacity with little costs. Since a cell is allocated a group of frequency carries (e.g. f_1-f_7) for each user, this range of frequencies is the bandwidth of that cell, BW. If that cell covers an area A_c, and each user has bandwidth B then the number of channels will be BW/B. The density of channels will be $\frac{\mathrm{BW}}{A_c \cdot \mathrm{B}}$. This formula shows that as the coverage area A_c is increased, the channel density decreases.

Channel holding time

Important parameters like the carrier-to-interference ratio (C/I), spectral efficiency and reuse distance determine the quality of service of a cellular network. Channel Holding Time is another parameter that can affect the quality of service in a cellular network, hence it is considered when planning the network. Calculating the channel holding time, however is not easy. (This is the time a Mobile Station (MS) remains in the same cell during a call). Channel holding time is therefore less than call holding time if the MS travels more than one cell as handover will take place and the MS relinquishes the channel. Practically, it is not possible to determine exactly the channel holding time. As a result, different models exist for the channel holding time distribution. In industry, a good approximation of the channel holding time is usually sufficient to determine the network traffic capability.

One of the papers in Key and Smith defines channel holding time as being equal to the average holding time divided by the average number of handovers per call plus one. Usually an exponential model is preferred to calculate the channel holding time for simplicity in simulations. This model gives the distribution function of channel holding time and it is an approximation that can be used to obtain estimates channel holding time. The exponential model may not be correctly modelling the channel holding time distribution as other papers may try to prove, but it gives an approximation. Channel holding time is not easily determined explicitly, call holding time and user's movements have to be determined in order to implicitly give channel holding time. The mobility of the user and the cell shape and size cause the channel holding time to have a different distribution function to that of call duration (call holding time). This difference is large for highly mobile users and small cell sizes. Since the channel holding time and call duration relationships are affected by mobility and cell size, for a stationary MS and large cell sizes, channel holding time and call duration are the same.

Chapter-3

Generic Cell Rate Algorithm

The **Generic Cell Rate Algorithm** (GCRA) is an algorithm that is used in Asynchronous Transfer Mode (ATM) networks to measure the timing of cells on Virtual Channels (VCs) and or Virtual Paths (VPs) against bandwidth and jitter limits contained in a traffic contract for the VC or VP to which the cells belong. Cells that do not conform may then be re-timed (delayed) in traffic shaping, or dropped or reduced in priority in traffic policing. Nonconforming cells that are reduced in priority may then be dropped, in preference to higher priority cells, by downstream components in the network that are experiencing congestion. Alternatively they may reach their destination (VC or VP termination) if there is enough capacity for them, despite them being excess cells as far as the contract is concerned.

The GCRA is given as the reference for checking the traffic on connections in the network, i.e. Usage/Network Parameter Control (UPC/NPC) at User–Network Interfaces (UNI) or Inter-Network Interfaces or Network-Network Interfaces (INI/NNI) . It is also given as the reference for the timing of cells transmitted (ATM PDU Data_Requests) onto an ATM network by a Network Interface Card (NIC) in a host, i.e. on the user side of the UNI . This ensures that cells are not then discarded by UPC/NCP in the network, i.e. on the network side of the UNI. However, as the GCRA is only given as a reference, the network providers and users may use any other algorithm that gives the same result.

Description of the GCRA

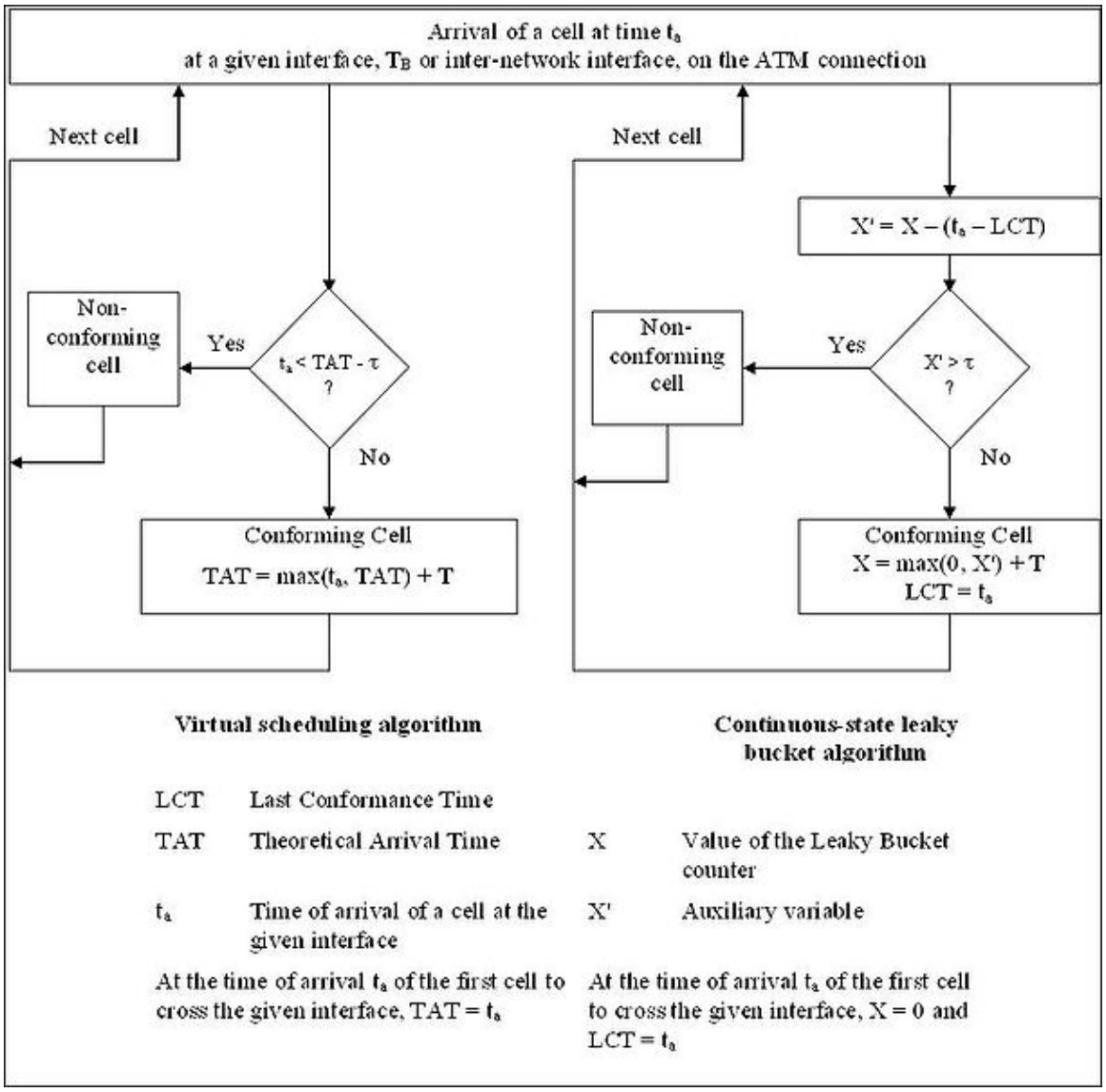

Figure 1: Equivalent versions of the generic cell rate algorithm

The GCRA is described by the ATM Forum in its *User-Network Interface (UNI)* and by the ITU-T in recommendation I.371 *Traffic control and congestion control in B-ISDN* . Both sources describe the GCRA in two equivalent ways: as a virtual scheduling algorithm and as a continuous state leaky bucket algorithm (figure 1).

Leaky bucket description

The description in terms of the leaky bucket algorithm may be the easier of the two to understand from a conceptual perspective, as it is based on a simple analogy of a bucket with a leak: see figure 1 on the leaky bucket page. However, there has been confusion in

the literature over the application of the leaky bucket analogy to produce an algorithm, which has crossed over to the GCRA. The GCRA should be considered as a version of the leaky bucket as a meter rather than the leaky bucket as a queue.

However, while there are possible advantages in understanding this leaky bucket description, it does not necessarily result in the best (fastest) code if implemented directly. This is evidenced by the relative number of actions to be performed in the flow diagrams for the two descriptions (figure 1).

The description in terms of the continuous state leaky bucket algorithm is given by the ITU-T as follows: “The continuous-state leaky bucket can be viewed as a finite capacity bucket whose real-valued content drains out at a continuous rate of 1 unit of content per time unit and whose content is increased by the increment T for each conforming cell... If at a cell arrival the content of the bucket is less than or equal to the limit value τ, then the cell is conforming; otherwise, the cell is non-conforming. The capacity of the bucket (the upper bound of the counter) is $(T + \tau)$” . It is worth noting that because the leak is one unit of content per unit time, the increment for each cell T and the limit value τ are in units of time.

Considering the flow diagram of the continuous state leaky bucket algorithm, in which T is the emission interval and τ is the limit value: What happens when a cell arrives is that the state of the bucket is calculated from its state when the last conforming cell arrived, X, and how much has leaked out in the interval, $t_a - LCT$. This current bucket value is then stored in X' and compared with the limit value τ. If the value in X' is not greater than τ, the cell did not arrive too early and so conforms to the contract parameters; if the value in X' is greater than τ, then it does not conform. If it conforms then, if it conforms because it was late, i.e. the bucket empty ($X' <= 0$), X is set to T; if it was early, but not too early, ($\tau >= X' > 0$), X is set to $X' + \tau$.

Thus the flow diagram mimics the leaky bucket analogy (used as a meter) directly, with X and X' acting as the analogue of the bucket.

Virtual scheduling description

The virtual scheduling algorithm, while not so obviously related to such an easily accessible analogy as the leaky bucket, gives a clearer understanding of what the GCRA does and how it may be best implemented. As a result, direct implementation of this version can result in more compact, and thus faster, code than a direct implementation of the leaky bucket description.

The description in terms of the continuous state leaky bucket algorithm is given by the ITU-T as follows: “The virtual scheduling algorithm updates a Theoretical Arrival Time (TAT), which is the 'nominal' arrival time of the cell assuming cells are sent equally spaced at an emission interval of T corresponding to the cell rate Λ [= $1/T$] when the source is active. If the actual arrival time of a cell is not 'too early' relative to the TAT and tolerance τ associated to the cell rate, i.e. if the actual arrival time is after its theoretical

arrive time minus the limit value ($t_a > TAT - \tau$), then the cell is conforming; otherwise, the cell is nonconforming" . If the cell is nonconforming then *TAT* is left unchanged. If the cell is conforming, and arrived before its TAT (equivalent to the bucket not being empty but being less than the limit value), then the next cell's *TAT* is simply $TAT + T$. However, if a cell arrives after its *TAT*, then the *TAT* for the next cell is calculated from this cell's arrival time, not its *TAT*. This prevents credit building up when there is a gap in the transmission (equivalent to the bucket becoming less than empty).

This version of the algorithm works because τ defines how much earlier a cell can arrive than it would if there were no jitter: delay variation tolerance. Another way to see it is that *TAT* represents when the bucket will next empty, so a time τ before that is when the bucket is exactly filled to the limit value. So, in either view, if it arrives more than τ before *TAT*, it is too early to conform.

Comparison with the token bucket

The GCRA, unlike implementations of the token bucket algorithm, does not simulate the process of updating the bucket (the leak or adding tokens regularly). Rather, each time a cell arrives it calculates the amount by which the bucket will have leaked since its level was last calculated or when the bucket will next empty (= *TAT*). This is essentially replacing the leak process with a (realtime) clock, which most hardware implementations are likely to already have.

This replacement of the process with an RTC is possible because ATM cells have a fixed length (53 bytes), thus *T* is always a constant, and the calculation of the new bucket level (or of *TAT*) does not involve any multiplication or division. As a result, the calculation can be done quickly in software, and while more actions are taken when a cell arrives than are taken by the token bucket, in terms of the load on a processor performing the task, the lack of a separate update process more than compensates for this. Moreover, because there is no simulation of the bucket update, there is no processor load at all when the connection is quiescent.

However, if the GCRA were to be used to limit to a bandwidth, rather than a packet/frame rate, in a protocol with variable length packets (Link Layer PDUs), it would involve multiplication: basically the value added to the bucket (or to TAT) for each conforming packet would have to be proportionate to the packet length: whereas, with the GCRA as described, the water in the bucket has units of time, for variable length packets it would have to have units that are the product of packet length and time. Hence, applying the GCRA to limit the bandwidth of variable length packets without access to a fast, hardware multiplier (as in an FPGA) may not be practical. However, it can always be used to limit the packet or cell rate, as long as their lengths are ignored.

Dual Leaky Bucket Controller

Multiple implementations of the GCRA can be applied concurrently to a VC or a VP, in a dual leaky bucket traffic policing or traffic shaping function, e.g. applied to a Variable

Bit Rate (VBR) VC. This can limit ATM cells on this VBR VC to a Sustained Cell Rate (SCR) and a Maximum Burst Size (MBS). At the same time, the dual leaky bucket traffic policing function can limit the rate of cells in the bursts to a Peak Cell Rate (PCR) and a maximum Cell Delay Variation tolerance (CDVt).

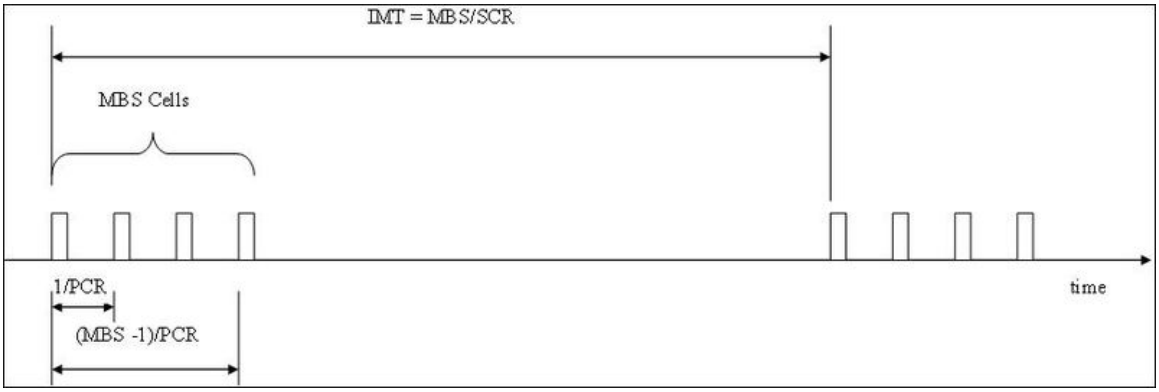

Figure 2: Example cell timings on a VBR connection

This may be best understood where the transmission on an VBR VC is in the form of fixed length messages (CPCS-PDUs), which are transmitted with some fixed interval or the Inter Message Time (IMT) and take a number of cells, MBS, to carry them; however, the description of VBR traffic and the use of the dual leaky bucket are not restricted to such situations. In this case, the average cell rate over the interval of IMT is the SCR (=MBS/IMT). The individual messages can be transmitted at a PCR, which can be any value between the bandwidth for the physical link ($1/\delta$) and the SCR. This allows the message to be transmitted in a period that is smaller than the message interval IMT, with gaps between instances of the message.

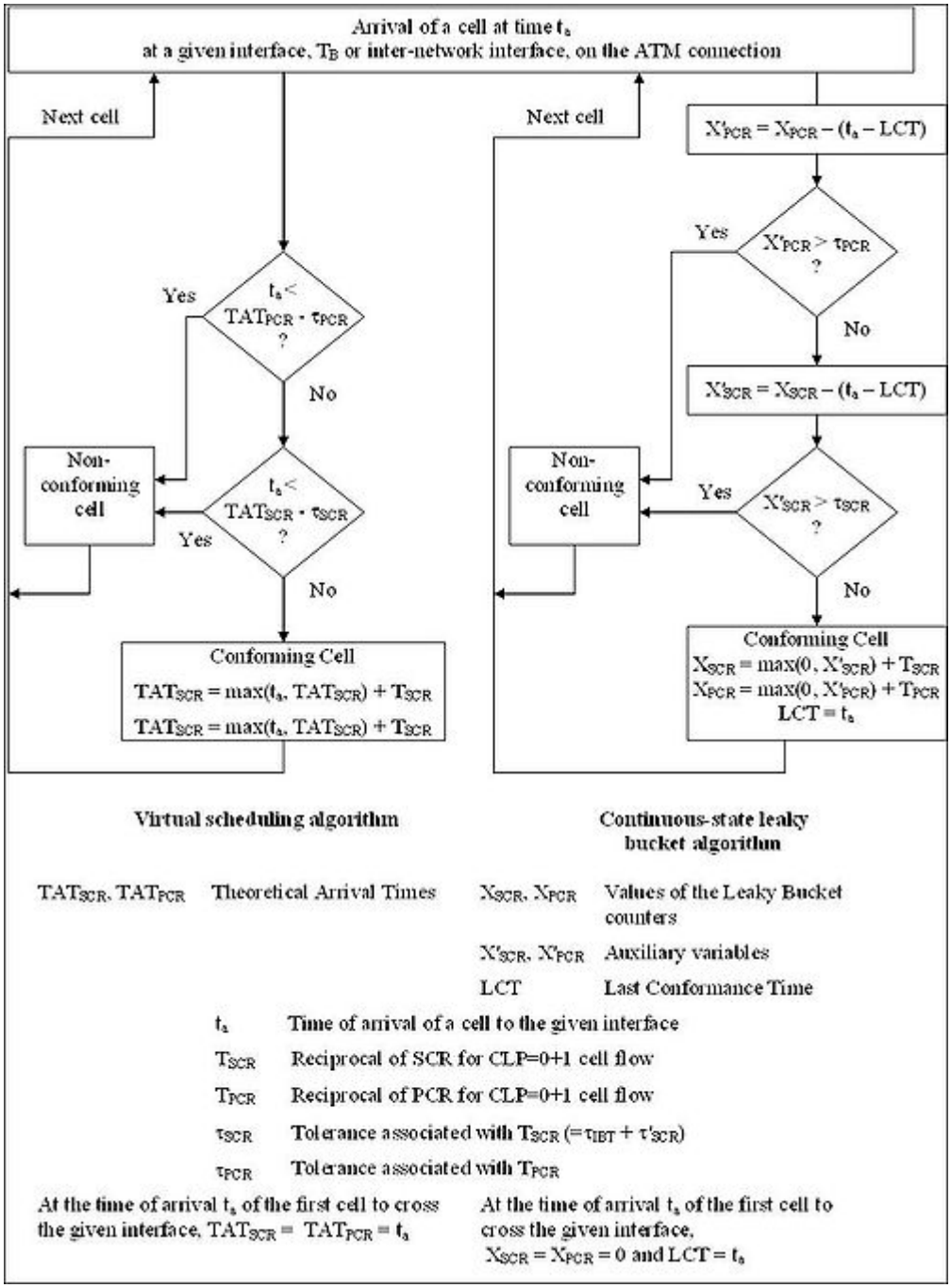

Figure 3: Reference algorithm for Sustainable Cell Rate (SCR) and Peak Cell Rate (PCR) for CLP = 0 + 1 cell flow

In the dual leaky bucket, one bucket is applied to the traffic with an emission interval of 1/SCR and a limit value τ_{SCR} that gives an MBS that is the number of cells in the message # Maximum Burst Size. The second bucket has an emission interval of 1/PCR and a limit value τ_{PCR} that allows for the CDV up to that point in the path of the connection: #Delay Variation Tolerance. Cells are then allowed through at the PCR, with jitter of τ_{PCR}, up to a

maximum number of MBS cells. The next burst of MBS cells will then be allowed through starting MBS x 1/SCR after the first.

If the cells arrive in a burst at a rate higher than 1/PCR (MBS cells arrive in less than (MBS - 1)/PCR - τ_{PCR}), or more than MBS cells arrive at the PCR, or bursts of MBS cells arrive closer than IMT apart, the dual leaky bucket will detect this and delay (shaping) or drop or de-prioritize (policing) enough cells to make the connection conform.

Figure 3 shows the reference algorithm for SCR and PCR control for both Cell Loss Priority (CLP) values 1 (low) and 0 (high) cell flows, i.e. where the cells with both priority values are treated the same. Similar reference algorithms where the high and low priority cells are treated differently are also given in Annex A to I.371 .

Chapter-4

Adaptive Quality of Service Multi-hop Routing and Grade of Service

Adaptive quality of service multi-hop routing

In multi-hop networks, **Adaptive Quality of Service routing** (AQoS or AQR) protocols have become increasingly popular and have numerous applications. One application in which it may be useful is in Mobile ad hoc networking (MANET).

Adaptive QoS routing is a cross-layer optimization adaptive routing mechanism. The cross-layer mechanism provides up-to-date local QoS information for the adaptive routing algorithm, by considering the impacts of node mobility and lower-layer link performance. The multiple QoS requirements are satisfied by adaptively using forward error correction and multipath routing mechanisms, based on the current network status. The complete routing mechanism includes three parts: (1) a modified dynamic source routing algorithm that handles route discovery and the collection of QoS related parameters; (2) a local statistical computation and link monitoring function located in each node; and (3) an integrated decision-making system to calculate the number of routing paths, coding parity length, and traffic distribution rates.

The adaptive cooperation concept has future promisses to overcome infrastructure loaded approaches and to get rid of central facilities with autonomous networks in industrial and home applications.

The United States Air Force is determining the best way to employ QoS protocols into Airborne Networking. Research composed shows that **Adaptive QoS** that uses cross-layer cooperation has provided the best results for military applications.

Introduction

A wireless ad hoc network consists of a collection of mobile nodes interconnected by multihop wireless paths with wireless transmitters and receivers. Such networks can be spontaneously created and operated in a self-organized manner, because they do not rely upon any preexisting network infrastructure.

The emergence of multimedia applications in communications has generated the need to provide mobile quality-of-service (QoS) support in ad hoc networks, and such applications require a stable path to guarantee QoS requirements. However, the topology of ad hoc networks is highly dynamic due to the unpredictable node mobility. In addition, wireless channel bandwidth is limited. So, QoS provisioning in such networks is complex and challenging.

QoS routing usually involves two tasks: collecting and maintaining up-to-date state information about the network and finding feasible paths for a connection based on its QoS requirements. Many approaches currently exist to perform QoS routing, most of which consist of routing across the Network layer of the OSI model only. Some approaches utilize both the Network and Data link layer but do not consider the cross layer behaviors. This makes quantifying the QoS parameters difficult and leads to considerations of QoS but does not guarantee QoS.

To address this problem, appropriate cross-layer cooperation is required. Adaptive QoS schemes provide QoS information by factoring the impacts of node mobility and lower-layer link parameters into QoS performance.

Traditional QoS Approaches

Most QoS approaches tend to focus on only one QoS parameter (e.g., packet loss, end-to-end delay, and bandwidth). For example, while many of the QoS-related schemes are successful in reducing packet loss by adding redundancy in the packet, they do this at the expense of end-to-end delay. Because packet loss and end-to-end delay are inversely related, it may not be possible to find a path that simultaneously satisfies the delay, packet loss, and bandwidth constraints. Some proposed QoS routing algorithms do consider multiple metrics, but without considering cross-layer cooperation. Multipath routing is another type of QoS routing that has received much attention, since it can provide load balancing, fault tolerance, and higher aggregate bandwidth. Although this approach decreases packet loss and end-to-end delay, it is only efficient and reliable if a relationship can be found between the number of paths and QoS constraints.

Adaptive QoS

Adaptive QoS is a cross-layer cooperation mechanism that supports adaptive multipath routing with multiple QoS constraints in an ad hoc network. The cross-layermechanism provides information on link performance for the QoS routing. It treats traffic distribution, wireless link characteristics, and node mobility in an integrated fashion. That is, it reflects the impacts of lower-layer parameters on QoS performance in higher layers, with emphasis on translating these parameters into QoS parameters for the higher-layer connections. Amultiobjective optimization algorithm is used to calculate routing parameters using the cross-layer mechanism. These parameters are adapted to the current network status, determining the number of routing paths and code parity lengths for Forward Error Correction (FEC). In addition, a traffic engineering strategy is used to evenly distribute traffic over multiple paths.

Adaptive QoS Scheme Overview

To implement an adaptive multipath routing scheme, three functions distributed in different parts of the network are needed. First, a modified dynamic source routing function is needed. It handles route discovery and collecting the local QoS-related information along the selected routes. Second, there is a local statistical computation and link monitoring function located in each node. This function is used to support the above routing function. It will manage and build the local routing information in each node, which includes a QoS-related table. The third function will be in charge of the final decision-making process. The adaptive routing parameters are derived from the decision-making algorithm based on the QoS constraints. They are the number N of selected paths, parity length k of the FEC, code and the set {R} of the traffic distribution rates on each path. With these functions, adaptive multipath QoS routing is implemented.

QoS requirements can be based on either a delay or a delay and bandwidth requirement, or a packet loss requirement. FEC parity length is derived from the difference between the QoS delay requirement and the average delay on selected paths under the packet-loss constraint. Average packet loss under this FEC scheme is achieved by using multiple routing paths. At the same time, the packet distribution rate on each path is determined under fair packet-loss and load-balance principles. Routing maintenance under the same QoS guarantees is achieved without increasing its computational complexity.

Adaptive QoS Performance

Three functions (routing function, local statistic computation and monitoring function, and integrated decisionmaking function) are implemented in the different parts of the mobile network. Due to the distributed structure, the computation and implementation complexity of the routing scheme are reduced. Also, since routes are discovered based on the up-to-date local information and selected by the optimization computation, routing parameters (e.g., number of paths, FEC parity length, and traffic distribution rate) are dynamic and optimized. In addition to supporting multiple QoS requirements, traffic balancing and bandwidth resources are factored into our decisionmaking process. The distributed structure of the local QoS statistics used in the routing enables this QoS support mechanism to be scalable in mobile networks. Simulation results indicate that the performance (i.e., packet loss and end-to-end delay) are much better and less susceptible to the state changes (i.e., node mobility, transmission power, channel characteristics, and the traffic pattern) of the network, compared to a nonadaptive routing strategy.

Grade of service

In telecommunication engineering, and in particular teletraffic engineering, the quality of voice service is specified by two measures: the **grade of service** (**GoS**) and the **quality of service** (**QoS**).

Grade of service is the probability of a call in a circuit *group* being blocked or delayed for more than a specified interval, expressed as a vulgar fraction or decimal fraction. This is always with reference to the busy hour when the traffic intensity is the greatest. Grade of service may be viewed independently from the perspective of incoming versus outgoing calls, and is not necessarily equal in each direction or between different source-destination pairs.

On the other hand, the **quality of service** which a *single* circuit is designed or conditioned to provide, e.g. voice grade or program grade is called the quality of service. Quality criteria for such circuits may include equalization for amplitude over a specified band of frequencies, or in the case of digital data transported via analogue circuits, may include equalization for phase. Criteria for mobile quality of service in cellular telephone circuits include the probability of abnormal termination of the call.

What is Grade of Service and how is it measured?

When a user attempts to make a telephone call, the routing equipment handling the call has to determine whether to accept the call, reroute the call to alternative equipment, or reject the call entirely. Rejected calls occur as a result of heavy traffic loads (congestion) on the system and can result in the call either being delayed or lost. If a call is delayed, the user simply has to wait for the traffic to decrease, however if a call is lost then it is removed from the system.

The Grade of Service is one aspect of the quality a customer can expect to experience when making a telephone call. In a Loss System, the Grade of Service is described as that proportion of calls that are lost due to congestion in the busy hour. For a Lost Call system, the Grade of Service can be measured using *Equation 1*.

$$\text{Grade of Service} = \frac{\text{number of lost calls}}{\text{number of offered calls}} \qquad (1)$$

For a delayed call system, the Grade of Service is measured using three separate terms:

- The mean delay t_d – Describes the average time a user spends waiting for a connection if their call is delayed.
- The mean delay t_o – Describes the average time a user spends waiting for a connection whether or not their call is delayed.
- The probability that a user may be delayed longer than time t while waiting for a connection. Time t is chosen by the telecommunications service provider so that they can measure whether their services conform to a set Grade of Service.

Where and when is Grade of Service measured?

The Grade of Service can be measured using different sections of a network. When a call is routed from one end to another, it will pass through several exchanges. If the Grade of Service is calculated based on the number of calls rejected by the final circuit group, then

the Grade of Service is determined by the final circuit group blocking criteria. If the Grade of Service is calculated based on the number of rejected calls between exchanges, then the Grade of Service is determined by the exchange-to-exchange blocking criteria.

The Grade of Service should be calculated using both the access networks and the core networks as it is these networks that allow a user to complete an end-to-end connection. Furthermore, the Grade of Service should be calculated from the average of the busy hour traffic intensities of the 30 busiest traffic days of the year. This will cater for most scenarios as the traffic intensity will seldom exceed the reference level.

Class of Service

Different telecommunications applications require different Qualities of Service. For example, if a telecommunications service provider decides to offer different qualities of voice connection, then a premium voice connection will require a better connection quality compared to an ordinary voice connection. Thus different Qualities of Service are appropriate, depending on the intended use. To help telecommunications service providers to market their different services, each service is placed into a specific class. Each Class of Service determines the level of service required.

To identify the Class of Service for a specific service, the network's switches and routers examine the call based on several factors. Such factors can include:

- The type of service and priority due to precedence
- The identity of the initiating party
- The identity of the recipient party

Quality of Service in broadband networks

In broadband networks, the Quality of Service is measured using two criteria. The first criterion is the probability of packet losses or delays in already accepted calls. The second criterion refers to the probability that a new incoming call will be rejected or blocked. To avoid the former, broadband networks limit the number of active calls so that packets from established calls will not be lost due to new calls arriving. As in circuit-switched networks, the Grade of Service can be calculated for individual switches or for the whole network.

Maintaining a Grade of Service

The telecommunications provider is usually aware of the required Grade of Service for a particular product. To achieve and maintain a given Grade of Service, the operator must ensure that sufficient telecommunications circuits or routes are available to meet a specific level of demand. It should also be kept in mind that too many circuits will create a situation where the operator is providing excess capacity which may never be used, or at the very least may be severely underutilized. This adds costs which must be borne by other parts of the network. To determine the correct number of circuits that are required,

telecommunications service providers make use of Traffic Tables. An example of a Traffic Table can be viewed in *Figure 1*. It follows that in order for a telecommunications network to continue to offer a given Grade of Service, the number of circuits provided in a circuit group must increase (non-linearly) if the traffic intensity increases.

Erlang's lost call assumptions

To calculate the Grade of Service of a specified group of circuits or routes, A.K. Erlang used a set of assumptions that relied on the network losing calls when all circuits in a group were busy. These assumptions are:

- All traffic through the network is pure-chance traffic, i.e. all call arrivals and terminations are independent random events
- There is statistical equilibrium, i.e., the average number of calls does not change
- Full availability of the network, i.e., every outlet from a switch is accessible from every inlet
- Any call that encounters congestion is immediately lost.

From these assumptions Erlang developed the Erlang-B formula which describes the probability of congestion in a circuit group. The probability of congestion gives the Grade of Service experienced.

Calculating the Grade of Service

To determine the Grade of Service of a network when the traffic load and number of circuits are known, telecommunications network operators make use of *Equation 2*, which is the Erlang-B equation.

$$\text{Grade of Service} = \frac{\left(\frac{A^N}{N!}\right)}{\left(\sum_{k=0}^{N} \frac{A^k}{k!}\right)} \qquad (2)$$

A = Expected traffic intensity in Erlangs, N = Number of circuits in group.

This equation allows operators to determine whether each of their circuit groups meet the required Grade of Service, simply by monitoring the reference traffic intensity.

(For delay networks, the Erlang-C formula allows network operators to determine the probability of delay depending on peak traffic and the number of circuits.)

Chapter-5

Long-tail Traffic

The terms ***long-range dependent***, ***self-similar*** and ***heavy-tailed*** are very close in meaning. Differences in nomenclature hint at the origins and application fields of the terms. These are somewhat different but closely related phenomena.

A **long-tailed** or **heavy-tailed** probability distribution is one that assigns relatively high probabilities to regions far from the mean or median. A more formal mathematical definition is given below. In the context of teletraffic engineering a number of quantities of interest have been shown to have a long-tailed distribution. For example, if we consider the sizes of files transferred from a web-server, then, to a good degree of accuracy, the distribution is heavy-tailed, that is, there are a large number of small files transferred but, crucially, the number of very large files transferred remains a major component of the volume downloaded.

Many processes are technically long-range dependent but not self-similar. The differences between these two phenomena are subtle. Heavy-tailed refers to a probability distribution, and long-range dependent refers to a property of a time series and so these should be used with care and a distinction should be made. The terms are distinct although superpositions of samples from heavy-tailed distributions aggregate to form long-range dependent time series.

Additionally there is Brownian motion which is self-similar but not long-range dependent.

Overview

The design of robust and reliable networks and network services has become an increasingly challenging task in today's Internet world. To achieve this goal, understanding the characteristics of Internet traffic plays a more and more critical role. Empirical studies of measured traffic traces have led to the wide recognition of self-similarity in network traffic.

Self-similar Ethernet traffic exhibits dependencies over a long range of time scales. This is to be contrasted with telephone traffic which is Poisson in its arrival and departure process .

With many time-series if the series is averaged then the data begins to look smoother. However, with self-similar data, one is confronted with traces which are spiky and bursty, even at large scales. Such behaviour is caused by strong dependence in the data: large values tend to come in clusters, and clusters of clusters, etc. This can have far-reaching consequences for network performance .

Heavy-tail distributions have been observed in many natural phenomena including both physical and sociological phenomena. Mandelbrot established the use of heavy-tail distributions to model real-world fractal phenomena, e.g. Stock markets, earthquakes, and the weather . Ethernet, WWW, SS7, TCP, FTP, TELNET and VBR video (digitised video of the type that is transmitted over ATM networks) traffic is self-similar .

Self-similarity in packetised data networks can be caused by the distribution of file sizes, human interactions and/or Ethernet dynamics . Self-similar and long-range dependent characteristics in computer networks present a fundamentally different set of problems to people doing analysis and/or design of networks, and many of the previous assumptions upon which systems have been built are no longer valid in the presence of self-similarity .

Short-range dependence vs. long-range dependence

Long-range and short-range dependent processes are characterised by their autocovariance functions.

In short-range dependent processes, the coupling between values at different times decreases rapidly as the time difference increases.

- The sum of the autocorrelation function over all lags is finite.
- As the lag increases, the autocorrelation function of short-range dependent processes decays quickly.

In long-range processes, the correlations at longer time scales are more significant.

- The area under the autocorrelation function summed over all lags is infinite .
- The decay of the autocorrelation function is often assumed to have the specific functional form,

$$\rho(k) \sim k^{-\alpha}$$

where $\rho(k)$ is the autocorrelation function at a lag k, α is a parameter in the interval (0,1) and the ~ means asymptotically proportional to as k approaches infinity.

The Poisson distribution and traffic

Before the heavy-tail distribution is introduced mathematically, the memoryless Poisson distribution, used to model traditional telephony networks, is briefly reviewed below.

Assuming pure-chance arrivals and pure-chance terminations leads to the following:

- The number of call arrivals in a given time has a Poisson distribution, i.e.:

$$P(a) = \left(\frac{\mu^a}{a!}\right) e^{-\mu},$$

where a is the number of call arrivals and μ is the mean number of call arrivals in time T. For this reason, pure-chance traffic is also known as Poisson traffic.

- The number of call departures in a given time also has a Poisson distribution, i.e.:

$$P(d) = \left(\frac{\lambda^d}{d!}\right) e^{-\lambda},$$

where d is the number of call departures and λ is the mean number of call departures in time T.

- The intervals, T, between call arrivals and departures are intervals between independent, identically distributed random events. It can be shown that these intervals have a negative exponential distribution, i.e.:

$$P[T \geq t] = e^{\frac{-t}{h}},$$

where h is the Mean Holding Time (MHT) .

Information on the fundamentals of statistics and probability theory can be found in the external links section.

The heavy-tail distribution

Heavy-tail distributions have properties that are qualitatively different from commonly used (memoryless) distributions such as the Poisson distribution.

The Hurst parameter H is a measure of the level of self-similarity of a time series that exhibits long-range dependence, to which the heavy-tail distribution can be applied. H takes on values from 0.5 to 1. A value of 0.5 indicates the data is uncorrelated or has only short-range correlations. The closer H is to 1, the greater the degree of persistence or long-range dependence.

Typical values of the Hurst parameter, *H*:

- Any pure random process has $H = 0.5$
- Phenomena with $H > 0.5$ typically have a complex process structure.

A distribution is said to be heavy-tailed if:

$$P[X > x] \sim x^{-\alpha}, \text{ as } x \to \infty, 0 < \alpha < 2$$

This means that regardless of the distribution for small values of the random variable, if the asymptotic shape of the distribution is hyperbolic, it is heavy-tailed. The simplest heavy-tail distribution is the Pareto distribution which is hyperbolic over its entire range. Complementary distribution functions for the exponential and Pareto distributions are shown below. Shown on the left is a graph of the distributions shown on linear axes, spanning a large domain. To its right is a graph of the complementary distribution functions over a smaller domain, and with a logarithmic range.

If the logarithm of the range of an exponential distribution is taken, the resulting plot is linear. In contrast, that of the heavy-tail distribution is still curvilinear. These characteristics can be clearly seen on the graph above to the right. A characteristic of long-tail distributions is that if the logarithm of both the range and the domain is taken, the tail of the long-tail distribution is approximately linear over many orders of magnitude. In the graph above left, the condition for the existence of a heavy-tail distribution, as previously presented, is not met by the curve labelled "Gamma-Exponential Tail".

The probability mass function of a heavy-tail distribution is given by:

$$p(x) = \alpha k^{\alpha} x^{-\alpha-1}, \ \alpha, k > 0, \ x \geq k$$

and its cumulative distribution function is given by:

$$F(x) = P[X \leq \ x] = 1 - (\frac{k}{x})^{\alpha}$$

where *k* represents the smallest value the random variable can take.

Readers interested in a more rigorous mathematical treatment of the subject are referred to the external links section.

What causes long-tail traffic?

In general, there are three main theories for the causes of long-tail traffic. First, is a cause based in the application layer which theorizes that user session durations vary with a long-tail distribution due to the file size distribution. If the distribution of file sizes is

heavy-tailed then the superposition of many file transfers in a client/server network environment will be long-range dependent. Additionally, this causal mechanism is robust with respect to changes in network resources (bandwidth and buffer capacity) and network topology. This is currently the most popular explanation in the engineering literature and the one with the most empirical evidence through observed file size distributions.

Second, is a transport layer cause which theorizes that the feedback between multiple TCP streams due to TCP's congestion avoidance algorithm in moderate to high packet loss situations causes self-similar traffic or at least allows it to propagate. However, this is believed only to be a significant factor at relatively short timescales and not the long-term cause of self-similar traffic.

Finally, is a theorized link layer cause which is predicated based on physics simulations of packet switching networks on simulated topologies. At a critical packet creation rate, the flow in a network becomes congested and exhibits 1/f noise and long-tail traffic characteristics. There have been criticisms on these sorts of models though as being unrealistic in that network traffic is long-tailed even in non-congested regions and at all levels of traffic.

showed in simulation that long-range dependence could arise in the queue length dynamics at a given node (an entity which transfers traffic) within a communications network even when the traffic sources are free of long-range dependence. The mechanism for this is believed to relate to feedback from routing effects in the simulation.

Modelling long-tail traffic

Modelling of long-tail traffic is necessary so that networks can be provisioned based on accurate assumptions of the traffic that they carry. The dimensioning and provisioning of networks that carry long-tail traffic is discussed in the next section.

Since (unlike traditional telephony traffic) packetised traffic exhibits self-similar or fractal characteristics, conventional traffic models do not apply to networks which carry long-tail traffic. Previous analytic work done in Internet studies adopted assumptions such as exponentially-distributed packet inter-arrivals, and conclusions reached under such assumptions may be misleading or incorrect in the presence of heavy-tailed distributions .

It has for long been realised that efficient and accurate modelling of various real world phenomena needs to incorporate the fact that observations made on different scales each carry essential information. In most simple terms, representing data on large scales by its mean is often useful (such as an average income or an average number of clients per day) but can be inappropriate (e.g. in the context of buffering or waiting queues).

With the convergence of voice and data, the future multi-service network will be based on packetised traffic, and models which accurately reflect the nature of long-tail traffic will be required to develop, design and dimension future multi-service networks . We seek an equivalent to the Erlang model for circuit switched networks .

There is not an abundance of heavy-tailed models with rich sets of accompanying data fitting techniques. A clear model for fractal traffic has not yet emerged, nor is there any definite direction towards a clear model. Deriving mathematical models which accurately represent long-tail traffic is a fertile area of research.

Gaussian models, even long-range dependent Gaussian models, are unable to accurately model current Internet traffic. Classical models of time series such as Poisson and finite Markov processes rely heavily on the assumption of independence, or at least weak dependence. Poisson and Markov related processes have, however, been used with some success. Nonlinear methods are used for producing packet traffic models which can replicate both short-range and long-range dependent streams.

A number of models have been proposed for the task of modelling long-tail traffic. These include the following:

- Fractional ARIMA
- Fractional Brownian motion
- Iterated Chaotic Maps
- Infinite Markov Modulated Processes
- Poisson Pareto Burst Processes (PPBP)
- Markov Modulated Poisson Processes (MMPP)
- Multi-fractal models
- Matrix models
- Wavelet Modelling

No unanimity exists about which of the competing models is appropriate , but the Poisson Pareto Burst Process (PPBP), which is an M/G/∞ process, is perhaps the most successful model to date. It is demonstrated to satisfy the basic requirements of a simple, but accurate, model of long-tail traffic .

Finally, results from simulations (taken from) using α-stable stochastic processes for modelling traffic in broadband networks are presented. The simulations are compared to a variety of empirical data (Ethernet, WWW, VBR Video).

The graph on the left shows the model's simulation results for Ethernet traffic. On its right is shown measured Ethernet traffic. The model appears to appear to represent the empirical traffic well.

The graph on the left shows the model's simulation results for WWW traffic. On its right is shown measured WWW traffic. Here, too, the model appears to appear to represent the empirical traffic well.

Network performance

In some cases an increase in the Hurst parameter can lead to a reduction in network performance. The extent to which heavy-tailedness degrades network performance is determined by how well congestion control is able to shape source traffic into an on-average constant output stream while conserving information . Congestion control of heavy-tailed traffic is discussed in the following section.

Traffic self-similarity negatively affects primary performance measures such as queue size and packet-loss rate. The queue length distribution of long-tail traffic decays more slowly than with Poisson sources. However, long-range dependence implies nothing about its short-term correlations which affect performance in small buffers . For heavy-tailed traffic, extremely large bursts occur more frequently than with light-tailed traffic . Additionally, aggregating streams of long-tail traffic typically intensifies the self-similarity ("burstiness") rather than smoothing it, compounding the problem .

The graph above right, taken from , presents a queueing performance comparison between traffic streams of varying degrees of self-similarity. Note how the queue size increases with increasing self-similarity of the data, for any given channel utilisation, thus degrading network performance.

In the modern network environment with multimedia and other QoS sensitive traffic streams comprising a growing fraction of network traffic, second order performance measures in the form of “jitter” such as delay variation and packet loss variation are of import to provisioning user specified QoS. Self-similar burstiness is expected to exert a negative influence on second order performance measures .

Packet switching based services, such as the Internet (and other networks that employ IP) are best-effort services, so degraded performance, although undesirable, can be tolerated. However, since the connection is contracted, ATM networks need to keep delays and jitter within negotiated limits .

Self-similar traffic exhibits the persistence of clustering which has a negative impact on network performance.

- With Poisson traffic (found in conventional telephony networks), clustering occurs in the short term but smooths out over the long term.
- With long-tail traffic, the bursty behaviour may itself be bursty, which exacerbates the clustering phenomena, and degrades network performance .

Many aspects of network quality of service depend on coping with traffic peaks that might cause network failures, such as

- Cell/packet loss and queue overflow
- Violation of delay bounds e.g. In video
- Worst cases in statistical multiplexing

Poisson processes are well-behaved because they are stateless, and peak loading is not sustained, so queues do not fill. With long-range order, peaks last longer and have greater impact: the equilibrium shifts for a while .

Due to the increased demands that long-tail traffic places on networks resources, networks need to be carefully provisioned to ensure that quality of service and service level agreements are met. The following subsection deals with the provisioning of standard network resources, and the subsection after that looks at provisioning web servers which carry a significant amount of long-tail traffic.

Network provisioning for long-tail traffic

For network queues with long-range dependent inputs, the sharp increase in queuing delays at fairly low levels of utilisation and slow decay of queue lengths implies that an incremental improvement in loss performance requires a significant increase in buffer size .

While throughput declines gradually as self-similarity increases, queuing delay increases more drastically. When traffic is self-similar, we find that queuing delay grows proportionally to the buffer capacity present in the system. Taken together, these two observations have potentially dire implications for QoS provisions in networks. To achieve a constant level of throughput or packet loss as self-similarity is increased, extremely large buffer capacity is needed. However, increased buffering leads to large queuing delays and thus self-similarity significantly steepens the trade-off curve between throughput/ packet loss and delay.

ATM can be employed in telecommunications networks to overcome second order performance measure problems. The short fixed length cell used in ATM reduces the delay and most significantly the jitter for delay-sensitive services such as voice and video.

Web site provisioning for long-tail traffic

Workload pattern complexities (for example, bursty arrival patterns) can significantly affect resource demands, throughput, and the latency encountered by user requests, in terms of higher average response times and higher response time variance. Without adaptive, optimal management and control of resources, SLAs based on response time are impossible. The capacity requirements on the site are increased while its ability to provide acceptable levels of performance and availability diminishes . Techniques to control and manage long-tail traffic are discussed in the following section.

The ability to accurately forecast request patterns is an important requirement of capacity planning. A practical consequence of burstiness and heavy-tailed and correlated arrivals is difficulty in capacity planning.

With respect to SLAs, the same level of service for heavy-tailed distributions requires a more powerful set of servers, compared with the case of independent light-tailed request traffic. To guarantee good performance, focus needs to be given to peak traffic duration because it is the huge bursts of requests that most degrade performance. That is why some busy sites require more head room (spare capacity) to handle the volumes; for example, a high-volume online trading site reserves spare capacity with a ratio of three to one.

Reference to additional information on the effect of long-range dependency on network performance can be found in the external links section.

Controlling long-tail traffic

Given the ubiquity of scale-invariant burstiness observed across diverse networking contexts, finding an effective traffic control algorithm capable of detecting and managing self-similar traffic has become an important problem. The problem of controlling self-similar network traffic is still in its infancy.

Traffic control for self-similar traffic has been explored on two fronts: Firstly, as an extension of performance analysis in the resource provisioning context, and secondly, from the multiple time scale traffic control perspective where the correlation structure at large time scales is actively exploited to improve network performance.

The resource provisioning approach seeks to identify the relative utility of the two principal network resource types - bandwidth and buffer capacity - with respect to their curtailing effects on self-similarity, and advocates a small buffer/ large bandwidth resource dimensioning policy. Whereas resource provisioning is open-loop in nature, multiple time scale traffic control exploits the long-range correlation structure present in self-similar traffic . Congestion control can be exercised concurrently at multiple time scales, and by cooperatively engaging information extracted at different time scales, achieve significant performance gains.

Another approach adopted in controlling long-tail traffic makes traffic controls cognizant of workload properties. For example, when TCP is invoked in HTTP in the context of web client/ server interactions, the size of the file being transported (which is known at the server) is conveyed or made accessible to protocols in the transport layer, including the selection of alternative protocols, for more effective data transport. For short files, which constitute the bulk of connection requests in heavy-tailed file size distributions of web servers, elaborate feedback control may be bypassed in favour of lightweight mechanisms in the spirit of optimistic control, which can result in improved bandwidth utilization.

found that the simplest way to control packet traffic is to limit the length of queues. Long queues in the network invariably occur at hosts (entities that can transmit and receive packets). Congestion control can therefore be achieved by reducing the rate of packet production at hosts with long queues.

It should be noted that long-range dependence and its exploitation for traffic control is best suited for flows or connections whose lifetime or connection duration is long lasting.

Chapter-6

Circuit Switching and Least-cost Routing

Circuit switching

Circuit switching is a telecommunications technology by which two network nodes establish a dedicated communications channel (circuit) connecting them for the duration of the communication session before the nodes may communicate. The circuit functions as if the nodes were physically connected with an electrical circuit.

The bit delay is constant during a connection, as opposed to packet switching, where packet queues may cause varying packet transfer delay. Each circuit cannot be used by other callers until the circuit is released and a new connection is set up. Even if no actual communication is taking place in a dedicated circuit that channel remains unavailable to other users. Channels that are available for new calls to be set up are said to be idle.

Virtual circuit switching is a packet switching technology that may emulate circuit switching, in the sense that the connection is established before any packets are transferred, and that packets are delivered in order.

There is a common misunderstanding that circuit switching is used only for connecting voice circuits (analog or digital). The concept of a dedicated path persisting between two communicating parties or nodes can be extended to signal content other than voice. Its advantage is that it provides for non-stop transfer without requiring packets and without most of the overhead traffic usually needed, making maximal and optimal use of available bandwidth for that communication. The disadvantage of inflexibility tends to reserve it for specialized applications, particularly with the overwhelming proliferation of internet-related technology.

The call

For call setup and control (and other administrative purposes), it is possible to use a separate dedicated signalling channel from the end node to the network. ISDN is one such service that uses a separate signalling channel while Plain Old Telephone Service (POTS) does not.

The method of establishing the connection and monitoring its progress and termination through the network may also utilize a separate control channel as in the case of links between telephone exchanges which use CCS7 packet-switched signalling protocol to communicate the call setup and control information and use TDM to transport the actual circuit data.

Early telephone exchanges are a suitable example of circuit switching. The subscriber would ask the operator to connect to another subscriber, whether on the same exchange or via an inter-exchange link and another operator. In any case, the end result was a physical electrical connection between the two subscribers' telephones for the duration of the call. The copper wire used for the connection could not be used to carry other calls at the same time, even if the subscribers were in fact not talking and the line was silent.

Compared to datagram packet switching

Since the first days of the telegraph it has been possible to multiplex multiple connections over the same physical conductor, but nonetheless each channel on the multiplexed link was either dedicated to one call at a time, or it was idle between calls.

With circuit switching, and virtual circuit switching, a route is reserved from source to destination. The entire message is sent in order so that it does not have to be reassembled at the destination. Circuit switching can be relatively inefficient because capacity is wasted on connections which are set up but are not in continuous use (however momentarily). On the other hand, the connection is immediately available and capacity is guaranteed until the call is disconnected.

Circuit switching contrasts with packet switching, which splits traffic data (for instance, digital representation of sound, or computer data) into chunks, called packets, that are routed over a shared network.

Packet switching is the process of segmenting a message/data to be transmitted into several smaller packets. Each packet is labeled with its destination and a sequence number (for ordering related packets), precluding the need for a dedicated path to help the packet find its way to its destination. Each is dispatched and many may go via different routes. At the destination, the original message is reassembled in the correct order, based on the packet number. Datagram packet switching networks do not require a circuit to be established and allow many pairs of nodes to communicate almost simultaneously over the same channel.

Examples of circuit switched networks

- Public Switched Telephone Network (PSTN)
- ISDN B-channel
- Circuit Switched Data (CSD) and High-Speed Circuit-Switched Data (HSCSD) service in cellular systems such as GSM
- Datakit

- X.21 (Used in the German DATEX-L and Scandinavian DATEX circuit switched data network)
- DisplayPort

Least-cost routing

In voice telecommunications, **least cost routing** (LCR) is the process of selecting the path of outbound communications traffic based on cost. Within a telecoms carrier, an LCR team might periodically (monthly, weekly or even daily) choose between routes from several or even hundreds of carriers for destinations across the world. This function might also be automated by a device or software program known as a "Least Cost Router."

Telecoms carriers as suppliers and customers

Telecoms carriers often buy and sell call termination services with other carriers. A carrier such as Telewest or France Telecom will be interconnected with other telecoms carriers and might have a number of routing options of different price, quality and capacity to a given country. In the de-regulated EU, these will be licensed alternative operators (e.g. Cable and Wireless / Colt in the UK or Jazztel in Spain) or the (PTT)'s of other countries, such as T-Systems (Germany), Telefonica (Spain), NTT (Japan) or Telstra (Australia), who establish offices or a point of presence (POP) in a major telecommunications hub city such as London, New York, Hong Kong or Amsterdam. The major US carriers, Sprint, Verizon, AT&T and Global Crossing in the US also have POPs in these hub cities. There are also *niche carriers* which specialise in providing termination to a small number of destinations, sometimes through the use of grey routes.

"Trading" in the telecom carrier-carrier market is very different than the "trading" conducted in financial markets by brokers and banks. Whereas brokers and banks may buy and sell the same stocks or bonds with each other in the same day, carriers have to be very careful not to do so. For example, if carrier A buys Venezuela from carrier B who buys it from A, one call will come in to carrier A, go to B and go back to A again, over and over until all the circuits are taken up with one call. If it does terminate on an overflow route, the carriers may bill each other many times over for the same call. This is called *looping* and is very undesirable.

Buying->costing->routing->pricing->margin management cycle

The LCR team in a carrier might follow a cycle:

(1) The buyers negotiate with their suppliers and get a new price schedule. (2) The prices are loaded into software to calculate and compare termination costs. (3) A route is chosen, fixing a cost-for-pricing, and new prices are issued based on the costs-for-

pricing. (4) The new routes are implemented on the switch and finally the traffic volumes and margins are monitored through reports from the billing system. (5) Loss-making traffic and odd routings are investigated, and either the billing system has its data corrected or routing and pricing action is taken.

Carriers sign interconnect agreements with each other specifying the terms under which they will do business. Such agreements define terms of payment, methods and procedures of dispute resolution, and the means by which the carriers will notify each other of pricing changes. The industry standard is currently seven days for price increases while price decreases often take effect on the day of notification. Because the margins in the carrier-carrier market are extremely slim, re-routes or price increases must be made quickly to a destination where the current route is going to increase in price. Since the price increase itself has seven days' notice, it must be issued within twenty-four hours of the cost increase to avoid losses. This can put significant pressures on a carrier's LCR team, who must process the offers from their suppliers quickly and accurately.

Impact of Mobile Number Portability in the VOIP and LCR Environments

Mobile number portability impacts the internet telephony, VOIP (Voice over IP) and least cost routing (LCR) businesses. Mobile number portability is a service that makes it possible for subscribers to keep their existing mobile phone number when changing the service provider (or mobile operator). With number portability now in place in many countries, LCR providers can no longer rely on using only a portion of the dialed telephone number to route a call. Instead, they now need to discover the actual current network of every number before routing the call. Thus, LCR solutions also need to handle MNP when routing a voice call. In countries without a central database like UK it might be necessary to query the GSM network about the home network a mobile phone number belongs to.

MNP checks are important to assure that this quality of service is met; by handling MNP lookups before routing a call and assuring that the voice call will actually work, VOIP companies give businesses the necessary reliability they look for in an internet telephony provider.

In countries such as Singapore, the most recent Mobile number portability solution is expected to open the doors to new business opportunities for non-traditional telecommunication service providers like wireless broadband providers and voice over IP (VOIP) providers.

In November 2008 the United States' FCC (Federal Communications Commission) released an order extending number portability obligations to interconnected VOIP providers and carriers that support VOIP providers.

Number plan management and analysis

Whereas markets in commodities such as pork bellies or oil have agreed definitions and arbitrating bodies for the commodities they trade, the carrier-carrier market has no agreed definitions of its destinations. Every carrier uses the International Telecommunication Union E.164 standard for country codes, but each carrier uses different codes for destinations within a country, usually because it is using different suppliers within that country. So one carrier's codes for Cali may not be the same as another's. This applies especially to mobile operators. While the difference in price between a call to a land-line and a call to a mobile may not seem much at 9 UK pence, the volumes can be one hundred thousand minutes a day or more, leading to losses of over £250,000 a month. When a carrier's dial code table can contain three thousand items, comparing codes is a critical and complex part of the process. The theory of dial code relationships actually involves the mathematical theory of lattices and code comparisons have to be done with computer software.

Number plan management monitors changes in suppliers' dial codes and adds or removes codes from the company's own code tables to improve costs. Implementing the changes across the company's switches, billing systems, calling card and other IN platforms is a significant task for the engineering and billing departments.

Cherry-picking

One aim of LCR teams is to *cherry-pick*. This happens when Carrier A's team finds that Carrier B defines a code range as being fixed-line and so cheap, while Carrier A defines it as mobile and so more expensive. Carrier A will send that range to Carrier B, pay a low fixed-line rate and charge at a high mobile rate - making much more profit. Carrier B will sustain losses if it does not notice that its supplier, C, also defines that range as belonging to a mobile operator and charges a higher rate. Caught in the middle, B can sustain five- or even six- figure losses in a very short time.

Route and call quality

The LCR team also has to take route and call quality into account. The quality of route to a destination can vary considerably between suppliers and even from week to week from the same supplier.

Quality is usually measured by the Answer-Seizure Ratio (ASR = call attempts answered / call attempts), Post-Dial Delay (PDD) and the Average Call Duration (ACD). If the average call duration is very low, it is taken to mean that the call quality is so poor that people cannot have a conversation and hang up. This matters to calling card operators because people do not re-purchase card services that give a low ACD. In case of significant discrepancies in ACD values across available routes, the carrier shall prioritize the routes offering higher ACD . A low ASR is taken to mean that callers cannot get through to the other end and hence that the route is congested or is of low-quality. The low ASR is not as bad as low ACD, because it suggests at least a proper answer

supervision (i.e. correct signaling), and therefore the handover mechanism can reroute calls via other available routes. An on-line monitoring system of a quality based routing is publicly available for a demo traffic . Post-dial delay is the time from dialing the last digit to the time a caller hears ringing.

Another, more sophisticated way of measuring the call quality is PESQ (Perceptual Evaluation of Speech Quality). Such measurements are rarely used in production switching systems, in particular due to the necessity of voice samples at both ends.

Additionally, the team may take into account the responsiveness of their supplier's technical team: if there is a fault or low quality, does the supplier fix it or just say that it is the best they can do?

LCR software

The key tasks LCR software must do are: load prices schedules and code tables automatically; compare dial codes correctly; turn the carriers' name-based price schedule into a dial code-dependent termination cost schedule; put costs in order; incorporate quality considerations; produce costing and routing schedules in a format suitable for pricing analysts and engineering; and transfer data into the billing system.

LCR software varies from home-grown Excel spreadsheets, through Access and Microsoft Visual Studio applications to commercial products offering integration with the switch and billing systems costing up to £500,000 for an installation. The simpler the software, the more complex the surrounding manual processes.

Routing Platforms

With VoIP becoming a predominate carrier messaging system, routing becomes available from stand alone products and hosted services. These services are built on top of the RFC 3261 3XX series messaging, which allows for stateless redirection of call signaling. These stand alone LCR systems, such as General Telecom's RouteNGN, integrate many powerful routing features such as: jurisdictional, profit margin protected routing along with standard LCR and offload routing from switching components. Other open source platforms such as OpenSER/Kamailo exist that are capable of performing redirect based routing, they are however less specifically geared toward exploiting the market niches involved in carrier telecommunications routing.

Related Ideas

Least Cost Routing is also used to describe a type of equipment installed on customers' premises. An *LCR box* is programmed with prices from the companies supplying telecoms services to that company and the box routes each call to the appropriate supplier.

Chapter-7

Network Congestion

In data networking and queueing theory, **network congestion** occurs when a link or node is carrying so much data that its quality of service deteriorates. Typical effects include queueing delay, packet loss or the blocking of new connections. A consequence of these latter two is that incremental increases in offered load lead either only to small increases in network throughput, or to an actual reduction in network throughput.

Network protocols which use aggressive retransmissions to compensate for packet loss tend to keep systems in a state of network congestion even after the initial load has been reduced to a level which would not normally have induced network congestion. Thus, networks using these protocols can exhibit two stable states under the same level of load. The stable state with low throughput is known as congestive collapse.

Modern networks use congestion control and network congestion avoidance techniques to try to avoid congestion collapse. These include: exponential backoff in protocols such as 802.11's CSMA/CA and the original Ethernet, window reduction in TCP, and fair queueing in devices such as routers. Another method to avoid the negative effects of network congestion is implementing priority schemes, so that some packets are transmitted with higher priority than others. Priority schemes do not solve network congestion by themselves, but they help to alleviate the effects of congestion for some services. An example of this is 802.1p. A third method to avoid network congestion is the explicit allocation of network resources to specific flows. One example of this is the use of Contention-Free Transmission Opportunities (CFTXOPs) in the ITU-T G.hn standard, which provides high-speed (up to 1 Gbit/s) Local area networking over existing home wires (power lines, phone lines and coaxial cables).

RFC 2914 addresses the subject of congestion control in detail.

Network capacity

The fundamental problem is that all network resources are limited, including router processing time and link throughput.

However:

- today's (2006) Wireless LAN effective bandwidth throughput (15-100Mbit/s) is easily filled by a single personal computer.
- Even on fast computer networks (e.g. 1 Gbit), the backbone can easily be congested by a few servers and client PCs.
- Because P2P scales very well, file transmissions by P2P have no problem filling and will fill an uplink or some other network bottleneck, particularly when nearby peers are preferred over distant peers.
- Denial of service attacks by botnets are capable of filling even the largest Internet backbone network links (40 Gbit/s as of 2007), generating large-scale network congestion

Congestive collapse

Congestive collapse (or **congestion collapse**) is a condition which a packet switched computer network can reach, when little or no useful communication is happening due to congestion. Congestion collapse generally occurs at choke points in the network, where the total incoming bandwidth to a node exceeds the outgoing bandwidth. Connection points between a local area network and a wide area network are the most likely choke points. A DSL modem is the most common small network example, with between 10 and 1000 Mbit/s of incoming bandwidth and at most 8 Mbit/s of outgoing bandwidth.

When a network is in such a condition, it has settled (under overload) into a stable state where traffic demand is high but little useful throughput is available, and there are high levels of packet delay and loss (caused by routers discarding packets because their output queues are too full) and general quality of service is extremely poor.

History

Congestion collapse was identified as a possible problem as far back as 1984 (RFC 896, dated 6 January). It was first observed on the early Internet in October 1986, when the NSFnet phase-I backbone dropped three orders of magnitude from its capacity of 32 kbit/s to 40 bit/s, and this continued to occur until end nodes started implementing Van Jacobson's congestion control between 1987 and 1988.

Cause

When more packets were sent than could be handled by intermediate routers, the intermediate routers discarded many packets, expecting the end points of the network to retransmit the information. However, early TCP implementations had very bad retransmission behavior. When this packet loss occurred, the end points sent *extra* packets that repeated the information lost; doubling the data rate sent, exactly the opposite of what should be done during congestion. This pushed the entire network into a 'congestion collapse' where most packets were lost and the resultant throughput was negligible.

Congestion control

Congestion control concerns controlling traffic entry into a telecommunications network, so as to avoid congestive collapse by attempting to avoid oversubscription of any of the processing or link capabilities of the intermediate nodes and networks and taking resource reducing steps, such as reducing the rate of sending packets. It should not be confused with flow control, which prevents the sender from overwhelming the receiver.

Theory of congestion control

The modern theory of congestion control was pioneered by Frank Kelly, who applied microeconomic theory and convex optimization theory to describe how individuals controlling their own rates can interact to achieve an "optimal" network-wide rate allocation.

Examples of "optimal" rate allocation are max-min fair allocation and Kelly's suggestion of proportional fair allocation, although many others are possible.

The mathematical expression for optimal rate allocation is as follows. Let x_i be the rate of flow i, C_l be the capacity of link l, and r_{li} be 1 if flow i uses link l and 0 otherwise. Let x, c and R be the corresponding vectors and matrix. Let $U(x)$ be an increasing, strictly convex function, called the utility, which measures how much benefit a user obtains by transmitting at rate x. The optimal rate allocation then satisfies

$$\max_x \sum_i U(x_i)$$
such that $Rx \leq c$

The Lagrange dual of this problem decouples, so that each flow sets its own rate, based only on a "price" signalled by the network. Each link capacity imposes a constraint, which gives rise to a Lagrange multiplier, p_l. The sum of these Lagrange multipliers,

$$y_i = \sum_l p_l r_{li},$$

is the price to which the flow responds.

Congestion control then becomes a distributed optimisation algorithm for solving the above problem. Many current congestion control algorithms can be modelled in this framework, with p_l being either the loss probability or the queueing delay at link l.

A major weakness of this model is that it assumes all flows observe the same price, while sliding window flow control causes "burstiness" which causes different flows to observe different loss or delay at a given link.

Classification of congestion control algorithms

There are many ways to classify congestion control algorithms:

- By the type and amount of feedback received from the network: Loss; delay; single-bit or multi-bit explicit signals
- By incremental deployability on the current Internet: Only sender needs modification; sender and receiver need modification; only router needs modification; sender, receiver and routers need modification.
- By the aspect of performance it aims to improve: high bandwidth-delay product networks; lossy links; fairness; advantage to short flows; variable-rate links
- By the fairness criterion it uses: max-min, proportional, "minimum potential delay"

Avoidance

The prevention of network congestion and collapse requires two major components:

1. A mechanism in routers to reorder or drop packets under overload,
2. End-to-end flow control mechanisms designed into the end points which respond to congestion and behave appropriately.

The correct end point behaviour is usually still to repeat dropped information, but progressively slow the rate that information is repeated. Provided all end points do this, the congestion lifts and good use of the network occurs, and the end points all get a fair share of the available bandwidth. Other strategies such as slow-start ensure that new connections don't overwhelm the router before the congestion detection can kick in.

The most common router mechanisms used to prevent congestive collapses are fair queueing and other scheduling algorithms, and random early detection, or RED, where packets are randomly dropped proactively triggering the end points to slow transmission before congestion collapse actually occurs. Fair queueing is most useful in routers at choke points with a small number of connections passing through them. Larger routers must rely on RED.

Some end-to-end protocols are better behaved under congested conditions than others. TCP is perhaps the best behaved. The first TCP implementations to handle congestion well were developed in 1984, but it was not until Van Jacobson's inclusion of an open source solution in the Berkeley Standard Distribution UNIX ("BSD") in 1988 that good TCP implementations became widespread.

UDP does not, in itself, have any congestion control mechanism. Protocols built atop UDP must handle congestion in their own way. Protocols atop UDP which transmit at a fixed rate, independent of congestion, can be troublesome. Real-time streaming protocols, including many Voice over IP protocols, have this property. Thus, special measures, such as quality-of-service routing, must be taken to keep packets from being dropped from streams.

In general, congestion in pure datagram networks must be kept out at the periphery of the network, where the mechanisms described above can handle it. Congestion in the Internet backbone is very difficult to deal with. Fortunately, cheap fiber-optic lines have reduced costs in the Internet backbone. The backbone can thus be provisioned with enough bandwidth to keep congestion at the periphery.

Practical network congestion avoidance

Implementations of connection-oriented protocols, such as the widely-used TCP protocol, generally watch for packet errors, losses, or delays in order to adjust the transmit speed. There are many different network congestion avoidance processes, since there are a number of different trade-offs available.

TCP/IP congestion avoidance

The TCP congestion avoidance algorithm is the primary basis for congestion control in the Internet.

Problems occur when many concurrent TCP flows are experiencing port queue buffer tail-drops. Then TCP's automatic congestion avoidance is not enough. All flows that experience port queue buffer tail-drop will begin a TCP retrain at the same moment - this is called TCP global synchronization.

Active Queue Management (AQM)

Purpose

"Recommendations on Queue Management and Congestion Avoidance in the Internet" (RFC 2309) states that:

- Fewer packets will be dropped with Active Queue Management (AQM).
- The link utilization will increase because less TCP global synchronization will occur.
- By keeping the average queue size small, queue management will reduce the delays and jitter seen by flows.
- The connection bandwidth will be more equally shared among connection oriented flows, even without flow-based RED or WRED.

Random early detection

One solution is to use random early detection (RED) on network equipments port queue buffer. On network equipment ports with more than one queue buffer, weighted random early detection (WRED) could be used if available.

RED indirectly signals to sender and receiver by deleting some packets, e.g. when the average queue buffer lengths are more than e.g. 50% (lower threshold) filled and deletes linearly more or (better according to paper) cubical more packets, up to e.g. 100% (higher threshold). The average queue buffer lengths are computed over 1 second at a time.

Flowbased-RED/WRED

Some network equipment are equipped with ports that can follow and measure each flow (**flowbased-RED/WRED**) and are hereby able to signal to a too big bandwidth flow according to some QoS policy. A policy could divide the bandwidth among all flows by some criteria.

IP ECN

Another approach is to use IP ECN. ECN is only used when the two hosts signal that they want to use it. With this method, an ECN bit is used to signal that there is explicit congestion. This is better than the indirect packet delete congestion notification performed by the RED/WRED algorithms, but it requires explicit support by both hosts to be effective. Some outdated or buggy network equipment drops packets with the ECN bit set, rather than ignoring the bit. More information on the status of ECN including the version required for Cisco IOS, by Sally Floyd, one of the authors of ECN.

When a router receives a packet marked as ECN capable and anticipates (using RED) congestion, it will set an ECN-flag notifying the sender of congestion. The sender then ought to decrease its transmission bandwidth; e.g. by decreasing the tcp window size (sending rate) or by other means.

Cisco AQM: Dynamic buffer limiting (DBL)

Cisco has taken a step further in their Catalyst 4000 series with engine IV and V. Engine IV and V has the possibility to classify all flows in "aggressive" (bad) and "adaptive" (good). It ensures that no flows fill the port queues for a long time. **DBL** can utilize **IP ECN** instead of packet-delete-signalling.

TCP Window Shaping

Congestion avoidance can also efficiently be achieved by reducing the amount of traffic flowing into a network. When an application requests a large file, graphic or web page, it usually advertises a "window" of between 32K and 64K. This results in the server

sending a full window of data (assuming the file is larger than the window). When there are many applications simultaneously requesting downloads, this data creates a congestion point at an upstream provider by flooding the queue much faster than it can be emptied. By using a device to reduce the window advertisement, the remote servers will send less data, thus reducing the congestion and allowing traffic to flow more freely. This technique can reduce congestion in a network by a factor of 40.

Side effects of congestive collapse avoidance

Radio links

The protocols that avoid congestive collapse are often based on the idea that data loss on the Internet is caused by congestion. This is true in nearly all cases; errors during transmission are rare on today's fiber based Internet. However, this causes WiFi, 3G or other networks with a radio layer to have poor throughput in some cases since wireless networks are susceptible to data loss due to interference. The TCP connections running over a radio based physical layer see the data loss and tend to believe that congestion is occurring when it isn't and erroneously reduce the data rate sent.

Short-lived connections

The slow-start protocol performs badly for short-lived connections. Older web browsers would create many consecutive short-lived connections to the web server, and would open and close the connection for each file requested. This kept most connections in the slow start mode, which resulted in poor response time.

To avoid this problem, modern browsers either open multiple connections simultaneously or reuse one connection for all files requested from a particular web server. However, the initial performance can be poor, and many connections never get out of the slow-start regime, significantly increasing latency.

Chapter-8

Routing in the PSTN and Self-similar Process

Routing in the PSTN

Routing in the PSTN is the process used to route telephone calls across the public switched telephone network. This process is the same whether the call is made between two phones in the same locality, or across two different continents.

Relationship between exchanges and operators

Telephone calls must be routed across a network of multiple exchanges, potentially owned by different telephone operators. The exchanges are all inter-connected together using trunks. Each exchange has many "neighbours", some of which are also owned by the same telephone operator, and some of which are owned by different operators. When neighbouring exchanges are owned by different operators, they are known as interconnect points.

This means that there is really only one virtual network in the world that enables any phone to call any other phone. This virtual network comprises many interconnected operators, each with their own exchange network. Every operator can then route calls directly to their own customers, or pass them on to another operator if the call is not for one of their customers.

The PSTN is not a fully meshed network with every operator connected to every other - that would be both impractical and inefficient. Therefore calls may be routed through intermediate operator networks before they reach their final destination. One of the major problems in PSTN routing is determining how to route this call in the most cost effective and timely manner.

Call routing

Each time a call is placed for routing, the destination number (also known as the called party) is entered by the calling party into their terminal. The destination number generally has two parts, a prefix which generally identifies the geographical location of the destination telephone, and a number unique within that prefix that determines the specific destination terminal. Sometimes if the call is between two terminals in the same local area (that is, both terminals are on the same telephone exchange), then the prefix may be omitted.

When a call is received by an exchange, there are two treatments that may be applied:

- Either the destination terminal is directly connected to that exchange, in which case the call is placed down that connection and the destination terminal rings.
- Or the call must be placed to one of the neighboring exchanges through a connecting trunk for onward routing.

Each exchange in the chain uses pre-computed routing tables to determine which connected exchange the onward call should be routed to. There may be several alternative routes to any given destination, and the exchange can select dynamically between these in the event of link failure or congestion.

The routing tables are generated centrally based on the known topology of the network, the numbering plan, and analysis of traffic data. These are then downloaded to each exchange in the telephone operators network. Because of the hierarchical nature of the numbering plan, and its geographical basis, most calls can be routed based only on their prefix using these routing tables.

Some calls however cannot be routed on the basis of prefix alone, for example non-geographical numbers, such as toll-free or freephone calling. In these cases the Intelligent Network is used to route the call instead of using the pre-computed routing tables.

In determining routing plans, special attention is paid for example to ensure that two routes do not mutually overflow to each other, otherwise congestion will cause a destination to be completely blocked.

According to Braess' paradox, the addition of a new, shorter, and lower cost route can lead to an increase overall congestion. The network planner must take this into account when designing routing paths.

One approach to routing involves the use of Dynamic Alternative Routing (DAR). DAR makes use of the distributed nature of a telecommunications network and its inherent randomness to dynamically determine optimal routing paths. This method generates a distributed, random, parallel computing platform that minimises congestion across the network, and is able to adapt to take changing traffic patterns and demands into account.

Self-similar process

Self-similar processes are types of stochastic processes that exhibit the phenomenon of self-similarity. A self-similar phenomenon behaves the same when viewed at different degrees of magnification, or different scales on a dimension (space or time). Self-similar processes can sometimes be described using heavy-tailed distributions, also known as long-tailed distributions. Example of such processes include traffic processes such as packet inter-arrival times and burst lengths. Self-similar processes can exhibit long-range dependency.

Overview

The design of robust and reliable networks and network services has become an increasingly challenging task in today's Internet world. To achieve this goal, understanding the characteristics of Internet traffic plays a more and more critical role. Empirical studies of measured traffic traces have led to the wide recognition of self-similarity in network traffic.

Self-similar Ethernet traffic exhibits dependencies over a long range of time scales. This is to be contrasted with telephone traffic which is Poisson in its arrival and departure process.

In traditional Poisson traffic, the short-term fluctuations would average out, and a graph covering a large amount of time would approach a constant value.

Heavy-tailed distributions have been observed in many natural phenomena including both physical and sociological phenomena. Mandelbrot established the use of heavy-tailed distributions to model real-world fractal phenomena, e.g. Stock markets, earthquakes, climate, and the weather. Ethernet, WWW, SS7, TCP, FTP, TELNET and VBR video (digitised video of the type that is transmitted over ATM networks) traffic is self-similar.

Self-similarity in packetised data networks can be caused by the distribution of file sizes, human interactions and/ or Ethernet dynamics. Self-similar and long-range dependent characteristics in computer networks present a fundamentally different set of problems to people doing analysis and/or design of networks, and many of the previous assumptions upon which systems have been built are no longer valid in the presence of self-similarity.

The Poisson distribution

Before the heavy-tailed distribution is introduced mathematically, the Poisson process with a memoryless waiting-time distribution, used to model (among many things) traditional telephony networks, is briefly reviewed below.

Assuming pure-chance arrivals and pure-chance terminations leads to the following:

- The number of call arrivals in a given time has a Poisson distribution, i.e.:

$$P(a) = \left(\frac{\mu^a}{a!}\right) e^{-\mu},$$

where *a* is the number of call arrivals in time *T*, and μ is the mean number of call arrivals in time *T*. For this reason, pure-chance traffic is also known as Poisson traffic.

- The number of call departures in a given time, also has a Poisson distribution, i.e.:

$$P(d) = \left(\frac{\lambda^d}{d!}\right) e^{-\lambda},$$

where *d* is the number of call departures in time *T* and λ is the mean number of call departures in time *T*.

- The intervals, *T*, between call arrivals and departures are intervals between independent, identically distributed random events. It can be shown that these intervals have a negative exponential distribution, i.e.:

$$P[T \geq t] = e^{-t/h},$$

where *h* is the mean holding time (MHT).

The heavy-tail distribution

A distribution is said to have a heavy tail if

$$\Pr[X > x] \sim x^{-\alpha}, \text{ as } x \to \infty, \qquad 0 < \alpha < 2.$$

This means that regardless of the distribution for small values of the random variable, if the asymptotic shape of the distribution is hyperbolic, it is heavy-tailed. The simplest heavy-tailed distribution is the Pareto distribution which is hyperbolic over its entire range.

Modelling self-similar traffic

Since (unlike traditional telephony traffic) packetised traffic exhibits self-similar or fractal characteristics, conventional traffic models do not apply to networks which carry self-similar traffic.

With the convergence of voice and data, the future multi-service network will be based on packetised traffic, and models which accurately reflect the nature of self-similar traffic will be required to develop, design and dimension future multi-service networks.

Previous analytic work done in Internet studies adopted assumptions such as exponentially-distributed packet inter-arrivals, and conclusions reached under such assumptions may be misleading or incorrect in the presence of heavy-tailed distributions.

Deriving mathematical models which accurately represent long-range dependent traffic is a fertile area of research.

Network performance

Network performance degrades gradually with increasing self-similarity. The more self-similar the traffic, the longer the queue size. The queue length distribution of self-similar traffic decays more slowly than with Poisson sources. However, long-range dependence implies nothing about its short-term correlations which affect performance in small buffers. Additionally, aggregating streams of self-similar traffic typically intensifies the self-similarity ("burstiness") rather than smoothing it, compounding the problem.

Self-similar traffic exhibits the persistence of clustering which has a negative impact on network performance.

- With Poisson traffic (found in conventional telephony networks), clustering occurs in the short term but smooths out over the long term.
- With self-similar traffic, the bursty behaviour may itself be bursty, which exacerbates the clustering phenomena, and degrades network performance.

Many aspects of network quality of service depend on coping with traffic peaks that might cause network failures, such as

- Cell/packet loss and queue overflow
- Violation of delay bounds e.g. in video
- Worst cases in statistical multiplexing

Poisson processes are well-behaved because they are stateless, and peak loading is not sustained, so queues do not fill. With long-range order, peaks last longer and have greater impact: the equilibrium shifts for a while.

Chapter-9

Quality of Service

In the field of computer networking and other packet-switched telecommunication networks, the traffic engineering term **quality of service** (*QoS*) refers to resource reservation control mechanisms rather than the achieved service quality. Quality of service is the ability to provide different priority to different applications, users, or data flows, or to guarantee a certain level of performance to a data flow. For example, a required bit rate, delay, jitter, packet dropping probability and/or bit error rate may be guaranteed. Quality of service guarantees are important if the network capacity is insufficient, especially for real-time streaming multimedia applications such as voice over IP, online games and IP-TV, since these often require fixed bit rate and are delay sensitive, and in networks where the capacity is a limited resource, for example in cellular data communication.

A network or protocol that supports QoS may agree on a traffic contract with the application software and reserve capacity in the network nodes, for example during a session establishment phase. During the session it may monitor the achieved level of performance, for example the data rate and delay, and dynamically control scheduling priorities in the network nodes. It may release the reserved capacity during a tear down phase.

A best-effort network or service does not support quality of service. An alternative to complex QoS control mechanisms is to provide high quality communication over a best-effort network by over-provisioning the capacity so that it is sufficient for the expected peak traffic load. The resulting absence of network congestion eliminates the need for QoS mechanisms.

In the field of telephony, quality of service was defined in the ITU standard X.902 as “A set of quality requirements on the collective behavior of one or more objects”. Quality of service comprises requirements on all the aspects of a connection, such as service response time, loss, signal-to-noise ratio, cross-talk, echo, interrupts, frequency response, loudness levels, and so on. A subset of telephony QoS is Grade of Service (GoS) requirements, which comprises aspects of a connection relating to capacity and coverage of a network, for example guaranteed maximum blocking probability and outage probability.

QoS is sometimes used as a quality measure, with many alternative definitions, rather than referring to the ability to reserve resources. Quality of service sometimes refers to the level of quality of service, i.e. the guaranteed service quality. High QoS is often confused with a high level of performance or achieved service quality, for example high bit rate, low latency and low bit error probability.

An alternative and disputable definition of QoS, used especially in application layer services such as telephony and streaming video, is requirements on a metric that reflects or predicts the subjectively experienced quality. In this context, QoS is the acceptable cumulative effect on subscriber satisfaction of all imperfections affecting the service. Other terms with similar meaning are the Quality of Experience (QoE) subjective business concept, the required "user perceived performance", the required "degree of satisfaction of the user" or the targeted "number of happy customers". Examples of measures and measurement methods are Mean Opinion Score (MOS), Perceptual Speech Quality Measure (PSQM) and Perceptual Evaluation of Video Quality (PEVQ).

History

Conventional Internet routers and LAN switches lack the ability to provide quality of service guarantees. This made Internet equipment less expensive, faster and thus more popular than competing more complex technologies that provided QoS mechanisms, for example X.25. Internet traditionally therefore runs at default QoS level, or "best effort". There were four "Type of service" bits and three "Precedence" bits provided in each IP packet header, but they were ignored. These bits were later re-defined as DiffServ Code Points (DSCP) and are sometimes honored in peered links on the modern Internet.

With the advent of IP TV and IP telephony, QoS mechanisms are increasingly available to the end user.

A number of attempts for layer 2 technologies that add QoS tags to the data have gained popularity during the years, but then lost attention. Examples are Frame relay and ATM. Recently, MPLS (a technique between layer 2 and 3) have gained some attention. However, today Ethernet may offer QoS and is, by far, the most popular layer 2 technology.

In Ethernet, Virtual LANs (VLAN) may be used to separate different QoS levels. For example in fibre-to-the-home switches typically offer several Ethernet ports connected to different VLAN:s. One VLAN may be used for Internet access (low priority), one for IP-TV (higher priority) and one for IP telephony (highest priority). Different Internet providers may use the different VLANs.

Key qualities of traffic

When looking at packet-switched networks, quality of service is affected by various factors, which can be divided into "human" and "technical" factors. Human factors include: stability of service, availability of service, delays, user information. Technical

factors include: reliability, scalability, effectiveness, maintainability, [[Grade of service|Grade of Service, etc.

Many things can happen to packets as they travel from origin to destination, resulting in the following problems as seen from the point of view of the sender and receiver:

Low throughput
: Due to varying load from other users sharing the same network resources, the bit rate (the maximum throughput) that can be provided to a certain data stream may be too low for realtime multimedia services if all data streams get the same scheduling priority.

Dropped packets
: The routers might fail to deliver (*drop*) some packets if their data is corrupted or they arrive when their buffers are already full. The receiving application may ask for this information to be retransmitted, possibly causing severe delays in the overall transmission.

Errors
: Sometimes packets are corrupted due to bit errors caused by noise and interference, especially in wireless communications and long copper wires. The receiver has to detect this and, just as if the packet was dropped, may ask for this information to be retransmitted.

Latency
: It might take a long time for each packet to reach its destination, because it gets held up in long queues, or takes a less direct route to avoid congestion. This is different from throughput, as the delay can build up over time, even if the throughput is almost normal. In some cases, excessive latency can render an application such as VoIP or online gaming unusable.

Jitter
: Packets from the source will reach the destination with different delays. A packet's delay varies with its position in the queues of the routers along the path between source and destination and this position can vary unpredictably. This variation in delay is known as jitter and can seriously affect the quality of streaming audio and/or video.

Out-of-order delivery
: When a collection of related packets is routed through a network, different packets may take different routes, each resulting in a different delay. The result is that the packets arrive in a different order than they were sent. This problem requires special additional protocols responsible for rearranging out-of-order packets to an isochronous state once they reach their destination. This is especially important for video and VoIP streams where quality is dramatically affected by both latency and lack of sequence.

Applications

A defined quality of service may be desired or required for certain types of network traffic, for example:

- Streaming media and specifically Internet protocol television (IPTV)
- IP telephony also known as Voice over IP (VoIP)
- Videoconferencing
- Circuit Emulation Service
- Safety-critical applications such as remote surgery where availability issues can be hazardous
- Network operations support systems either for the network itself, or for customers' business critical needs
- Online games where real-time lag (latency) can be a factor
- Industrial control systems protocols such as Ethernet/IP which are used for real-time control of machinery

These types of service are called *inelastic*, meaning that they require a certain minimum level of bandwidth and a certain maximum latency to function. By contrast, *elastic* applications can take advantage of however much or little bandwidth is available. Bulk file transfer applications that rely on TCP are generally elastic.

Obtaining QoS

- In advance: When the expense of mechanisms to provide QoS is justified, network customers and providers typically enter into a contractual agreement termed a service level agreement (SLA) which specifies guarantees for the ability of a network/protocol to give guaranteed performance/throughput/latency bounds based on mutually agreed measures, usually by prioritizing traffic.
- Reserving resources: Resources are reserved at each step on the network for the call as it is set up. An example is RSVP, Resource Reservation Protocol.

Over-provisioning

An alternative to complex QoS control mechanisms is to provide high quality communication by generously over-provisioning a network so that capacity is based on peak traffic load estimates. This approach is simple and economical for networks with predictable and light traffic loads. The performance is reasonable for many applications. This might include demanding applications that can compensate for variations in bandwidth and delay with large receive buffers, which is often possible for example in video streaming.

Commercial VoIP services are often competitive with traditional telephone service in terms of call quality even though QoS mechanisms are usually not in use on the user's connection to his ISP and the VoIP provider's connection to a different ISP. Under high load conditions, however, VoIP may degrade to cell-phone quality or worse. The mathematics of packet traffic indicate that network requires just 60% more raw capacity under conservative assumptions.

The amount of over-provisioning in interior links required to replace QoS depends on the number of users and their traffic demands. This is an important factor that limits usability

of over-provisioning. Newer more bandwidth intensive applications and the addition of more users results in the loss of over-provisioned networks. This then requires a physical update of the relevant network links which is an expensive process. Thus over-provisioning cannot be blindly assumed on the Internet.

QoS mechanisms

Early work used the "IntServ" philosophy of reserving network resources. In this model, applications used the Resource reservation protocol (RSVP) to request and reserve resources through a network. While IntServ mechanisms do work, it was realized that in a broadband network typical of a larger service provider, Core routers would be required to accept, maintain, and tear down thousands or possibly tens of thousands of reservations. It was believed that this approach would not scale with the growth of the Internet, and in any event was antithetical to the notion of designing networks so that Core routers do little more than simply switch packets at the highest possible rates.

The second and currently accepted approach is "DiffServ" or differentiated services. In the DiffServ model, packets are marked according to the type of service they need. In response to these markings, routers and switches use various queuing strategies to tailor performance to requirements. — At the IP layer, differentiated services code point (DSCP) markings use the 6 bits in the IP packet header. At the MAC layer, VLAN IEEE 802.1Q and IEEE 802.1p can be used to carry essentially the same information.

Routers supporting DiffServ use multiple queues for packets awaiting transmission from bandwidth constrained (e.g., wide area) interfaces. Router vendors provide different capabilities for configuring this behavior, to include the number of queues supported, the relative priorities of queues, and bandwidth reserved for each queue.

In practice, when a packet must be forwarded from an interface with queuing, packets requiring low jitter (e.g., VoIP or VTC) are given priority over packets in other queues. Typically, some bandwidth is allocated by default to network control packets (e.g., ICMP and routing protocols), while best effort traffic might simply be given whatever bandwidth is left over.

Additional bandwidth management mechanisms may be used to further engineer performance, to include:

- Traffic shaping (rate limiting):
  - Token bucket
  - Leaky bucket
  - TCP rate control—artificially adjusting TCP window size as well as controlling the rate of ACKs being returned to the sender
- Scheduling algorithms:
  - Weighted fair queuing (WFQ)
  - Class based weighted fair queuing
  - Weighted round robin (WRR)

    - Deficit weighted round robin (DWRR)
    - Hierarchical Fair Service Curve (HFSC)
- Congestion avoidance:
    - RED, WRED — Lessens the possibility of port queue buffer tail-drops and this lowers the likelihood of TCP global synchronization
    - Policing (marking/dropping the packet in excess of the committed traffic rate and burst size)
    - Explicit congestion notification
    - Buffer tuning

As mentioned, while DiffServ is used in many sophisticated enterprise networks, it has not been widely deployed in the Internet. Internet peering arrangements are already complex, and there appears to be no enthusiasm among providers for supporting QoS across peering connections, or agreement about what policies should be supported in order to do so.

One compelling example of the need for QoS on the Internet relates to this issue of congestion collapse. The Internet relies on congestion avoidance protocols, as built into TCP, to reduce traffic load under conditions that would otherwise lead to Internet Meltdown. QoS applications such as VoIP and IPTV, because they require largely constant bitrates and low latency cannot use TCP, and cannot otherwise reduce their traffic rate to help prevent meltdown either. QoS contracts limit traffic that can be offered to the Internet and thereby enforce traffic shaping that can prevent it from becoming overloaded, hence they're an indispensable part of the Internet's ability to handle a mix of real-time and non-real-time traffic without meltdown.

Asynchronous Transfer Mode (ATM) network protocol has an elaborate framework to plug in QoS mechanisms of choice. Shorter data units and built-in QoS were some of the unique selling points of ATM in the telecommunications applications such as video on demand, voice over IP.

Protocols that provide quality of service

- The Type of Service (ToS) field in the IP(v4) header (now superseded by DiffServ)
- IP Differentiated services (DiffServ)
- IP Integrated services (IntServ)
- Resource reSerVation Protocol (RSVP)
- Multiprotocol Label Switching (MPLS) provides eight QoS classes
- RSVP-TE
- Frame relay
- X.25
- Some ADSL modems
- Asynchronous Transfer Mode (ATM)
- IEEE 802.1p
- IEEE 802.1Q

- IEEE 802.11e
- HomePNA Home networking over coax and phone wires
- The ITU-T G.hn standard provides QoS by means of "Contention-Free Transmission Opportunities" (CFTXOPs) which are allocated to flows which require QoS and which have negotiated a "contract" with the network controller. G.hn also supports non-QoS operation by means of "Contention-based Time Slots".

QoS solutions

The research project "Multi Service Access Everywhere" (MUSE) defined a QoS concept in Phase I which was further worked out in another research project PLANETS. The new idea of this solution is to agree on a discrete jitter value per QoS class which is imposed on network nodes. Including best effort, four QoS classes were defined, two elastic and two inelastic. The solution has several benefits:

- End-to-end delay and packet loss rate can be predicted
- It is easy to implement with simple scheduler and queue length given in PLANETS
- Nodes can be easily verified for compliance
- End users do notice the difference in quality

The MUSE project finally elaborated its own QoS solution which is primarily based in:

- The usage of traffic classes
- Selective CAC concept
- Appropriate network dimensioning

Quality of service procedures

Unlike the Internet2 Abilene Network, the Internet is actually a series of exchange points interconnecting private networks and not a network in its own right. Hence the Internet's core is owned and managed by a number of different Network Service Providers, not a single entity. Its behavior is much more stochastic or unpredictable. Therefore, research continues on QoS procedures that are deployable in large, diverse networks.

There are two principal approaches to QoS in modern packet-switched networks, a parameterized system based on an exchange of application requirements with the network, and a prioritized system where each packet identifies a desired service level to the network. On the Internet, Integrated services ("IntServ") implements the parameterized approach. In this model, applications use the Resource Reservation Protocol (RSVP) to request and reserve resources through a network.

Differentiated services ("DiffServ") implements the prioritized model. DiffServ marks packets according to the type of service they desire. In response to these markings, routers and switches use various queueing strategies to tailor performance to

expectations. — At the IP layer, DiffServ Code Point (DSCP) markings use the first 6 bits in the ToS field of the IP(v4) packet header. At the MAC layer, VLAN IEEE 802.1q and IEEE 802.1p can be used to carry essentially the same information. — DiffServ internally assumes over-provisioning within its delay-sensitive Expedited Forwarding, class. This assumption is not always justifiable in the Internet, making it a contributing factor to the lack of DiffServ implementations in networks that transit the Internet.

Cisco IOS NetFlow and the Cisco Class Based QoS (CBQoS) Management Information Base (MIB) can both be leveraged within a Cisco network device to obtain visibility into QoS policies and their effectiveness on network traffic.

Non-IP protocols, especially those intended for voice transmission, such as ATM or GSM, have already implemented QoS in the core protocol and don't need additional procedures to achieve it.

End-to-end quality of service

End-to-end quality of service usually requires a method of coordinating resource allocation between one autonomous system and another. Research consortia such as EuQoS and fora such as IPsphere have developed mechanisms for handshaking QoS invocation from one domain to the next. IPsphere defined the Service Structuring Stratum (SSS) signaling bus in order to establish, invoke and (attempt to) assure network services. EuQoS conducted experiments to integrate Session Initiation Protocol, Next Steps in Signaling and IPsphere's SSS.

The Internet Engineering Task Force (IETF) defined the Resource Reservation Protocol (RSVP) for bandwidth reservation. RSVP is an end-to-end bandwidth reservation protocol that is also useful to end-to-end QoS. The traffic engineering version, RSVP-TE, is used in many networks today to establish traffic-engineered MPLS label-switched paths.

The IETF also defined NSIS with QoS signalling as a target. NSIS is a development and simplification of RSVP.

Quality of service circumvention

Strong cryptography network protocols such as Secure Sockets Layer, I2P, and virtual private networks obscure the data transferred using them. As all electronic commerce on the Internet requires the use of such strong cryptography protocols, unilaterally downgrading the performance of encrypted traffic creates an unacceptable hazard for customers. Yet, encrypted traffic is otherwise unable to undergo deep packet inspection for QoS.

Doubts about quality of service over IP

Gary Bachula, Vice President for External Affairs for Internet2, asserts that specific QoS protocols are unnecessary in the core network as long as the core network links are "over-provisioned" to the point that network traffic never encounters delay. In "quality of service" engineering, this formulation is guaranteed by the *admission control* feature. It is important to note that this only refers to core networks and not end-to-end connections. Recent studies point to a relatively low end-to-end bandwidth availability even on Internet2.

The Internet2 QoS Working Group concluded that increasing bandwidth is probably more practical than implementing QoS.

The Internet2 project found, in 2001, that the QoS protocols were probably not deployable inside its Abilene network with equipment available at that time. While newer routers are capable of following QoS protocols with no loss of performance, equipment available at the time relied on software to implement QoS. The Internet2 Abilene network group also predicted that "logistical, financial, and organizational barriers will block the way toward any bandwidth guarantees" by protocol modifications aimed at QoS. In essence, they believe that the economics would be likely to make the network providers deliberately erode the quality of best effort traffic as a way to push customers to higher priced QoS services.

The Abilene network study was the basis for the testimony of Gary Bachula to the Senate Commerce Committee's Hearing on Network Neutrality in early 2006. He expressed the opinion that adding more bandwidth was more effective than any of the various schemes for accomplishing QoS they examined.

Bachula's testimony has been cited by proponents of a law banning quality of service as proof that no legitimate purpose is served by such an offering. This argument is dependent on the assumption that over-provisioning isn't a form of QoS and that it is always possible. Cost and other factors affect the ability of carriers to build and maintain permanently over-provisioned networks.

Mobile cellular QoS

Mobile cellular service providers may offer **mobile QoS** to customers just as the fixed line PSTN services providers and Internet Service Provides (ISP) may offer QoS. QoS mechanisms are always provided for circuit switched services, and are essential for non-elastic services, for example streaming multimedia. It is also essential in networks dominated by such services, which is the case in today's mobile communication networks, but not necessarily tomorrow.

Mobility adds complication to the QoS mechanisms, for several reasons:

- A phone call or other session may be interrupted after a handover, if the new base station is overloaded. Unpredictable handovers make it impossible to give an absolute QoS guarantee during a session initiation phase.
- The pricing structure is often based on per-minute or per-megabyte fee rather than flat rate, and may be different for different content services.
- A crucial part of QoS in mobile communications is Grade of Service, involving outage probability (the probability that the mobile station is outside the service coverage area, or affected by co-channel interference, i.e. crosstalk) blocking probability (the probability that the required level of QoS can not be offered) and scheduling starvation. These performance measures are affected by mechanisms such as mobility management, radio resource management, admission control, fair scheduling, channel-dependent scheduling etc.

Chapter-10

Teletraffic Engineering in Broadband Networks and Traffic Generation Model

Teletraffic engineering in broadband networks

Teletraffic engineering is a well-understood discipline in the traditional telephone network, where traffic patterns are established, growth rates can be predicted, and vast amounts of detailed historical data are available for analysis. However, for modern broadband networks, the teletraffic engineering methodologies used for voice networks no longer suffice. Various aspects relating to **teletraffic engineering in broadband networks** are discussed here.

Firstly, the nature of broadband traffic is different from that of traditional voice networks. Many of the methodologies developed for traditional networks were based on the nature of voice calls, and are therefore not applicable to broadband networks. The nature of broadband traffic (broadband traffic characteristics) is discussed in the following sub-section.

The inherent nature of broadband networks is also different from that of traditional voice networks. Broadband networks have:

- high speeds,
- small cell sizes (in ATM networks), and
- limited information in the header.

These factors make teletraffic engineering in broadband networks more difficult than in traditional voice networks. A few more factors that further complicate teletraffic engineering in broadband networks are:

- A wide range of applications with diverse Quality of Service (QoS) requirements must be catered for.
- Much of the traffic (e.g., voice, video) is not amenable to flow control.
- The feedback within the network is “slow”.
- There are a large variety of traffic patterns.

Broadband traffic characteristics

In the traditional voice network, the study of traffic characteristics has matured over many years following the seminal work of Erlang in 1909. However, the teletraffic theory that has evolved relied heavily on the facts that

- Only one type of connection is offered, a circuit-switched connection, and
- the order of magnitude of the call holding time is relatively stable and well-known, namely the few minutes of a telephone conversation.

The diversity of broadband service connections and the variety of holding times make the application of teletraffic theory in voice networks to broadband networks very difficult . Figure 1 and Figure 2 show some applications and the variation in holding time and burstiness that may be expected for each one.

To manage the traffic implications of all these types of connections, we must return to the basic principles of traffic statistics. This has been extensively studied in recent years, and there is a large volume of published work on the subject. A Poisson process with one parameter does accurately model telephony traffic. However, to account for the changes with broadband traffic, alternatives to the Erlang formula have been considered . Two methods of modelling the traffic are considered, namely the Bernoulli-Poisson-Pascal approximation and the Maximum Entropy method. These methods use two parameters for describing the traffic – one for the mean demand and another to characterize the variability of the traffic.

Mechanisms used for teletraffic engineering in broadband networks

There are two primary mechanisms that are used for teletraffic engineering in broadband networks, namely:

- Traffic Control and Management, and
- Congestion Control

These two mechanisms are described in the following two sub-sections.

Traffic control and Management

Traffic control and management *is defined as the set of actions performed by the network to:*

- **avoid congestion**
- ensure that users get their required Quality of Service (QoS) guarantees.

The basic control problem is related to the efficient allocation of network resources so as to:

- satisfy different QoS requirements; and
- provide fair access to the network resources for all users

Traffic control and management provides the means by which:

- a user is ensured that the offered cell flows meet the rate specified in the traffic contract; and
- the network is ensured that the traffic contract rates are enforced such that the Quality of Service guarantees are provided for all users.

Congestion Control

Congestion control is defined as the set of actions performed by the network to **prevent or reduce congestion**. Congestion control is the most important part of the traffic management issue.

A network that controls congestion must:

- be responsive to the different utility functions of the users
- be able to manage the resources so that there is no loss of utility as the load increases

A simple example

The process of how a broadband network provides different levels of service to different types of traffic while avoiding congestion will be described briefly by means of a simple example. Links are provided for the reader interested in a more detailed description of the various concepts involved.

The concept of traffic control in broadband networks (particularly in ATM networks) is very simple: an application that requires the network to transport traffic from one location to another with a specific Quality of Service (QoS) follows the following procedure :

- The application declares the traffic's characteristics and the Quality of Service required by the traffic in a traffic contract (in order to make a connection request).
- The network judges whether it has enough resources available to accept the connection, and then either accepts or rejects the connection request. This is known as admission control.
- If there are insufficient resources available in the network at the time, the connection request is rejected.
- If there are sufficient resources available in the network at the time, the connection request is accepted and the network assigns the resources necessary to the connection.
- During communication, the network monitors the conformity between the declared traffic characteristics (in the traffic contract) and the characteristics of the actual traffic entering the network. This is known as traffic policing.

- The network has the capability to discard the non-conformant traffic in the network (using Priority Control).
- The overall rate at which all the traffic is admitted into the network is governed by the network's traffic shaping policy.

The ATM Traffic Management Framework

The ITU-T has defined a collection of ATM control mechanisms that operate across a spectrum of timing intervals . These control mechanisms are summarised in table 1.

Table 1: Summary of control mechanisms defined in ATM

Response Time	Traffic Control Functions	Congestion Control Functions
Long term	Network Resource Management	
Connection Duration	Connection Admission Control (CAC)	
Round-trip propagation time	Fast Resource Management	Explicit forward congestion indication (EFCI),ABR Flow Control
Cell insertion time	UPC and NPC, Priority Control, Traffic Shaping	Selective Cell Discard, Frame discard

Traffic generation model

A **traffic generation model** is a stochastic model of the traffic flows or data sources in a communication network, for example a cellular network or a computer network. A **packet generation model** is a traffic generation model of the packet flows or data sources in a packet-switched network. For example, a web traffic model is a model of the data that is sent or received by a user's web-browser. These models are useful during the development of telecommunication technologies, in view to analyse the performance and capacity of various protocols, algorithms and network topologies.

Application

The network performance can be analysed by network traffic measurement in a testbed network, using a **network traffic generator** such as iperf, bwping and Mausezahn. The traffic generator sends dummy packets, often with a unique packet identifier, making it possible to keep track of the packet delivery in the network.

Numerical analysis using network simulation is often a less expensive approach.

An analytical approach using queueing theory may be possible for simplified traffic model, but is often too complicated if a realistic traffic model is used.

The greedy source model

A simplified packet data model is the greedy source model. It may be useful in analyzing the maximum throughput for best-effort traffic (without any quality-of-service guarantees). Many traffic generators are greedy sources.

Poisson traffic model

Another simplified traditional traffic generation model for circuit-switched data as well as packet data, is the Poisson process, where the number of incoming packets or calls per time unit follows the Poisson distribution. The length of each phone call is typically modelled as en exponential distribution. The number of simultaneously ongoing phone calls follows the Erlang distribution.

Long-tail traffic models

However, the Poisson traffic model is memoryless, which means that is does not reflect the bursty nature of packet data, also known as the long-range dependency. For a more realistic model, a self-similar process such as the Pareto distribution can be used as a long-tail traffic model.

Payload data model

The actual content of the payload data is typically not modelled, but replaced by dummy packets. However, if the payload data is to be analyzed on the receiver side, for example regarding bit-error rate, a Bernoulli process is often assumed, i.e. a random sequence of independent binary numbers. In this case a channel model reflects channel impairments such as noise, interference and distortion.

Standardized Internet traffic models

There are at least two standardized traffic generation models for packet-switched wireless networks: the 3GPP2 model and the 802.16 model. The 3GPP2 model is much more complex to implement but it is supposed to give more precise results. The 802.16 model is much simpler in realization.

3GPP2 model

The 3GPP2 model is described in . This document describes the following types of traffic generators:

- Downlink:

    - HTTP/TCP
    - FTP/TCP
    - WAP
    - near real-time Video
    - Voice
- Uplink:
    - HTTP/TCP
    - FTP/TCP
    - WAP
    - Voice
    - Mobile Network Gaming

The main idea is to partly implement HTTP, FTP and TCP protocols. For example, an HTTP traffic generator simulates the download of a web-page, consisting of a number of small objects (like images). A TCP stream (that's why TCP generator is a must in this model) is used to download these objects according to HTTP1.0 or HTTP1.1 specifications. These models take into account the details of these protocols' work. The Voice, WAP and Mobile Network Gaming are modelled in a less complicated way.

802.16 model

The 802.16 model is much simpler. It was proposed in several 802.16 TG3 contributions. The idea is to define three basic models:

- Interrupted Poisson Process (IPP)
- Interrupted Discreet Process (IDP)
- Interrupted Renewal Process (IRP)

and mix them together in order to simulate different kinds of web-traffic. Every interrupted process may be either in ON or OFF state. The packets are generated only in ON state. The lengths of ON and OFF periods, sizes of the packets and intervals between them are defined separately in each model, so these models differ in the way their parameters are defined. These models may be mixed together, for example: 4IPP means a mix of four IPP flows with different parameters. HTTP and FTP is simulated as 4IPP; VoIP is simulated as IDP, 2IDP, 4IDP; Video is simulated as 2IRP.