A New Approach to Service Provisioning in ATM Networks

Steven H. Low, Member, IEEE, and Pravin P. Varaiya, Fellow, IEEE

Abstract—We formulate and solve a problem of allocating resources among competing services differentiated by user traffic characteristics and maximum end-to-end delay. The solution leads to an alternative approach to service provisioning in an ATM network, in which the network offers directly for rent its bandwidth and buffers and users purchase freely resources to meet their desired quality. Users make their decisions based on their own traffic parameters and delay requirements and the network sets prices for those resources. The procedure is iterative in that the network periodically adjusts prices based on monitored user demand, and is decentralized in that only local information is needed for individual users to determine resource requests. We derive network's adjustment scheme and users' decision rule and establish their optimality. Since our approach does not require the network to know user traffic and delay parameters, it does not require traffic policing on the part of the network.

I. INTRODUCTION

In this paper, we study a different approach to provisioning services in an ATM network. A service is specified by a one-way connection (source, destination, route) and two sets of service parameters. A connection is used to transport a data stream or message from the source to the destination; the traffic parameters specify constraint on a user's traffic "burstiness"; the quality parameters specify maximum end-to-end delay and cell loss rate. We assume that the network offers for sale different types of services, differentiated by the triplet (connection, traffic parameters, quality parameters), and that there is a flow, depending on the service cost, of user requests for these services. A service request is admitted if sufficient resources can be allocated along the connection's route to guarantee service quality. By resources, we mean the bandwidth and buffers in each node along the route. Before transmission, a message is segmented into small, fixed size units called cells. The bandwidth and buffers allocated to a connection can vary over links in its route.

More specifically, a network offers a set \( S \) of services. A unit of type \( s \) service is provided by a type \( s \) connection with the associated traffic and quality parameters, and is sold for a unit price of \( w_s \). The network can produce any amount of type \( s \) service provided the required resources do not exceed capacity. Users request services to maximize their individual utilities subject to budget constraints, and the result of this maximization is summarized by an aggregate demand function \( D_s(w_s) \). Further assume that the network is regulated to charge its services so as to maximize a welfare function:

\[
\max_{w,x} W'(w, x) \tag{1}
\]

subject to \( f(x) \leq 0 \)

\[
x_s \leq D_s(w_s), \quad s \in S \tag{2}
\]

where \( W'(w, x) \) is the sum of network revenue and user surplus, and (2) is capacity constraint. Standard price-adjustment schemes [1], pp. 188–189] can be used to reach an equilibrium, in which the network sets a price \( w \) and supply \( x \), users present their demands \( D_s(w_s) \), and the network increases or decreases price according as demand is greater or less than supply, until \( (w, x) \) converges.

A key observation is that bandwidth and buffers are "substitutable resources" to meet a service quality (see below). In this paper, we exploit this tradeoff of different resource combinations to optimize an overall measure of network performance. For each allocation indexed by \( \mu \), let \( W(\mu) \) denote the welfare achieved in (1) by an equilibrium price—supply vector \((w(\mu), x(\mu))\). Our objective is to derive an iterative and decentralized algorithm that solves for a \( \mu^* \) that meets service quality and maximizes \( W(\mu) \). The algorithm leads to our service provisioning procedure in which the network offers directly for rent its bandwidth and buffers, and the users purchase freely resources to meet its desired quality. A user bases its decision on the knowledge of its own traffic and quality parameters, and on the resource price. The network periodically adjusts the prices based on the monitored user request for resources. Unlike the common price adjustment scheme based on the law of demand and supply, our scheme involves a minimization by the network. Furthermore, users of type \( s \) service effectively do two optimization in each period: one selects a resource combination along the route to minimize service cost \( w_s \); the other selects a demand \( D_s(w_s) \) that maximizes surplus. It is decentralized in that each user only needs to know the resource price at nodes along its route in addition to its own traffic and quality parameters. The solution makes critical use of the bandwidth-buffer tradeoff.
described by burstiness curve in [16], [17]: for the network, it
determines the resource combination to maximize welfare;
for individual users, it provides a simple rule for requesting
resources (see (19)).

Our approach differs in three ways from the conventional
service provisioning approach in which the network decides
beforehand resources to be allocated to the users. First, under
our approach, users freely rent resources and package them
into services that best meet their needs. Second, since service
price is the rent a user pays for the resources it reserves, our
procedure ties this price to network performance measured by
the welfare function. Third, since the network only guarantees
the availability of purchased resources, it is the users' responsi-
bility to shape their traffic in order that the allocated resources
can provide the desired quality. Our approach relieves
the network of the difficult task of traffic policing and enforcement
and can potentially adapt to time-varying user needs expressed
by the traffic and quality parameters.

Several sets of previous work are relevant here. In con-
ventional packet-switched networks, bandwidth and buffers
are typically not reserved for connections but shared on
an on-demand basis. Unrestricted sharing makes it difficult
for the network to offer a guaranteed delay to a particular
connection, since a more bursty connection can monopolize
network resources to the detriment of other connections,
possibly leading to congestion. As a consequence of the need
to guarantee service quality in an ATM network, various
bandwidth allocation schemes have recently been designed
to guarantee certain amount of bandwidth to a connection
despite changes in the number and burstiness of concurrent
connections, e.g., [10], [25], [5], [20], [2], [7], [22]. This
justifies our approximation of each network node as allocating
a fixed bandwidth to a connection. The next question is how
much bandwidth to allocate to each connection. To secure
service quality, a common proposal is to allocate to each
connection enough bandwidth to accommodate its peak rate.
This simple proposal leads to inefficiency if the peak rate is
much larger than its average rate. This may be unnecessary
if the tradeoff between bandwidth and buffers as substitutable
resources for service provisioning can be exploited. A more
sophisticated proposal is based on the notion of 'equivalent
bandwidth' [13], [6], [14], [3]. References [11],
[12] study the effect on some overall measure of network
performance of varying resource capacities in the network
and of varying routing. We vary the resource allocation to the
connections rather than the resource capacity, exploiting the
bandwidth-buffer tradeoff. Multiple-service multiple-resource
model is also considered in [8]. There the resource allocation
is fixed, and the problem is to find a revenue-maximizing
admission control policy that may reject a request even when
sufficient resources are available in the hope of more profitable
requests in the future. In the present formulation resources
cannot be reserved in anticipation of future requests. Instead
we explore alternative feasible allocations. Pricing has been
used previously for control and optimization in communication
networks, e.g., [4], [15].

The tradeoff between bandwidth and buffers is easily illus-
trated. Consider the transfer over a single link of a sequence of
video frames; see Fig. 1. Suppose a frame contains 512×512
8-bit pixels, and is generated at the video source every 33 ms.
The service is to deliver a frame every 33 ms to the display.
Suppose the source rate is m(t) as shown in the figure. If
we allocate the peak rate 150 Mbps then the received rate
is the same as m(t) (except for propagation and processing
delay). An alternative is to allocate 65 Mbps and some buffers
to the connection; the received rate m'(t) is as shown. The
alternative allocation achieves the same service as shown in
the figure.

In Section II, we introduce our network model and service
parameters. In Section III, we define optimality and formulate
optimal allocation as a game problem. Its solution leads to our
iterative and decentralized service provisioning procedure that
does not require the network to know user traffic and quality
parameters. Concluding remarks are collected in Section IV,
and proofs are relegated to the Appendix.

While our argument is preliminary, it does suggest an
alternative to the popular approach in which the network
decides which services will best meet user needs. Here, the
users decide the resources they need based on their own traffic
and quality parameters, and the network coordinates their
choices via resource pricing in order to optimize an overall
measure of network performance.

II. MODEL

We consider a network with a set L of links. Link l ∈ L
comprises a transmission capacity of C l cells per second, or
cps, and buffers for B l cells, see Fig. 2. A set R of routes
is specified; r ∈ R also denotes the set of links along that
route. The network offers a set S of services. A unit of
type s service is sold at a price of w s and is provided
by a connection over route r s for one unit of time. Under
our service provisioning scheme, this unit price w s is related
to the cost of resources needed to provide the service, and
is adjusted periodically to achieve an optimal allocation, as
elaborated in Section III. We shall assume that user demand,
or requests, for type s service is given by the aggregate demand
function D s (w s ) = v s exp(−w s ), w s := λ s T s , λ s exp(−w s ) is
the (average) rate type s requests arrive and T s is the (average)
duration of a type s connection. A type s request is admitted
"
but the allocation positive real number that bounds the average message rate. It is compliant if

can be freely chosen provided the service quality constraint is met as explained next.

Once a type \( s \) request is admitted, a connection is set up, and the user sends a message. A message is a “fluid flow,” \( m(t), 0 \leq t \leq T \), where \( m(t) \) is the instantaneous rate in cps, and \( T \) is the duration. A type \( s \) message must satisfy two constraints denoted by \( (b_s(p), \mu_s) \). The parameter \( \mu_s \) is a positive real number that bounds the average message rate. It is also the minimum bandwidth required for a type \( s \) message at each node along its route. The parameter \( b_s(p), \mu \geq 0 \), is a nonnegative, decreasing, convex function that bounds the message “burstiness.” A type \( s \) message \( m \) is said to be compliant if

\[
\eta := \frac{1}{T} \int_0^T m(\tau) d\tau \leq \mu_s
\]

and

\[
\max_{0 \leq s \leq T} \int_0^T [m(\tau) - \mu] d\tau \leq b_s(\mu), \quad \mu \geq \eta
\]

Inequality (3) says that the average message rate \( \eta \) cannot exceed \( \mu_s \). The left-hand side of (4) is the maximum backlog if \( m \) is transmitted over a link at a constant speed \( \mu \geq \eta \). Hence, inequality (4) says that if \( m \) is allocated a bandwidth of \( \mu \), then a buffer of size \( b_s(\mu) \) is sufficient to prevent cell loss. Note that the larger \( \mu \) the smaller is \( b_s(\mu) \). Thus the function \( b_s \), called burstiness curve, gives the bandwidth-buffer tradeoff for zero cell loss. To incorporate cell loss, we may relax (4) and let \( b_s(\mu) \) be the buffer required to have no more than certain number of lost cells if \( m \) is transmitted over a link at a constant speed \( \mu \) [23].

A word is in order on how this traffic characterization may be used in practice. First, if the duration of a message is very long, e.g., a video program, we divide its duration into disjoint periods \( T_i, i = 1, \ldots, k \). A type \( s \) message is compliant if the portion of message on every period \( T_i \) satisfies (3) and (4) with \( T \) replaced by \( T_i \). Inequality (3) then guarantees that no cell backlog carries over to the next period. Second, we do not assume that a user of type \( s \) service knows its own message \( m(t) \) or the burstiness of \( m(t) \). We only assume that they know a bound \( b_s(\mu), \mu \geq \eta \), on the burstiness. In fact, condition (4) can be easily enforced by passing an arbitrary user message through a leaky bucket policing device before being admitted into the network. The two parameters of a leaky bucket (and the bound on peak message rate in each period) define a piecewise linear burstiness curve that bounds the burstiness of the output message from the leaky bucket [17, Proposition 3].

For the rest of this paper, we will use the following vector notation. \( \mu_s \) denotes the vector \( \{\mu_s, l \in r_s\} \) of bandwidth allocation for a type \( s \) connection, and \( \mu \) denotes the vector \( \{\mu_s, s \in S\} \). Similarly, \( b_s(\mu_s) \) denotes the vector \( \{b_s(\mu_s), l \in r_s\} \) of buffers required for a type \( s \) connection, and \( b(\mu) \) denotes the vector \( \{b_s(\mu_s), s \in S\} \). \( \mu_s \) denotes the vector \( \{\mu_s, s \in S\} \). We may abuse notation and use “\( \mu \geq \mu \)” to mean \( \eta[\mu_s \geq \mu_s; l \in r_s, s \in S] \). Finally, \( x, y > \) denotes the inner product of vectors \( x \) and \( y \).

To specify the quality of service, the maximum end-to-end delay, we use two results from the theory of burstiness curve in [16]–[18] (see also [1], [20], [24]). Suppose a type \( s \) compliant message is transmitted over a connection with route \( r_s \). Suppose that a bandwidth of \( \mu_s \) cps and buffer of \( b_s \) cells are allocated to that connection at each link \( l \in r_s \). Suppose that the allocation \( \mu_s, b_s, l \in r_s \) satisfies

\[
\mu_s \geq \mu_s, \quad b_s \geq b_s(\mu_s), \quad l \in r_s
\]

i.e., at each link, the allocated bandwidth exceeds the minimum bandwidth \( \mu_s \), and the allocated buffer exceeds the burstiness constraint. Then, (i) no cells will be lost at any link \( l \in r_s \) [16, Proposition 4], and (ii) the end-to-end delay is at most [16, Theorem 3]

\[
\frac{b_s(\mu_s)}{b_s} + \text{propagation and processing delay}
\]

We shall assume that the “propagation and processing delay” is constant and omit it from further consideration. Consequently, a maximum end-to-end delay translates into the minimum bandwidth \( \mu_s \) required at each link along the route.

In summary, there are two service parameters for type \( s \) service: the burstiness curve \( b_s \), and the minimum bandwidth \( \mu_s \) required at each link along the route. A user message is compliant if it satisfies (3) and (4). An allocation \( \{\mu_s, b_s, l \in r_s\} \) is compliant if it satisfies (5). Given service parameters, \( \{b_s, \mu; s \in S\} \), we want to find a compliant allocation \( \{\mu, b_s; l \in r_s, s \in S\} \) that is “optimal.” We will restrict ourselves to allocations with \( b_s = b_s(\mu_s) \), since this is sufficient to prevent cell loss. We henceforth represent an allocation by a vector \( \mu = \{\mu_s, b_s(\mu_s); l \in r_s, s \in S\} \).

We now formulate the problem and present a solution, which leads to a different approach to service provisioning.

III. OPTIMAL ALLOCATION AND SERVICE PROVISIONING PROCEDURE

The network can produce any amount \( x_s \) of type \( s \) service, provided that sufficient resources are available, i.e.,

\[
x_s \leq D_s(w), \quad s \in S
\]
\[
\sum_{s} x_s \frac{\mu_s}{T_s} \leq C_i, \quad \sum_{s} x_s \frac{b_s(\mu_s)}{T_s} \leq B_l, \quad l \in L
\]

and expects a revenue of \( \sum x_s w_s \). The aggregate demand function summarizes the users’ utility such that

\[
\sum_{w_s} \int_{w_s}^{\infty} D_s(v)dv
\]

is the user surplus [21]. Take as social welfare

\[
W'(w, x, \mu) := \int D(v)dv + \sum x_s w_s
\]

so the problem is to maximize \( W'(w, x, \mu) \) subject to (6) and (7). Consider initially a fixed allocation \( p \).

**Definition 1:** A set of prices and amounts of service produced \( \{w_s(\mu), x_s(\mu)\}; s \in S \) form an equilibrium if, for all \( \{x_s\} \) satisfying (6) and (7),

\[
x_s(\mu) = \nu_s \exp\left[-w_s(\mu)\right]
\]

(9)

The conditions say that, in equilibrium, user demand is met and the network maximizes revenue.

**Proposition 1:** \( \{w_s(\mu), x_s(\mu)\} \) is an equilibrium if and only if there exist \( (\alpha(\mu), \beta(\mu)) \) such that

\[
x_s(\mu) = \nu_s \exp\left[-w_s(\mu)\right]
\]

(8)

\[
\sum_n x_s(\mu) \frac{\mu_s}{T_s} \leq C_i, \quad \sum_n x_s(\mu) \frac{b_s(\mu_s)}{T_s} \leq B_l
\]

(11)

\[
w_s(\mu) = \frac{1}{T_s} (\alpha(\mu), \mu_s > + < \beta(\mu), b_s(\mu_s) >)
\]

\[
\sum_n x_s(\mu) w_s(\mu) = \alpha(\mu), C > + < \beta(\mu), B > \tag{12}
\]

Note that (12) says that the equilibrium price equals the resource cost for providing that service. The cost is estimated by taking as the “shadow” price or rent of \( \alpha(\mu) \) per cps of bandwidth and \( \beta(\mu) \) per cell of buffer, in link \( l \).

It can be verified that there is a unique equilibrium, that the equilibrium maximizes \( W'(w, x, \mu) \) over \( w \geq 0, x \geq 0 \) subject to (7), and that the maximum welfare is

\[
W(\mu) = \sum_n x_s(\mu) + \alpha(\mu), C > + < \beta(\mu), B > \tag{13}
\]

Now suppose \( \mu \geq \mu \) can be freely chosen. We can now formally define an optimal allocation.

**Definition 2:** An allocation \( \mu \) is optimal, or welfare maximizing, if it maximizes the welfare \( W(\mu) \) in (13).

To directly maximizing (13) the network needs to know user traffic and quality parameters \( \{b_s(\mu_s)\} \). We propose a different approach which does not require such knowledge, and hence does not require any traffic policing and enforcement on the part of the network, though users may still want to shape their messages to comply with \( \{b_s(\mu_s)\} \) so that the end-to-end delay is met. From (10) and (11) and the convexity of \( G(\mu, \alpha, \beta) \) in \( \alpha, \beta \), we obtain an alternative expression for the maximum welfare,

\[
W(\mu) = \min_{(\alpha, \beta) \geq 0} G(\mu, \alpha, \beta)
\]

where

\[
G(\mu, \alpha, \beta) = \sum_s \nu_s \exp\left[-\frac{1}{T_s} (\alpha, \mu_s > - < \beta, b_s(\mu_s) >)\right]
\]

\[
\quad + < \alpha, C > + < \beta, B >
\]

Hence, \( \mu^* \) is a welfare-maximizing allocation if

\[
W(\mu^*) = \max W(\mu) = \max_{\mu \geq \mu} \min_{(\alpha, \beta) \geq 0} G(\mu, \alpha, \beta) \tag{14}
\]

The following result is key to our solution. Note that \( G \) is convex in \( (\alpha, \beta) \) but not generally concave in \( \mu \).

**Proposition 2:** There exists a saddle-point \( (\mu^*, \alpha^*, \beta^*) \) to the max-min problem (14) that is welfare-maximizing, i.e.,

\[
G(\mu^*, \alpha^*, \beta^*) = \max_{\mu \geq \mu} \min_{(\alpha, \beta) \geq 0} G(\mu, \alpha, \beta)
\]

\[
= \min_{(\alpha, \beta) \geq 0} \max_{\mu \geq \mu} G(\mu, \alpha, \beta)
\]

Note that the max–min problem in the proposition is equivalent to the following game: for all \( s \in S \),

\[
\text{player } U_s : \min_{\mu_s \geq \mu_s} \sum_{i \in T_s} (\alpha(\mu_s) + \beta b_s(\mu_s)) \tag{15}
\]

\[
\text{player } N : \min_{(\alpha, \beta) \geq 0} \max_{\mu \geq \mu} G(\mu, \alpha, \beta) \tag{16}
\]

where we recall that \( \mu_s = \{\mu_s, l \in T_s\} \). The proposition says that if player \( N \) chooses the minimizer \( (\alpha^*, \beta^*) \), then player \( U_s \) will choose the optimal \( \mu^*_s \) since \( (\mu^*, \alpha^*, \beta^*) \) is a saddle point. Note again that \( \alpha \) and \( \beta \) in (15) can be conveniently interpreted as the rent for one unit of bandwidth and buffer, respectively. Even though the minimizer \( (\alpha^*, \beta^*) \) for \( \min_{(\alpha, \beta)} G(\mu^*, \alpha, \beta) \) may not be unique, it can be shown that the unit price \( w^*_s = \frac{1}{T_s} \sum_{i \in T_s} (\alpha^*_i \mu^*_s + \beta^*_i b_s(\mu^*_s)) \) for type \( s \) service is the same regardless of which minimizer is used as resource price.

This interpretation suggests the following service provisioning procedure to reach \( (\mu^*, \alpha^*, \beta^*) \). The procedure is based on an algorithm to solve the following equivalent game problem

\[
\min_{(\alpha, \beta) \geq 0} \max_{\mu \geq \mu} G(\mu, \alpha, \beta) \tag{17}
\]

Problem (17) is equivalent to (15) since the objective function in (15) is separable in \( \mu_s \).

Suppose the network charges each user during connection setup a rent of \( \alpha_l \) per cps of bandwidth and \( \beta_l \) per cell of buffer in link \( l \). The expected cost to a user per request of
A. Service Provisioning Procedure

Network Algorithm:

1) Network initializes the update period \( n = 0 \).
2) It posts a rent \((\alpha^0, \beta^0)\) in period \( n \) for resources at each link.
3) It monitors the requested bandwidth and buffers \((\mu_s^0, b_s^0)\) in the entire network, and uses this observed \((\mu^0, b^0)\) to solve
\[
\min_{(\alpha, \beta) \geq 0} \sum_s \nu_s \exp[-\frac{1}{T_s} (\alpha < \alpha_s, \mu_s^0 > - \beta, b_s^0)] + \alpha, C > + \beta, B >
\]
4) It uses any minimizer as rent \((\alpha^{n+1}, \beta^{n+1})\) in the next period. It increments \( n \) and go to step 2.

User Algorithm:

1) A user of type \( s \) service solves (using (19) below)
\[
\min_{\mu_s^0, b_s^0 \geq 0} \alpha^0 \mu_s^0 + \beta^0 b_s^0(\mu_s^0), \quad l \in r_s
\]
for minimizer \( \mu_s^0 \) and \( b_s^0 = b_s(\mu_s^0) \).
2) It requests \((\mu_s^0, b_s^0)\) and is admitted if resources are available, and rejected otherwise.
3) If admitted, it pays the rent \( \sum_{l \in r_s} (\alpha^0 \mu_s^0 + \beta^0 b_s^0) \).

The optimality of the procedure is assured by the following theorem.

**Theorem 1:** Given \((\alpha^0, \beta^0)\), construct a sequence \((\mu^n, \alpha^{n+1}, \beta^{n+1})\), \( n \geq 0 \), by
\[
\mu_s^n = \arg \min_{\mu_s^0, b_s \geq 0} \alpha^0 \mu_s^0 + \beta^0 b_s(\mu_s^0), \quad l \in r_s, s \in S
\]

Then any accumulation point of the sequence \((\mu^n, \alpha^{n+1}, \beta^{n+1})\) is a saddle point of \( G \).

The decentralized nature of the procedure is striking: given the price \((\alpha, \beta)\), the bandwidth and buffer request

\[
(\mu_s, b_s(\mu_s)) \text{ at link } l \text{ of route } r_s \text{ that minimizes the service cost (18) depends only on the rent at link } l. \]

In fact, given \((\alpha, \beta)\), the cost minimizing \(\mu_s\) satisfies, by the Kuhn-Tucker theorem [19],
\[
\mu_s = \mu_s^0 \text{ if } - \alpha^0 \beta^0 < \frac{d}{d\mu_s} b_s(\mu_s)
\]
\[
\mu_s = M_s \text{ if } - \alpha^0 \beta^0 > \frac{d}{d\mu_s} b_s(M_s)
\]

Here, \( M_s \) is the peak message rate at which \( b_s(M_s) = 0 \). Since \( b_s(\mu_s) \) is strictly decreasing and convex for \( \mu_s \), the optimal \( \mu_s^0 \) is unique for each \((\alpha^0, \beta^0)\). Hence, in each period \( n \), a user minimizes the expected cost (18) by requesting resources according to (19); and this computation can be done locally for each link along the route. By Theorem 1, the service provisioning procedure will achieve a welfare-maximizing allocation. The optimal price \((\alpha^*, \beta^*)\) determines a minimum-cost allocation \(\mu^*\). \((\mu^*, \alpha^*, \beta^*)\) yields the service price \( x^* \) and the amount \( x^* = D_s(w^*) \) of service produced that form an equilibrium.

IV. CONCLUSION

We have studied an alternative approach to service provisioning in an ATM network, which does not require the network to know user needs. In this approach, the network offers directly for rent its bandwidth and buffers and the users purchase them freely to meet their desired quality. The service provisioning procedure is based on a solution of the problem of allocating bandwidth and buffers to meet several types of service requests, differentiated by bounds on the average rate and burstiness of the message and on the end-to-end delay.

The model has one serious deficiency. Once network resources are allocated to a service request, those resources cannot be shared by other connections. Such “exclusion” is necessary since no cell loss is allowed. However, if one does permit cell loss (as a service quality parameter), then it is possible, indeed desirable, to share resources among concurrent connections, i.e. to permit statistical multiplexing. This will require an understanding of the interaction between resource allocations, cell loss, and delay. A preliminary attempt in this direction is reported in [23].

APPENDIX

**Proof of Proposition 1:** By the duality theorem of linear programming, \((x_s(\mu))\) satisfies (9) if and only if there exists \((\alpha(\mu), \beta(\mu)) \geq 0 \) such that
\[
\frac{1}{F_s} (\alpha(\mu), \mu_s > < \beta(\mu), b_s(\mu_s) > ) \geq w_s(\mu)
\]
with equality if \( x_s(\mu) > 0 \). Moreover,
\[
\sum x_s(\mu) w_s(\mu) = < \alpha(\mu), C > + < \beta(\mu), B >
\]

The result now follows upon noting that, in equilibrium, \( x_s(\mu) = \nu_s \exp[-w_s(\mu)] > 0 \).
We now proceed to the proof of Proposition 2 and Theorem 1. Since $G(\mu, \alpha, \beta)$ is not concave in $\mu$, we will use a generalization of von Neumann's minimax theorem due to Kakutani [9].

Suppose that $M_s$ is the bound on the peak rate for a type $s$ message, i.e., $h_s(M_s) = 0$. Then, in the zero-sum game of the proposition, we may restrict the strategy set for players $U_s, s \in S$, who maximize $G(\mu, \alpha, \beta)$, to the following compact and convex set

$$S_1 := \{ \mu | \mu_s \leq M_s; \mu \in \mathbb{R}^n, s \in S \}$$

Since $G(\mu, 0, 0) = \sum_s \nu_s$, we may restrict the strategy set for player $N$, who minimizes $G(\mu, \alpha, \beta)$, to the following compact and convex set

$$S_2 := \{ (\alpha, \beta) \geq 0 | \sum_s (\alpha_s C_s + \beta_s B_s) \leq \sum_s \nu_s \}$$

For each $(\alpha, \beta)$, player $U_s$'s optimal strategy $R_1(\alpha, \beta)$, given by (19), is unique. For each $\mu$, denote by $R_2(\mu)$ the set of minimizers for (16). Both $R_1$ and $R_2$ are nonempty. The convexity of $G$ in $(\alpha, \beta)$ implies that $R_2(\mu)$ is convex for each $\mu$.

**Lemma 1:** $R_1$ is continuous on $S_2$, $R_2$ is upper semi-continuous on $S_1$, i.e., if $\mu^n \to \mu^*$, $(\alpha^n, \beta^n) \in R_2(\mu^n)$ and $(\alpha^n, \beta^n) \to (\alpha^*, \beta^*)$, then $(\alpha^*, \beta^*) \in R_2(\mu^*)$.

**Proof:** We shall prove that $R_2$ is upper semi-continuous. The proof for $R_1$ is similar. Suppose in contradiction that $(\alpha^*, \beta^*) \notin R_2(\mu^*)$. Fix $(\alpha, \beta) \in R_2(\mu^*)$. Then

$$\epsilon := G(\mu^*, \alpha^*, \beta^*) - G(\mu^*, \alpha^*, \beta^*) > 0$$

By continuity of $G$, for sufficiently large $n$,

$$G(\mu^n, \alpha^n, \beta^n) > G(\mu^*, \alpha^*, \beta^*) - \frac{\epsilon}{2}$$

Continuity of $G$ at $(\mu^*, \alpha^*, \beta^*)$ guarantees a neighborhood $N(\mu^*, \alpha^*, \beta^*)$ of $(\mu^*, \alpha^*, \beta^*)$ in $S_1 \times S_2$ such that for sufficiently large $n$, for all $(\mu^n, \alpha^n, \beta^n) \in N(\mu^*, \alpha^*, \beta^*)$,

$$G(\mu^n, \alpha^n, \beta^n) \leq G(\mu^*, \alpha^*, \beta^*) + \frac{\epsilon}{2}$$

$$= G(\mu^*, \alpha^*, \beta^*) - \frac{\epsilon}{2} \quad \text{(By definition of $\epsilon$)}$$

$$< G(\mu^n, \alpha^n, \beta^n)$$

contradicting $(\alpha^n, \beta^n) \in R_2(\mu^n)$.

**Proof of Proposition 2:** To prove the existence of a saddle point, define the mapping $\chi : S_1 \times S_2 \to S_1 \times S_2$

$$\chi(\mu, \alpha, \beta) = (R_1(\alpha, \beta), R_2(\mu))$$

where $S_1 \times S_2$ denotes the collection of all subsets of $S_1 \times S_2$. $(\alpha^*, \beta^*)$ is a saddle point if and only if it is a fixed point of $\chi$, because

$$(\mu^*, \alpha^*, \beta^*) \in \chi(\mu^*, \alpha^*, \beta^*)$$

is equivalent to

$$\mu^* \in R_1(\alpha^*, \beta^*)$$

$$(\alpha^*, \beta^*) \in R_2(\mu^*)$$

which is equivalent to

$$G(\mu, \alpha^*, \beta^*) \leq G(\mu^*, \alpha^*, \beta^*) \leq G(\mu^*, \alpha^*, \beta^*)$$

for all $(\mu, \alpha, \beta) \in S_1 \times S_2$ We will prove that $\chi$ has a fixed point using Kakutani's theorem [9] which states that, given a compact and convex set $X$, if $f$ is an upper semi-continuous function which assigns to each $x \in X$ a closed and convex subset of $X$, then there exists some $x \in X$ such that $x \in f(x)$. Now $S_1 \times S_2$ is compact and convex, and by Lemma 1 $\chi$ is upper semi-continuous. $\chi(\mu, \alpha, \beta) = R_1(\alpha, \mu) \times R_2(\mu)$ is convex since each component is. We only need to show that $\chi(\mu, \alpha, \beta)$ is closed. $R_1(\alpha, \beta)$ is a singleton and hence closed. The continuity of $G$ implies that $R_2(\mu)$ is closed. Hence $\chi(\mu, \alpha, \beta)$ is closed.

**Proof of Theorem 1:** Since $S_1 \times S_2$ is compact, accumulation points of $(\mu^n, \alpha^{n+1}, \beta^{n+1})$ always exist.

Define the mapping $\psi : S_1 \times S_2 \to S_1 \times S_2$

$$\psi(\mu, \alpha, \beta) = (R_1(\alpha, \beta), R_2(\mu))$$

Note that if player $N$ starts with the strategy $(\alpha, \beta)$, then players $U_s, s \in S$, will pick $R_2(\alpha, \beta)$, to which $N$ will respond by picking a strategy in $R_2(R_1(\alpha, \beta))$, and so forth. The following three observations complete the proof:

1) The sequence in the theorem can be constructed, given $(\nu_s, (\alpha^n, \beta^n) \in R_2(\mu^n)$ and $(\alpha^n, \beta^n) \to (\alpha^*, \beta^*)$, then $(\alpha^*, \beta^*) \in R_2(\mu^*)$.

2) If $(\nu^n) \in \gamma((\alpha^n, \beta^n) \to (\alpha^*, \beta^*)$, then by the upper semi-continuity of $\psi$ and the first observation, $(\mu^*, \alpha^*, \beta^*)$ is a fixed point of $\psi$, i.e., $(\mu^*, \alpha^*, \beta^*) \in \psi(\mu^*, \alpha^*, \beta^*)$.

3) As in the proof of Proposition 2, $(\mu^*, \alpha^*, \beta^*)$ is a fixed point of $\psi$ if and only if it is a saddle point.

**REFERENCES**


Steven H. Low (S'88-M'92) received the B.S. degree from Cornell University in 1987, M.S. and Ph.D. degrees from the University of California-Berkeley in 1990 and 1992, respectively. He is a member of technical staff at AT&T Bell Labs. His areas of research are in the control and optimization of communication networks and protocols.