ABSTRACT
In this paper, we propose a novel method of charging in a packet switched network. The idea is to redefine the network capacity as seen by individual packets. Then, based on this redefined capacity, each packet is to be charged a different amount - depending upon the tariffs. We have considered two simple network models to show that the variation of traffic load does indeed alter the way a packet is treated and the same may be reflected in the billing procedure. As an example, we have proposed a workable charging mechanism, contending the fact that there are many ways that the main idea of this paper may be used. The method is potentially useful in ATM wide area networks (WAN) offering ABR/UBR and VBR services.

1 INTRODUCTION
The concept of equivalent capacity is not new. It has been successfully applied in the modeling of the available capacity of a network [6] for admitting or rejecting a new call (Call Admission Control). The issue is of paramount importance to evaluate the multiplexing gain when variable bit rate (VBR) sources use ATM network multiplexer. Owing to the variations in the VBR traffic, a major fraction of the bandwidth can go unused if the multiplexing gain is not evaluated and utilized. There have been two service categories devised to utilize the "fluctuating remaining capacity" of an ATM network. These are the available bit rate (ABR) and the unspecified bit rate (UBR) services. In this paper, we will address the issue of fair charging for these two services.

ABR is to target the traffic sources with reliability constraints and no significant delay or jitter requirements. The user of ABR provides three parameters which can be used to determine the acceptance of the call. These are, the minimum rate \( R_{min} \), maximum rate \( R_{max} \) and, depending upon the network provider, the cell loss ratio (CLR) [1]. These parameters may be unique for each traffic source and may be decided by the designer of the equipment at the time of its design. These parameters, together with the network congestion state, help an ATM/ABR switch determine the new bit rate every time a new decision is needed. The method of determining the new bit rate is key to the design of an ABR switch and has not been standardized yet. Since it may not have any effect on integration of multivendor networks, it may never be standardized.

UBR is simpler in concept and places less demand on the network. A source using UBR service may be shut out of the service without any warning or may voluntarily reduce its speed in case of congestion. In other word, it is for a traffic which can wait, and perhaps, be lost as well. UBR is closer to the classical ATM idea, of having dumb switches which have no "say" in the network congestion control. It is much easier to implement as compared to rate adaptive ABR. A discussion of which one is better in what sense continues till today and there are no hard and fast rules of deciding.

While implementing ABR and UBR, the network provider has to make many decisions independently. Charging is one such crucial decision, not only for revenue generation but for also for fairness of charging and proper account of resource utilization of the network.

1.1 ALTERNATIVES FOR CHARGING MECHANISMS
A rather primitive method is to charge in proportion to the amount of time that it takes to complete the call, just as in PSTN. Let's call it holding-time-based charging (HBC). A second mechanism is the one used in some PSDN's, that is, charging based on a packet (cell) count (CBC, count based charging). One logical way is to base charging on the traffic parameters of each call. Thus, a cost of a call is given by \( f(R_{min}, R_{max}, CLR) \) and will always be the same as long as \( R_{min}, R_{max} \) and CLR remain the same, regardless of how long it takes to complete the call. Since delay is not one of the strict requirements in this service category, it may sound fair. Let's call it parameter-based charging (PBC). Another, modified, way is to use some equivalent capacity formula for the incoming call based on its \( R_{min}, R_{max}, CLR \) and the actual rate. Let's call it equivalent capacity-based charging (ECBC).

1.1.1 SIMPLE COMPARISON OF DIFFERENT ALTERNATIVES
It is possible to compare the above four alternatives based on simple logic. The holding time based charging (HBC) is good only for dedicated capacity allocation, such as
CBR. In ABR/UBR, it will result in punishing the user for a bad network condition. This is contrary to the logic of fairness. PBC does not consider the time it takes to complete a call. Delay may not be a criterion in deciding the quality of an ABR/UBR call, but two calls delayed by widely different times cannot be regarded equal. Hence, PBC can be rejected solely on this basis. ECBC entails the complication arising from mixing traffic from many sources and then finding out the effect of all other calls on the one under-consideration, and that too in real time. It will add yet another headache to the already bogged down designer of the ABR switch. However, this is the only way to have a justifiable billing criterion. So, research is needed in this area.

In this paper, we present a concept of equivalent capacity, which is simple, logical and not very complex to implement. In fact, the implementation complexity depends on network provider, and will consist of a simple logic circuit. As will be clear later, a better name for such charging mechanism will be transmission-time-based charging (TBC), as it will take into account the actual end to end transmission time.

2. TRANSMISSION-TIME-BASED CHARGING (TBC)

In conventional circuit-switched networks, such as PSTN, when a call is setup, a constant bandwidth is allocated to the call which is at the disposal of user throughout the holding time. It has been extensively quoted that a voice user really utilizes the system resources only about 40% of the time [2]. However, if the line is used for 100% of the time, the user doesn't have to pay more - as maybe the case in using a data modem. The customer is billed for the time spent 'on-line'.

To understand the concept of TBC, consider a traffic source which emanates a message of L bits. If the network allocates a capacity of C bps to this message, the transmission time is $L/C$ seconds. Such a source can be charged for C bps. However, if the message has to wait in a transmission queue, it will take longer for the same message. The total transmission completion time will be the sum of the $L/C$ and the waiting time in the queue W. Let this time be called the equivalent transmission time $T_e$. Thus, we can write:

$$T_e = \frac{L}{C} + W \quad \ldots 1$$

Now let $C_e$ be the capacity required to transmit the same message in $T_e$ time without waiting. Then, we can write:

$$T_e = \frac{L}{C_e} \quad \ldots 2$$

From equations 1 and 2, we easily see that:

$$C_e = \frac{LC}{L + CW} \quad \ldots 3$$

It is obvious from equation 3 that $C_e = C$, when there is no waiting delay, and $C_e = 0$, when there is infinite delay. For $0 < W < \infty$, we have, $C > C_e > 0$.

If the message is provided with a dedicated capacity of $C_e$ bps, it will always take a fixed $T_e$, given by equation 2. However actual $C_e$ may be a stochastically varying quantity, depending upon W, so, it will have a distribution and moments.

In TBC, the billing is done based on $C_e$. Since ABR requests a delivery within a range of offered bit rate, it can be billed based on $C_e$. Other services, which are affected by a variance in delay cannot use $C_e$ directly as a charging parameter. In some cases, a guarantee must be made about the maximum delay. This guarantee can be met with some probability $\pi$, which can help refine the formula for $C_e$ for such services. In the following sections, we will demonstrate the application of $C_e$ for charging.

2.1 SYSTEM MODELS AND ASSUMPTIONS

We will apply our method to two types of systems, the M/M/1 queuing system and the M/D/1 system. Message arrival in both cases is governed by Exponential interarrival times, while the messages (or better yet, the packets) are distributed exponentially in the former case and have a constant length in the later case. Even though the ATM network has constant length cells, there are two reasons for employing the M/M/1 model, that is, 1. If the charging is applied at an upper layer, such as AAL, the model may be useful.

2. M/M/1 is the simplest queueing system and helps understand matters better as compared to other models. Also, it is a loose upper bound for some other models.

At this point, it may be remarked that the traffic modeling of ON/OFF type will not appear as such to a common buffer, as a large number of sources are emanating cells asynchronously. Thus, to a common buffer, the cell stream may still appear to be continuous. We also assume that sufficient buffer space is available at the UNI. The unit of time will be normalized to the average packet length.

2.2 THE EQUIVALENT CAPACITY EVALUATION

As clear from the above discussion, the actual $C_e$ is not only a random quantity, but also depends upon many factors, such as, the source traffic characteristics, message service time characteristics, service mechanism, and priority classes. In this section, we will first define a quantity $C_\pi$ called the $\pi$-guaranteed $C_e$. Here's the definition:
Definition 1: Let $C_x$ be a continuous stochastic quantity with a probability distribution $F_{C_x}(x)$. Then $C_x$ is such that $P[x > C_x] = \pi$, or in other words, $F_{C_x}(C_x) = 1 - \pi$.

A simple interpretation of $C_x$ is that $x$ is guaranteed to have a minimum value equal to $C_x$ with probability $\pi$.

A better parameter for communications services is the total transmission time. Accordingly, if a packet is $L$ bits long, a charge for $C_x$ ensures a total system response time of $T_x \leq L/C_x$ seconds. Obviously, there is an error of $1-\pi$ in the assurance. It should be noted here that the actual value of the available capacity (and hence the response time) is still stochastic and therefore, will result in a variance. As a result, this method is applicable only to those (data) services which are generally non-sensitive to jitter.

Corollary 1: A direct result of the definition of $C_x$ is the that the jitter will be a function of $\pi$. Hence, for infinite buffer situations (such as UBR services require), jitter will be exponentially increasing with $\pi$, while for limited buffer situations, it will be a convex function of $\pi$.

Corollary 2: Let $T_x$ be the equivalent transmission time corresponding to $C_x$ with a cumulative distribution function $F_{T_x}(x)$. Then $\pi = 1 - F_{C_x}(C_x) = F_{T_x}(T_x) = P[T_x \leq T_x]$ where $T_x = L/C_x$.

2.3 APPLICATION OF TBC

As the first example of application, we will consider the most basic queueing model, that of Poisson arrivals, Exponential service time, and a single server (M/M/1) with infinite waiting room available. Let $f_W(x)$ and $f_T(x)$ be the probability density functions of the waiting times, and time in systems respectively. It is well-known that \[ f_T(t) = (1 - \lambda) e^{-(1-\lambda)t} \]

where $L = 1 = C$.

So,

\[ P[T \leq T_x] = \int_0^{T_x} f_T(t) \, dt = \pi \]

\[ = 1 - e^{-(1-\lambda)T_x} \]

Which gives

\[ 1/T_x = C/C' = \frac{1 - \rho}{\ln(1 - \pi)} \]

Let's define $\eta = C/C'$ as the capacity goodput efficiency, then

\[ \eta = \frac{1 - \rho}{\ln(1 - \pi)} \quad \text{...4} \]

Two plots in Figure 1(a) and (b) show the effect of varying $\rho$ and $\pi$ on $\eta$. For a given value of $\pi$, $\eta$ varies linearly with $\rho$. This result is rather interesting. The equivalent delay would vary non-linearly with $\rho$ as $(1 - \rho)$ factor would be in the denominator.

The effect of varying $\pi$ is linear for most of the load as seen from Figure 1(b). From these graphs it is easily seen that charging for the same amount under different traffic loads is not fair, as the actual capacity offered to the users is much less than the line capacity.

A more realistic queueing model for contemporary network nodes is an M/D/1, with Poisson arrivals,
constant packet length and a single server. Such model is more suited to ABR and UBR with sufficient amount of buffer. In such a case, the service time is always the same, say equal to unity, and the waiting time has two parts, a discrete part owing to packets queued ahead and a continuous part owing to the continuous arrival process. The distribution function of the waiting time is given by [4]:

$$F_w(k + x) = (1 - \lambda) \sum_{j=0}^{k} \frac{[-\lambda(j + x)]^{k-j}}{(k-j)!} e^{\lambda(j+x)}; \ldots 5$$

$k=0,1, \ldots$ and $0 \leq x < 1$

Let $k_e$ and $x_e$ be such that $T_e = 1 + k_e + x_e$, then

$$\pi = (1 - \lambda) \sum_{j=0}^{k_e} \frac{[-\lambda(j + x_e)]^{k-e-j}}{(k_e-j)!} e^{\lambda(j+x_e)} \ldots 6$$

Since the time is normalized over $L/C$, $\lambda$ is the load in the above equation. Two plots of equation 6 are shown in Figures 2(a) and 2(b) below. Figure 2(a) shows $\eta$ as a function of $\rho(= \lambda)$ for $\pi = 0.8$, 0.85, 0.9, and 0.95 while Figure 2(b) depicts the effect of $\pi$ on $\eta$ for a set of values of 0.2, 0.4, 0.6 and 0.8 of the load. From these graphs we note again that there is wide difference of system response time for different load values.

One prominent difference between the curves of Figure 1 and those of Figure 2 is that in Figure 2, the equivalent capacity is general much higher than in Figure 1. Thus, at a moderate load of $\rho = 0.2$, and $\pi = 0.95$, $\eta$ for $M/M/1$ queue is about 20%, while for $M/D/1$ system, its value is around 30%. This 10% more capacity is due to the fact that in $M/D/1$ queue there is no variance of the transmission time, thus reducing the randomness in the total transmission plus waiting time as compared to the $M/M/1$ case.

2.4 LOOSE QOS CONTROL

The above idea can be applied by using the average service and waiting times and applying the Pollackz-Kinhchin formula for the mean values. However, a little reflection will reveal that the above formulas are not only more accurate but also contain the averages as a special case.

3. EXAMPLE BILLING ALGORITHM

The use of TBC for actual billing is subjective, in the sense that each network provider can use a different formula to evaluate $\eta$, and give different weight to each parameter, depending upon the traffic models used. Therefore, there is no unique way of utilizing these equations. Here is one such way:

Step 1: At the subscription time, the user is told that the billing will be done for a known value of $\pi$. Such as $\pi = 0.95$, which means that the curve ($\eta$ vs. $\rho$) with $\pi = 0.95$ will be used.

Step 2: To avoid continuous measurement of the traffic load, intermittent measurements may be performed. A measurement performed at time $t_0$ may be valid till $t_0 + \tau$, where $\tau$ is the measuring cycle time.

Step 3: A packet counter will count the packets for each call during a measurement time. At the end of the measurement cycle, the number of counted packets are multiplied by the charge per packet for the measured load.
Step 4: Let \( n_i \) be the number of packets during the \( i \)th measurement and let \( d_i \) be the cost per packet during this time, then if \( N \) is the total number of measurement cycles for which the call was in progress, the total charge \( D \) is given by:

\[
D = \sum_{i=1}^{N} n_i d_i
\]

which is a simple formula and does not require extraordinary measurements. In fact, most of the information is already available in the network management and congestion control system. There is reasonably small overhead for a logically fair charging system.

4. PRACTICAL CONSIDERATIONS

The system models presented in this paper are not applicable directly to real networks. These are simple models easily understandable and used to clarify the idea. Actual networks are far more complex and the traffic models are not an approximation of Poisson traffic either [5]. Not only that, each connection may consist of many links within a single network, and many networks too. Traffic on each link may be different, resulting in different curves applicable to different links for a single connection. An example connection is shown in Figure 3 below.

In a network like this, the measurements may be based on some feedback received from the management system. In ATM service category ABR, resource management (RM) cells are used to inform the traffic source of a new increased or decreased bit rate. The source rate is varies inversely with the load. By performing relevant studies on such networks, a price factor can be added to each RM cell driven from the congestion level. This price can be applied until the next RM cell. In this way, an acyclic mechanism can be realized. The diversity of algorithms that can be applied to TBC stems from the fact that the only promise the network provider is making is on the worst case scenario, i.e., \( C_x \) as minimum capacity. This is just like the ABR service category in concept.

4.1 TRAFFIC CONTROL BASED ON TBC

While on the one hand transmission-time-based charging is a fair billing mechanism, it can also be used to control the traffic entering a network. By merely setting up a minimum available capacity \( C_x \), the network designer, in fact, is limiting the maximum admissible load to some \( \rho_x \). The call admission control system, then, has to evaluate the equivalent capacity\(^1\) of the existing calls. Based on the two equivalent capacities, the \( C_x \) and the existing calls, a decision can be made on the acceptance or otherwise of a call.

We leave for the future to show that such a utilization of \( C_x \) is possible.

5. CONCLUSION

In this paper, we have taken the case of charging in a packet switched network when the available capacity is fluctuating randomly. In such situations, it is not fair to bill the user for a fixed capacity, as is done in PSTN, or on flat rate packet count, as is done in some PSDNs. The most scientific and fair way is to charge for the capacity seen by each packet. We have shown that it is possible with certain probability. We have demonstrated the applicability of this idea for systems modeled as M/M/1 and M/D/1 queues. More advanced models are possible in a similar way. In the end we have shown an example algorithm for fair billing.

REFERENCES

1. FORUM, Traffic Management 4.0", , ATM Forum 1996

---

\(^1\) The meaning of equivalent capacity here is different. It is the capacity used by the existing calls which is calculated by using any of the proposed methods, such as a fluid flow model [6]. A better term for this will be equivalent accumulated bit rate.