Bandwidth Allocation for Guaranteed versus Best Effort Service Categories

Eitan Altman[1] Ariel Orda[2] Nahum Shimkin[3]

Abstract— **Modern computer architectures evolve towards integration of guaranteed-performance and best-effort service types. The co-existence of these two service types offers substantial benefits, such as resource sharing between service classes, and the ability of the user to select an appropriate service class according to its individual requirements and preferences. Notwithstanding, such interaction potentially complicates the system behavior, and gives rise to subtle optimization questions, which need to be explored and understood in order to allow efficient network operation.**

In this paper we address some essential performance and flow control issues associated with such service interactions. We propose a fluid model for session flow, which captures the two interaction mechanisms of resource sharing. In particular, our model incorporates the possibility of *session migration*, where sessions may shift from best effort to guaranteed performance service due to congestion experienced in the former. Within this model, we analyze the system performance and characterize its steady state behavior. We further show that under certain conditions the system exhibits bistable behavior, where some transient congestion may stir the system from a stable and efficient operating point to an inefficient and congested one, which might persist indefinitely. For the latter case, we propose a call admission control scheme that, while not interfering with normal system operation, it prevents the system from getting trapped in a congested-type equilibrium.

I. Introduction

Broadband networks are designed to offer several service categories. For example, in ATMs, real time traffic would typically use CBR (Constant Bit Rate) and rt-VBR (real-time Variable Bit Rate) services, whereas non-real-time traffic would use the nrt-VBR (non-real-time Variable Bit Rate), ABR (Available Bit Rate), UBR (Unspecified Bit Rate) or ABT (ATM Block Transfer) services. Such trends are being experienced also with IP technology, towards the support of QoS flows and provision of various service guarantees (see, e.g., [3], [12], [5]).

We classify such services into two basic categories. In the first, which we call *guaranteed performance*, a fixed amount of bandwidth is reserved for the whole duration of the session. In particular, the allocated bandwidth does not change according to the congestion state of the system. In the second, which we call *best effort*, the bandwidth allocation may change dynamically according to both the momentary session requirements and the current bandwidth availability in the network. For example, the CBR service in ATM networks belongs to the guaranteed performance category, while the ABR service belongs to the best effort category.

The main goal of this paper is to analyze some basic performance issues that result from the co-existence of and interaction between best effort and guaranteed performance services. In particular, we explore the consequences of *dynamic resource allocation* and *flow migration*. Employing dynamic resource allocation allows to efficiently share network resources, mainly bandwidth and buffer space, between the two service classes. While there are various possibilities for implementing such allocations, they typically share a common property: resources that are not used for guaranteed performance traffic may be (momentarily) used by best effort traffic. Flow migration relates to the option of best effort clients to turn to guaranteed performance service when the former is congested and does not supply satisfactory performance. For example, for non real time traffic in ATM, one might prefer nrt-VBR instead of ABR when the number of ABR connections becomes large.

We formulate a fluid flow model, which effectively incorporates the essential characteristics of such systems. We investigate the performance measures of the network, namely delays and throughputs, and characterize its steady-state behavior. Our analysis reveals that multiple and persistent equilibria may exist. Specifically, the system may move from a non-congested to a congested equilibrium due to transient congestion, which results with an unfavorable and persistent change in the the distribution of resources between service classes. We obtain conditions for the stable operation of the network, and identify the cases where multiple equilibria exist. For such cases, we propose and validate a call admission control scheme that prevents the system from getting trapped in a congested-type equilibrium.

Our model is described in Section II, and equilibrium analysis is presented in Section III. Section IV describes the proposed admission control mechanism. Finally, Section V presents concluding remarks.

II. The model

We consider a communication network which offers two types of service classes, namely *guaranteed performance* and *best effort*. Guaranteed performance is maintained by reserving appropriate network resources for the entire session duration, according to the service guarantees negotiated initially. The users, *i.e.*, incoming sessions, are free to choose the service class according to their service requirements and preferences.

A continuous-time fluid approximation is used to model

[1]INRIA B.P. 92, 2004 Route des Lucioles, 06903 Sophia-Antipolis Cedex, France. `Eitan.Altman@sophia.inria.fr`

[2]Department of Electrical Engineering, Technion—Israel Institute of Technology, Haifa 32000, Israel. {`ariel`}`@ee.technion.ac.il`

[3]Department of Electrical Engineering, Technion—Israel Institute of Technology, Haifa 32000, Israel. {`shimkin`}`@ee.technion.ac.il`

the system. The number of sessions in each service class (or the workload associated with these sessions) is represented by a continuous variable, and the arrival and service processes, as well as the session length, are described by their averages. This fluid model can be formally obtained by considering a sequence of discrete stochastic models, where the arrival rates, service rates, and total available bandwidth are scaled by the same factor N; our fluid model is then obtained by considering the behavior of the system for large N (see e.g. [4]). Specifically, The model comprises of the following elements.

1. *Guaranteed-Performance (GP) service class:*
(a) $x(t)$ denotes the number of active sessions at time t in service class GP.
(b) λ_{gp} denotes the rate of *external* (new) sessions arriving to GP, and is assumed to be constant.
(c) *GP service rate:* Each session which is admitted to GP service is allocated a fixed amount of network resources, so that its service rate is not degraded by congestion. Let μ_1^{-1} denote the average duration of a GP session. Then the service rate in GP, expressed in sessions per second, is $\mu_{gp}(t) := \mu_1 x(t)$.
(d) *GP admission control:* The number of sessions $x(t)$ in GP is bounded by a constant $x_{\max}$. This bound may be enforced through an appropriate admission control mechanism. The bound could be equal to the whole available bandwidth. It is, however, useful to allow for that bound to be strictly smaller than the total available bandwidth, so as to avoid periods in which BE traffic cannot be transmitted.
2. *Best Effort (BE) service class:*
(a) $y(t)$ denotes the number of active sessions at time t in service class BE.
(b) λ_{be} denotes the rate of *external* (new) sessions arriving to BE, and is assumed to be constant.
(c) Let μ_{be} denote the available service rate, in sessions per second, at the BE service class. This service rate depends of course on the available network resources, which in turn depend on the load at GP. To quantify this dependence, we shall make the simplistic assumption that the system has a finite and fixed amount W_{max} of available effective bandwidth (which consolidates the restrictions imposed by link capacities, finite buffers, etc. and their interaction with service quality requirements), which is utilized by BE and GP in a linear proportion to their momentary service rate. Thus, a service rate μ_{gp} in GP requires $\alpha_{gp}\mu_{gp}$ units of effective bandwidth, and similarly μ_{be} in BE requires $\alpha_{be}\mu_{be}$. Here α_{gp} and α_{be} are fixed parameters which are not necessarily equal, due to different data transmission efficiences, partial utilization of reserved resources in GP, etc. This is further elaborated at the end of this section. It follows that the following constraint applies: $\alpha_{gp}\mu_{gp} + \alpha_{be}\mu_{be} \leq W_{max}$. Denoting $\mu_{\max} := W_{max}/\alpha_{be}$ and $\alpha := \alpha_{gp}/\alpha_{be}$, this constraint reads:

$$\mu_{be} \leq \mu_{\max} - \alpha\mu_{gp}\,. \tag{1}$$

We refer to the parameter α as the *bandwidth ratio.* Hence, assuming that all leftover bandwidth from GP is applied to BE, the total service rate available for BE is given by

$$\mu_{be} = \mu_{\max} - \alpha\mu_{gp} = \mu_{\max} - \alpha\mu_1 x\,. \tag{2}$$

With service distributed evenly among sessions, the service rate per unit of BE session is then $\mu_2 := \mu_{be}/y = (\mu_{\max} - \alpha\mu_1 x)/y$.
3. *Session migration:* The service rate per session in BE decreases with the system load. For some applications that normally use best effort services, this amount may be too low if x and/or y are large, and might result in unacceptable performance. Thus, some sessions which experience degraded performance in BE might prefer to shift to GP. We call this the *migration phenomenon.* Quantitatively, the rate of BE sessions shifting to GP is given by

$$\lambda_{mi} := y \cdot c(x,y), \qquad 0 \leq x \leq x_{max}, 0 \leq y, \tag{3}$$

where c is the per-session migration rate function, which is positive and non-decreasing in both arguments. Further specification of c is not required at the moment.
For convenience we assume that the average service requirement at GP of a migrating BE session is identical to that of an original GP session; this can always be arranged by re-scaling $y(t)$.

Some assumptions are required regarding the system parameters. First, note that the resource constraint (1) implies that μ_{gp} cannot exceed $\mu_{\max}/\alpha$, so that the maximal allowed load at GP must satisfy:

$$x_{\max} \leq \mu_{\max}/\alpha\mu_1\,. \tag{4}$$

Additionally, we assume that GP has enough bandwidth to handle its externally arriving traffic (namely, without migrating sessions from BE) without overflow. This requires $\lambda_{gp} < \mu_1 x_{\max}$, since the latter is the maximal service rate at GP. This requirement can be expressed as

$$x_{\min} := \frac{\lambda_{gp}}{\mu_1} < x_{\max}\,. \tag{5}$$

$x_{\min}$ is the minimal level required at GP to handle the external traffic alone, and hence can serve as an effective lower bound for $x(t)$. Both (4) and (5) are assumed to hold in the sequel.

The resulting flow dynamics at the two service classes may now be expressed by the following pair of coupled differential equations:

$$\frac{d}{dt}x = \lambda_{gp} - \mu_{gp}(x) + \lambda_{mi}(x,y)\,, \quad 0 \leq x < x_{\max}\,, \tag{6}$$

$$\frac{d}{dt}y = \lambda_{be} - \mu_{be}(x) - \lambda_{mi}(x,y)\,, \quad y > 0\,, \tag{7}$$

where

$$\mu_{gp}(x) = \mu_1 x\,, \tag{8}$$

$$\mu_{be}(x) = \mu_{\max} - \alpha\mu_{gp}(x) = \mu_{\max} - \alpha\mu_1 x\,, \tag{9}$$

$$\lambda_{mi}(x,y) = yc(x,y)\,. \tag{10}$$

The boundaries $x = x_{\max}$ and $y = 0$ require special attention to prevent the state variables from exceeding their feasible regions. (The boundary at $x = 0$ presents no such problem since the service rate $\mu_{gp} = \mu_1 x$ in nullified there.) Consider first the case of $y = 0$. Since the derivative of y cannot be negative, and noting that $\lambda_{mi} = 0$ there, we obtain:

$$\begin{aligned}
\frac{d}{dt}x &= \lambda_{gp} - \mu_{gp}(x)\,, && (11)\\
\frac{d}{dt}y &= [\lambda_{be} - \mu_{be}(x)]^+ && (12)\\
&= [\lambda_{be} - \mu_{\max} + \alpha\mu_1 x]^+\,, \quad \text{for } y = 0\,, && (13)
\end{aligned}$$

where $[\cdot]^+$ denotes the positive part.

The boundary at $x = x_{\max}$ requires additional care, regarding its effect on the migration term λ_{mi} and hence on the variation of y.

- The derivative of x should obviously be non-positive.
- We assume that BE sessions that cannot migrate to GP because of rejection, remain in BE, and thus remain in the system. Thus, in practice, only real GP sessions suffer from actual rejection.
- When the total arrival rate to GP is larger than the service rate, *the surplus* $[\lambda_{gp}+\lambda_{mi}-\mu_{gp}]$ *must be rejected.* The migration behavior will clearly change and will not follow (3). The actual arrival rates $\tilde{\lambda}_{mi}$ and $\tilde{\lambda}_{gp}$ must now satisfy the constraint $\tilde{\lambda}_{gp}+\tilde{\lambda}_{mi}-\mu_{gp} = 0$, Thus, in particular, $\tilde{\lambda}_{mi}$ will be upper bounded by μ_{gp} (in contrast with (3), where $\tilde{\lambda}_{mi}$ is seen to tend to infinity when $y \to \infty$). Since some of the attempts of BE to join the GP service fail due to rejection, we shall have $\tilde{\lambda}_{mi} \le \lambda_{mi}$. The actual value of λ_{mi} will depend on the actual type of BE applications (e.g. the willingness to retry to migrate) as well as on parameters such as pricing.

It follows that $0 \le \tilde{\lambda}_{gp} \le \lambda_{gp}$, and that $\tilde{\lambda}_{mi} = \tilde{\lambda}_{mi}(y)$ must satisfy

$$\mu_{gp} - \lambda_{gp} \le \tilde{\lambda}_{mi} \le \min\{\mu_{gp}, \lambda_{mi}\}\,, \qquad (14)$$

when $\lambda_{gp} + \lambda_{mi} > \mu_{gp}$. We shall further assume that $\tilde{\lambda}_{mi}$ is continuous and increasing in y. (We note that in the particular case where the rejection proportion of the two arrival types is the same, and rejected BE sessions do not retry to join GP, we obtain $\tilde{\lambda}_{mi} = \frac{\mu_{gp}}{\lambda_{gp}+\lambda_{mi}}\lambda_{mi}$.) To complete the specification of $\tilde{\lambda}_{mi}$ we define $\tilde{\lambda}_{mi} = \lambda_{mi}$ when $\lambda_{gp} + \lambda_{mi} \le \mu_{gp}$. The above two constraints on $\tilde{\lambda}_{mi}$ can be summarized as follows:

$$\min\{\mu_{gp} - \lambda_{gp}, \lambda_{mi}\} \le \tilde{\lambda}_{mi} \le \min\{\mu_{gp}, \lambda_{mi}\}\,, \qquad (15)$$

where all quantities are computed at the point $(x_{\max}, y)$. Thus, finally we obtain

$$\begin{aligned}
\frac{d}{dt}x &= \max\{0,\, \lambda_{gp} - \mu_{gp} + \lambda_{mi}\}\,. && (16)\\
\frac{d}{dt}y &= \lambda_{be} - \mu_{be} + \tilde{\lambda}_{mi}\,, \quad \text{for } x = x_{\max}\,. && (17)
\end{aligned}$$

Having specified the dynamic model, we turn now to discuss the bandwidth ratio parameter $\alpha = \alpha_{gp}/\alpha_{be}$. This constant reflects the fact that a session which migrates from BE to GP might require more ($\alpha > 1$) or less ($\alpha < 1$) system resources (effective bandwidth) than required by it originally. This depends on the exact nature of the traffic, the exact service category (CBR, VBR etc), and the resource reservation scheme. Here are some possible scenaria that illustrate the above phenomenon.

- $\alpha < 1$: Consider *data traffic*, or more precisely, file transfer, that uses initially the ABR service category. Assume that due to congestion, the user prefers to switch to CBR. When using CBR service, the whole available bandwidth is consumed. However, in the ABR mode, most controllers adapt only to a *fraction* of the available bandwidth. In particular: (i) Controllers that are based on congestion indication cause oscillations of the throughputs and under-utilize the available bandwidth [9], [10]; this is partially due to the end-to-end delays in the system and to the fact that the source is limited in the rate at which it is allowed to increase (and also decrease) its transmission rate [1]. As evident from [9], [10], during part of the time, the queues may be empty; at such epochs, the throughput equals to the source transmission rate and might be smaller than the available bandwidth. (ii) Controllers that are based on explicit rate indication, generated by the switch, often try to achieve a fraction, smaller than one, of the available bandwidth (e.g., link utilization of the controller ERICA [8]).
- $\alpha > 1$: Although ABR is primarily a service category for non-real time traffic, when the load is small, users might be tempted to use it also for real-time applications, e.g., due to price considerations. When such an application does run with VBR service (at times of congestion), it will typically require a peak cell rate higher than its sustainable cell rate, especially if it is bursty. The bandwidth that should be allocated to that source should thus take into account its burstiness requirements as well as the loss probabilities (see, e.g., [6], [11]).

III. Equilibrium Conditions and their Stability

In this section we characterize the equilibrium conditions at which our system can operate.

A basic quantity which influences the stability properties of the system is the (overall) load factor, which is defined as follows. Let

$$\lambda := \lambda_{gp} + \lambda_{be}\,, \quad \mu(x) := \mu_{gp}(x) + \mu_{be}(x) \qquad (18)$$

denote the combined arrival and available service rates. Define ρ as their ratio:

$$\begin{aligned}
\rho(x) &:= \frac{\lambda}{\mu(x)} = \frac{\lambda}{(\mu_1 x) + (\mu_{\max} - \alpha\mu_1 x)} && (19)\\
&= \frac{\lambda}{\mu_{\max} + (1-\alpha)\,\mu_1 x}\,.
\end{aligned}$$

It may be seen that ρ depends on the load x at GP, and may be increasing or decreasing in x, depending on α being larger or smaller than unity. As we shall see, these two possibilities will lead to different system behaviors.

As a first step, we make the following distinction between systems based on their overall loading conditions.

a. Under-loaded system: $\rho(x) < 1$ for all feasible x.
b. Over-loaded system: $\rho(x) > 1$ for all feasible x.
c. Non-definite loading: Both $\rho(x) > 1$ and $\rho(x) < 1$ are possible, depending on x.

As may be expected, the first case leads to stable system operation with minimal session backlog. However, this case may correspond to over-conservative system design, especially when α is significantly different from unity. The second case will inevitably lead to buffer overflow and session rejection. For completeness, these two cases are briefly treated below. Our main concern and interest shall lie in the case of non-definite loading, where the analysis will be conducted separately for $\alpha > 1$ and $\alpha < 1$.

Before proceeding, we need to define what we mean by "feasible x" in the above definitions of system loading conditions. As explained in the previous section, x is upper bounded by $x_{\max}$, and lower bounded by $x_{\min}$ (any value $x < x_{min}$ is transient, and after some finite time, it is never visited by the state trajectory). Thus, x is said to be feasible if $x_{\min} \le x \le x_{\max}$.

We now turn to analyze the equilibrium conditions which may prevail in our system for the different cases mentioned above. Note that the term *equilibrium* refers to a point $p_e = (x_e, y_e)$ for which $\dot{x} = \dot{y} = 0$. Also, as we shall see, under certain (overflow) conditions a diverging trajectory $\{x = x_{\max}$ and $y \to \infty\}$ is obtained; for convenience we also refer to that behavior as an equilibrium of the system. (Obviously, in practice y cannot increase without bound, and will stabilize around some large, finite value.)

A. Under-loaded system

We assume here that $\rho(x) < 1$ for $x_{\min} \le x \le x_{\max}$. Summing (6) and (7), we obtain

$$\frac{d}{dt}(x+y) = \lambda - \mu(x) < 0 \tag{20}$$

for $x < x_{\max}$ and $y > 0$ (the inequality follows from $\rho < 1$). Similarly, for $y > 0$ and $x = x_{\max}$ we obtain $\frac{d}{dt}(x+y) \le \lambda - \mu(x) < 0$. Evidently, this precludes any equilibrium (x_e, y_e) with $y_e > 0$.

Consider then $y \equiv 0$. The corresponding equilibrium value for x is obtained from (11), namely $\lambda_{gp} - \mu_{gp}(x) = \lambda_{gp} - \mu_1 x = 0$ yields $x = \lambda_{gp}/\mu_1 \equiv x_{\min}$. Furthermore, since $\dot{y} = [\lambda_{be} - \mu_{be}(x)]^+$ for $y = 0$, it is required that $\lambda_{be} - \mu_{be}(x_{\min}) \le 0$; however this follows since, by assumption, $\rho(x_{\min}) < 1$ and since, as we just observed, $\lambda_{gp} = \mu_{gp}$ at $x = x_{\min}$.

It follows that the system has a single equilibrium point $p_e = (x_e, y_e)$ at $(x_{\min}, 0)$. Furthermore, it can be seen that this equilibrium is globally asymptotically stable, namely for any feasible initial conditions of $p(t) = \Big(x(t), y(t)\Big)$ we obtain $\lim_{t\to\infty} p(t) = p_e$. This can be verified as follows. First, we note that

$$\frac{d}{dt}x = \lambda_{gb} - \mu_1 x + \lambda_{mi} \ge \lambda_{gb} - \mu_1 x = \mu_1(x_{\min} - x) , \tag{21}$$

with equality holding if $y = 0$. Assuming first that $x(0) > x_{\min}$, it follows that $x(t) \ge x_{\min}$ for all $t \ge 0$. We now show that $(x + y)$ is monotonically decreasing. Since $\rho(x) < 1$ over $x_{\min} \le x \le x_{\max}$, it follows that $\lambda - \mu(x) \le -\varepsilon_o$ there, for some fixed $\varepsilon_o > 0$. Thus, for $y > 0$ we have (see (20)):

$$\frac{d}{dt}(x+y) = \lambda - \mu(x) < -\varepsilon_o . \tag{22}$$

If $y = 0$, then $\dot{x} = \mu_1(x_{\min} - x) \le 0$, and thus

$$\begin{aligned} \frac{d}{dt}(x+y) &= \Big[\lambda_{be} - \mu_{be}(x)\Big]^+ + \mu_1(x_{\min} - x) \\ &= \max\{\lambda - \mu(x) ,\ \mu_1(x_{\min} - x)\} \\ &\le \max\{-\varepsilon ,\ \mu_1(x_{\min} - x)\} \le 0 . \end{aligned}$$

It follows then that $(x + y)$ (which can be viewed as a Lyapunov function) decreases at least exponentially fast to $x_{\min}$, implying that x decreases to $x_{\min}$ and y to 0. In fact, it is easily verified that y reaches 0 within a finite time $t_0 \le [x(0) - x_{\min} + y(0)]/\varepsilon_o$, and then x converges exponentially to $x_{\min}$ with time constant $(\mu_1)^{-1}$.

Finally, if $x(0) < x_{\min}$, it follows from (21) that $x(t)$ converges at least exponentially to $\{x \ge x_{\min}\}$, and the above argument may be repeated.

B. Over-loaded system

Here we assume that $\rho(x) > 1$ for $x_{\min} \le x \le x_{\max}$, implying that $\lambda - \mu(x) > \varepsilon > 0$ in that region. The total flow equation now yields

$$\frac{d}{dt}(x+y) = \lambda - \mu(x) > \varepsilon , \tag{23}$$

for $x < x_{\max}$ and $y > 0$.

Similarly, for $y = 0$ we obtain $\frac{d}{dt}(x+y) \ge \lambda - \mu(x) > \varepsilon$. Thus, beyond some finite time we must have $x = x_{\max}$, or $y(t) \to \infty$. However, since $\lambda_{mi} = yc(x, y)$ increases to infinity with y, it follows that $x = x_{\max}$ is obtained in the second case as well.

Given $x = x_{\max}$, the asymptotic behavior of y may be obtained from (17), namely $\dot{y} = \lambda_{be} - \mu_{be}(x_{\max}) - \tilde{\lambda}_{mi}(y) := f(y)$, and may depend on the specific definition of $\tilde{\lambda}_{mi}(y)$. We note first that $f(y)$ decreases in y since $\tilde{\lambda}_{mi}$ increases in y; thus, if $f(y) = 0$ is solvable for some y_0 (equivalently, $\lambda^*_{mi} := \lim_{y\to\infty} \tilde{\lambda}_{mi}(y) \ge \lambda_{be} - \mu_{be}(x_{\max})$), then $y(t) \to y_0$. Otherwise, $y(t) \to \infty \overset{\text{def}}{=} y_0$. In either case, we say that the system is in *overflow equilibrium*, and

$$\text{Rejection rate of actual GP } = \frac{\tilde{\lambda}_{gp}(y_0)}{\lambda_{gp}} = \frac{\mu_{max} - \tilde{\lambda}_{mi}(y_0)}{\lambda_{gp}} .$$

C. Non-definite loading, with $\alpha > 1$

For $\alpha > 1$, we can see from (19) that $\rho(x)$ is increasing in x. The non-definite loading condition is thus equivalent to $\rho(x_{\min}) < 1$, and $\rho(x_{\max}) > 1$. Under these conditions the system will have multiple equilibrium points, as summarized below.

proposition 1: For the case of non-definite loading and $\alpha > 1$, there exist exactly three equilibrium points (x_e, y_e), with the following characteristics:

(i) $x_e = x_{\min}$, $y_e = 0$.
(ii) $x_e \in (x_{\min}, x_{\max})$ satisfying $\rho(x_e) = 1$, and $y_e > 0$.
(iii) $x_e = x_{\max}$ (overflow equilibrium).

Proof: Consider $y_e = 0$ first. From (11) and $\dot{x} = 0$ we obtain $x_e = x_{\min} = \frac{\lambda_{gp}}{\mu_1}$. The condition for this point to be an equilibrium is then $\dot{y} = [\lambda_{be} - \mu_{be}(x_{\min})]^+ = 0$. However, $\rho(x_{\min}) < 1$ implies that $\mu(x_{\min}) := \mu_{be}(x_{\min}) + \mu_{gp}(x_{\min}) > \lambda := \lambda_{be} + \lambda_{gp}$, and $\mu_{gp}(x_{\min}) = \lambda_{gp}$ by definition of $x_{\min}$. It follows that indeed $[\lambda_{be} - \mu_{be}(x_{\min})] < 0$, and $(x_{\min}, 0)$ is an equilibrium.

Next, consider a possible equilibrium with $y_e > 0$ and $x_e < x_{\max}$. Equating the total flow to zero gives here $\frac{d}{dt}(x + y) = \lambda - \mu(x) = 0$, namely $\rho(x_e) = 1$. However, by continuity and monotonicity of ρ and the non-definite loading condition $\rho(x_{\min}) < 1$ and $\rho(x_{\max}) > 1$, it follows that $\rho(x_e) = 1$ is satisfied for a unique $x_e \in (x_{\min}, x_{\max})$. The equilibrium equation for y is then $\dot{y} = \lambda_{be} - \mu_{be}(x_e) - \lambda_{mi}(x_e, y) = 0$. Recalling that $\lambda_{mi}(x_e, y) = yc(x_e, y)$ is strictly increasing in y, this equation will have a unique solution $y_e > 0$ provided that $\lambda_{be} - \mu_{be}(x_e) > 0$. However, the latter follows from $\mu(x_e) = \lambda$ and $\mu_{gp}(x_e) > \mu_{gp}(x_{\min}) = \lambda_{gp}$. Thus, there exists a unique equilibrium (x_e, y_e) with $x_e < x_{\max}$ and $y_e > 0$, which is defined by $\rho(x_e) = 1$ and $\lambda_{mi}(x_e, y_e) = \lambda_{be} - \mu_{be}(x_e)$.

Finally, consider a possible "overflow equilibrium" with $x_e = x_{\max}$. Noting that $\rho(x_{\max}) > 1$ by assumption, the situation here is similar to the overloaded system case discussed in the previous subsection. Thus, if $f(y) := \lambda_{be} - \mu_{be}(x_{\max}) - \tilde{\lambda}_{mi}(y) = 0$ is solvable for some y_0, then $(x_{\max}, y_0)$ is an equilibrium point. Otherwise, $y(t) \to \infty$ results. □

To understand the long-term system behavior, we need to determine the stability properties of these equilibrium points. Our stability definitions follow the standard definitions in the sense of Lyapunov (c.f. [7] or [13]). Namely, an equilibrium point $p_e = (x_e, y_e)$ is *stable* if for any neighborhood (or open ball) B_ε of p_e there exists another (small enough) neighborhood B_δ so that $p(0) \in B_\delta$ implies $p(t) \in B_\varepsilon$ for all $t \geq 0$. p_e is unstable if it is not stable in the above sense. p_e is *asymptotically stable* if it is stable and, in addition, there exists some neighborhood D of p_e such that $p(0) \in D$ implies $p(t) \to p_e$. Stability of the overflow equilibrium $\{x = x_{\min}, y(t) \to \infty\}$ may be defined similarly with respect to the x coordinate only.

proposition 2: The equilibria (i) and (iii) are asymptotically stable, while (ii) is unstable.

Proof: Stability of (i) follows by noting that $\rho(x_{\min}) < 1$, so that $\rho(x) < 1$ holds in some neighborhood $(x_{\min} - \varepsilon, x_{\min} + \varepsilon)$ of $x_{\min}$. Applying locally the stability arguments used in Section III-A, for the under-loaded system, yields the (local) stability of $(x_{\min}, 0)$.

Stability of the overflow-equilibrium (iii) is argued similarly to the over-loaded system case in Section III-B. For the case of $y_e(t) \to \infty$, since $\lambda_{mi} \to \infty$ it follows that any downward deviation from $x = x_{\max}$ is offset in finite time. For $y_e = y_0 < \infty$, we argue similarly that $\frac{d}{dt}x > \varepsilon$ for some $\epsilon > 0$, for all $x < x_{\max}$ and y close enough to $(x_{\max}, y_0)$. Indeed, this follows by continuity from $\rho(x_{\max}) > 1$, implying that $\frac{d}{dt}(x + y) > \varepsilon > 0$ for (x, y) close enough to $(x_{\max}, y_0)$, while $\dot{y} = 0$ at $y = y_e$ by the equilibrium condition. Thus, here also x converges to $x_{\max}$ in finite time, and then y converges to its equilibrium value according to (12), as shown in Section III-B.

The final equilibrium point (ii), being internal, is a continuity point of the flows $(\dot{x}, \dot{y})$, and therefore its stability may be determined by direct linearization (cf. [13]). Denoting $\frac{\partial}{\partial x}\lambda_{mi} := \beta_x$ and $\frac{\partial}{\partial y}\lambda_{mi} := \beta_y$ the Jacobian of $(\dot{x}, \dot{y})$, as defined in (6) and (7), is

$$J = \begin{bmatrix} -\mu_1 + \beta_x & , & \beta_y \\ -\alpha\mu_1 - \beta_x & , & -\beta_y \end{bmatrix} . \tag{24}$$

The corresponding eigenvalues are the solutions of the characteristic equation

$$\lambda^2 + (\mu_1 - \beta_x + \beta_y)\lambda - (\alpha - 1)\,\mu_1\beta_y = 0 . \tag{25}$$

Since $\mu_1\beta_y > 0$ and $\alpha > 1$ at least one eigenvalue λ is positive, implying that the internal equilibrium (x_e, y_e) cannot be stable. □

The preceding stability results may be interpreted in terms of the load factor $\rho(x)$. Thus $\rho(x_{\min}) < 1$ accounts for the stability of $(x_{\min}, 0)$ while $\rho(x_{\max}) > 1$ accounts for stability of the overflow equilibrium at $x = x_{\max}$. The instability of the internal equilibrium (ii), where $\rho(x_e) = 1$, can be understood by noting that $\rho(x)$ increases in x, hence when x deviates from x_e a positive feedback mechanism results which further contributes to this deviation. Of course, this interpretation ignores the interaction between x and y, and thus in some cases might not coincide with the actual results, as shall be seen in the next subsection.

The observed stability properties of the system have the following implication on its operation. Initially, the system may be operating in the first equilibrium $(x_{\min}, 0)$, resulting in satisfactory performance in both service classes. However, if the load (x or y) increases momentarily beyond a certain value, the system may revert to operating in the overflow equilibrium condition, resulting in a large load in both service classes and session rejection. Specific measures may be required to mitigate this phenomena.[1]

D. Non-definite loading, with $\alpha < 1$

For $\alpha < 1$, we observe from (19) that $\rho(x)$ is decreasing in x. The non-definite loading requirement is now equivalent to $\rho(x_{\min}) > 1$ and $\rho(x_{\max}) < 1$. The equilibrium properties for this case are summarized below.

proposition 3: Consider the case of non-definite loading and $\alpha < 1$.
(i) There exists a single equilibrium point (x_e, y_e), which is internal ($y > 0$, $x < x_{\max}$) and satisfies $\rho(x_e) = 1$.
(ii) This point may be stable or unstable, depending on the specific system parameters.

[1]We observe that this type of stability characterization somewhat resembles that of an Aloha system [2]: there too, three equilibria exist, two of which are stable, and only one of which is desirable.

Proof: The possibility of equilibrium with $y=0$ (and hence $x=x_{\min}$) is easily negated by noting that $\rho(x_{\min}) > 1$ implies that $\frac{d}{dt}(x+y) > 0$ for $x=x_{\min}$. Similarly the possibility of an overflow equilibrium with $x=x_{\max}$ is negated since $\rho(x_{\max}) < 1$ implies that $\frac{d}{dt}(x+y) < 0$ there. It remains to consider a possible internal equilibrium (x_e, y_e), for which (6) and (7) hold. As seen in the previous subsection, this immediately implies that $\rho(x_e)=1$, yielding a unique value for x_e, while y_e is determined as the unique solution of $\lambda_{mi}(x_e, y) = \lambda_{be} - \mu_{be}(x_e)$.

Stability of this equilibrium may again be determined through linearization, leading to the characteristic equation (25) for the linearized system, repeated here as follows:

$$\lambda^2 + (\mu_1 - \beta_x + \beta_y)\lambda + (1-\alpha)\,\mu_1\beta_y = 0\ . \qquad (26)$$

Since $\mu_1\beta_y > 0$ and $\alpha < 1$, the last term is positive, and the stability properties depend on the value of $\gamma := \mu_1 - \beta_x + \beta_y$ at (x_e, y_e). Thus if $\gamma > 0$ the equilibrium is stable, while $\gamma < 0$ implies instability. However, recalling that $\beta_x = \frac{\partial\lambda_{mi}}{\partial x}$ and $\beta_y = \frac{\partial\lambda_{mi}}{\partial y}$, both β_x and β_y are positive according to the migration characteristics, and depending on their specific values γ may be either negative or positive. □

As observed in the above proof, the equilibrium point will be unstable if $\gamma := \mu_1 - \frac{\partial\lambda_{mi}}{\partial x} + \frac{\partial\lambda_{mi}}{\partial y} > 0$, which required that the effect of x on the migration rate λ_{mi} will be larger than the effect of y.

IV. Global stabilization of the non-congested equilibrium

As we have seen, for non-definite loading with $\alpha > 1$ the system exhibits an unfavorable bistable behavior. Unfortunately, this case is seemingly the most important one. Although $\alpha < 1$ is feasible, as explained at the end of Section II, we propose that $\alpha > 1$ would be the more common case. Furthermore, non-definite loading reflects a network which has sufficient resources to support the service demands of normal incoming traffic, albeit under proper resource usage. In this light, an under-loaded design indeed reflects an overdesign.

Specifically, we have seen that in this case there exist three equilibrium points, two of which are asymptotically stable. While the first stable equilibrium, $(x_{\min}, 0)$, provides satisfactory performance to both classes, the second is an overflow equilibrium. In this section we show that, by exercising a simple call admission control (CAC) scheme, the system globally stabilizes at the efficient (and now unique) equilibrium $(x_{\min}, 0)$.

A CAC scheme should operate on the GP class only, and refrain from rejecting BE flow[2]. Moreover, a reasonable scheme would reject sessions only when *both* x and y become large, *i.e.*, exceed some thresholds $\hat{x}$ and $\hat{y}$, respectively. This means that while y is small, e.g. $y=0$, no sessions would be rejected, even if x grows to its maximal value $x_{\max}$. Accordingly, consider the following scheme.

[2]Ideally, it should start by rejecting sessions that migrated from the BE class; however this is not possible, since the network cannot distinguish between "genuine" GP sessions and those that migrated from the BE class.

CAC Scheme

- If $x < \hat{x}$ or $y < \hat{y}$, then all newly arriving GP sessions are admitted.
- Otherwise ($x \geq \hat{x}$ and $y \geq \hat{y}$), reject a portion ϕ ($0 < \phi < 1$) of the newly arriving GP sessions, where $\phi = \phi(x, y)$ may depend on the number of GP and BE sessions.

As shall be shown, the above scheme, with the proper choice of parameters $\hat{x}$, $\hat{y}$ and ϕ, and under the standard assumptions of our model, provides the required stability result. In order to define the appropriate values of the parameters, the following discussion is in place.

Consider the original system, without the application of the CAC scheme. For a small $\epsilon > 0$, let x_ρ be such that $\rho(x_\rho) = 1-\epsilon$, *i.e.*, $x_\rho = \frac{\lambda(1-\epsilon)^{-1} - \mu_{\max}}{(1-\alpha)\mu_1}$. We proceed to characterize the region on the (x, y) plane in which $x \geq x_\rho$ and $\frac{d}{dt}x \leq -\epsilon$. Let (x, y), $x \geq x_\rho$, be in that region, *i.e.* (see (6)):

$$\lambda_{mi}(x, y) = yc(x, y) \leq \mu_1 x - \lambda_{gp} - \epsilon, \qquad (27)$$

for ϵ sufficiently small. For a fixed $x > x_\rho$, the right hand side of (27) is a positive constant, while the left hand side is monotonously increasing in y, and takes the values of 0 and ∞, respectively, for $y=0$ and $y \to \infty$. For each $x_\rho \leq x \leq x_{\max}$, define $y_\epsilon(x) \stackrel{\text{def}}{=} \sup\{y : yc(x, y) \leq \mu_1 x - \lambda_{gp} - \epsilon\}$. Note that $0 < y_\epsilon(x) < \infty$.

We proceed to introduce some additional notation. Let $\lambda_{gp}^{tot} = \lambda_{gp} + \lambda_{mi}(x, y)$ denote the total arrival rate to the GP class (prior to the application of the CAC). Denote by $\bar{\lambda}_{gp}^{tot}$ the actual arrival rate to the GP service, after the rejection imposed by the CAC scheme, that is, $\bar{\lambda}_{gp}^{tot} = (1-\phi)\cdot\lambda_{gp}^{tot}$. The CAC will be employed at states (x, y) for which in the original system we have $\frac{d}{dt}x \geq 0$, *i.e.* for which we have, instead of (27) (again by (6)):

$$\lambda_{gp}^{tot} = \lambda_{mi}(x, y) + \lambda_{gp} \geq \mu_1 x.$$

We shall choose the fraction ϕ so that when replacing in (6) $\lambda_{mi}(x, y) + \lambda_{gp}$ by $\overline{\lambda}_{gp}^{tot}$, then $\frac{d}{dt}x$ will be no larger than $-\epsilon$. In other words, we choose ϕ such that $(1-\phi(x, y))\lambda_{gp}^{tot}(x, y) \leq \mu_1 x - \epsilon$ for all (x, y) such that $x \in [x_\rho, x_{max}]$ and $y = y_\epsilon(x)$. Finally, denote by $\bar{\lambda}_{gp}$ the actual arrival rate of *external* sessions to GP.

Theorem 1: Let the above CAC scheme be applied to the system with

$$\hat{x} = x_\rho,\ \hat{y} = \min_{x_\rho \leq x \leq x_{\max}} y_\epsilon(x) \ \text{ and } \ \phi(x, y) = 1 - \frac{\mu_1 x - \epsilon}{\lambda_{gp}^{tot}(x, y)}.$$

Then, the system globally stabilizes at $p_e = (x_{\min}, 0)$, *i.e.*: for any initial conditions $p(t_0) = (x(t_0), y(t_0))$, $x(t_0) \geq 0$, $y(t_0) \geq 0$, we have $\lim_{t\to\infty} p(t) = p_e$.

The theorem is proved thorough the following sequence of lemmas.

Lemma 1: Whenever $x \geq x_\rho$, $\frac{d}{dt}x \leq -\epsilon$.

Proof: Consider $x \geq x_\rho$. If $y < \hat{y}$, then the lemma follows from the definition of $\hat{y}$. Otherwise, we have $x \geq x_\rho = \hat{x}$ and $y \geq \hat{y}$, meaning that the CAC scheme is applied. The actual arrival rate to the GP service is decreased to $\bar{\lambda}_{gp}^{tot} = (1-\phi)\cdot\lambda_{gp}^{tot} = \mu_1 x - \epsilon$. Replacing λ_{gp}^{tot} with $\bar{\lambda}_{gp}^{tot}$ in (6), the lemma follows. □

Corollary 1: Suppose that, at some time $t_1 \geq t_0$, $x(t_1) > x_\rho$. Then, there is a time t_2, $t_1 < t_2 < \infty$, such that $x(t_2) = x_\rho$.

Lemma 2: Suppose that, at some time $t_1 \geq t_0$, $x(t_1) < x_\rho$. Then, either (i) there is some time t_2, $t_1 < t_2 < \infty$, such that $x(t_2) = x_\rho$, in which case $(x+y)$ is monotonically decreasing for $t_1 \leq t < t_2$, or else (ii) for all $t \geq t_1$, $x(t) < x_\rho$ and $(x+y)$ is monotonically decreasing. Moreover, in the later case $\lim_{t\to\infty} p(t) = p_e$.

Proof: Let τ, $\tau > t_1$, be such that $x(t) < x_\rho$ for all $t_0 \leq t < \tau$. Following the same lines as in the stability proof for the under-loaded case, it can be established that $(x+y)$ is monotonically decreasing for all $t_0 \leq t < \tau$.

If there is some t_2, $t_1 \leq t_2 < \infty$, such that $x(t_2) = x_\rho$, then the above argumentation, applied to $\tau = t_2$, establishes case *(i)*.

Otherwise $x(t) < x_\rho$ for all $t \geq t_1$, and it follows from the above argumentation (taking $\tau \to \infty$) that $(x+y)$ is monotonically decreasing for all $t \geq t_1$. Following the same lines as in the stability proof for the under-loaded case, it can be shown that $\lim_{t\to\infty} p(t) = p_e$, thus establishing case *(ii)*. □

Lemma 3: Suppose that, at some time $t_1 \geq t_0$, $x(t_1) = x_\rho$. Then, at $x = x(t_1)$, $\frac{d}{dt}x \leq -\epsilon$ and $(x+y)$ is monotonically decreasing.

Proof: The first part of the lemma follows from Lemma 1. We proceed to establish the second part. For $x = x_\rho$, we have:

$$\begin{aligned}\frac{d}{dt}(x+y) &= \bar{\lambda}_{gp} + \lambda_{be} - \mu_{gp}(x) - \mu_{be}(x) \\ &\leq \lambda_{gp} + \lambda_{be} - \mu_{gp}(x) - \mu_{be}(x) < 0,\end{aligned}$$

where the first inequality follows from $\bar{\lambda}_{gp} \leq \lambda_{gp}$ and the second from $\rho(x_\rho) = 1 - \epsilon < 1$. □

Proof of Theorem 1: By Corollary 1, there is some time $t_1 \geq t_0$, such that $x(t_1) \leq x_\rho$. According to Lemma 3, at $x = x_\rho$, $\frac{d}{dt}x \leq -\epsilon < 0$, therefore for all $t \geq t_1$, $x(t) \leq x_\rho$ holds. Thus, by Lemmas 2 and 3, $\frac{d}{dt}(x+y) < 0$ for $t \geq t_1$, and the proof follows as in the under-loaded case. □

Remark 1: In the presence of CAC, one may consider again the possibility that the migration of BE to GP will actually be at a smaller rate than λ_{mi}, as in the case discussed above (expression (14)). It is easy to verify that the assertions of the the lemmas and theorem of this section would still hold. Moreover, the above phenomenon would have an additional stabilizing effect on the system. For example, the time derivative of x in Lemma 2 will become smaller.

To conclude, we stress the practical implication of the above result. Without the application of a CAC scheme, the system is bi-stable, meaning that a transient congestion may lock it in an over-loaded equilibrium. However, by exercising a simple CAC rule, while the system can still be temporarily driven to any feasible state due to transient conditions, it is guaranteed to converge back to the efficient, under-loaded equilibrium.

Our CAC mechanism has a "smooth" behavior: only a fraction of calls that use GP service are rejected when the undesirable equilibria occurs. The amount of rejection is linear in the excess of the actual overall arrival rate (λ_{gp}^{tot}) beyond the available bandwidth for GP. Thus if the excess is small then the rejection rate is small too, so that the non-congested equilibrium can be reached with a minimal intervention from the network.

An alternative way to handle occasional undesirable overflow equilibria could be to reject, during short periods, all arriving sessions, till the system stabilizes again in the desirable equilibrium (as done, for example, in the TCP/IP Tahoe congestion avoidance mechanism, where the window size for the transmission of packets is sharply reset to one when congestion is detected). The advantage of such a drastic approach is in its simplicity. However, our analysis shows that in order to avoid congestion it suffices to use our proposed CAC.

Finally, we point out that the CAC only rejects calls that use GP service. One could propose to reject also BE traffic. We are not enthusiastic about such a solution, since an application that uses a best effort service (with no guarantee on the minimum bandwidth) is already penalized by having to accept a very low throughput during congestion periods. The gain on the overall performance, which would be obtained by rejecting a session that uses just a small part of the bandwidth, would be negligible. Moreover, our approach is in agreement with the ATM Forum specification concerning ABR traffic, which has no minimum cell rate guarantee, namely: "the CAC will not block the connection attempt because of bandwidth allocated to other connections" ([1], p. 85).

V. Concluding remarks

We have analyzed in this paper the behavior of a network that provides both best-effort as well as guaranteed-performance services. In such networks, during congested periods, some best-effort applications might prefer to use guaranteed-performance service instead. We analyzed the possible overall equilibria behavior of the network due to that phenomenon, and showed how different equilibria, namely overflow equilibria and non-congested equilibria, are obtained depending on the network's and traffic's parameters. We identified four qualitative behaviors of the system: an under-loaded regime, an overloaded regime, and two non-definite regimes. We identified a case of bi-equilibria behavior, where one of the equilibria corresponds to a congested system. We then presented a call admission mechanism that ensures that the system stabilizes in the non-congested equilibrium. In all cases of overflow (congested) equilibria, the number of guaranteed-performance sessions reaches in finite time the available upper limit x_{max}. In the non-congested equilibria, the number of guaranteed-performance sessions converges, geometrically fast, to the value λ_{gp}/μ_1.

The congestion in the overflow-equilibrium case is experienced differently by best-effort and guaranteed-performance sessions. As the number of best-effort sessions becomes large, the throughput available for each session becomes small, and its sojourn time, *i.e.* the time it takes to transmit all the packets of a session, becomes unacceptably large. When the number of guaranteed-performance sessions becomes large and attains x_{max}, the congestion that the guaranteed-performance sessions experience results in an increased rejection probability of new guaranteed-performance sessions.

References

[1] The ATM Forum Technical Committee, *Traffic Management Specification*, Version 4.0, April 1996.

[2] D. Bertsekas et R. Gallager, *Data Networks*, Prentice-Hall, 1987.

[3] R. Braden, L. Zhang, S. Berson, S. Herzog and S. Jamin, "Resource reservation protocol (RSVP) - version 1 functional specification," Internet Draft, Internet Engineering Task Force, May 1997. ftp://ds.internic.net/internet-drafts/draft-ietf-rsvp-spec-15.txt.

[4] J. D. Dai and S. P. Meyn, "Stability and convergence of moments for multiclass queueing networks via fluid limit models", *IEEE Trans. Automatic Control* **40**, pp. 1889-1904, 1995.

[5] R. Guérin, S. Kamat, A. Orda, T. Przygienda and D. Williams, "QoS Routing Mechanisms and OSPF Extensions", Internet Draft, Internet Engineering Task Force, April 1997. ftp://ds.internic.net/internet-drafts/draft-guerin-qos-routing-ospf-01.txt.

[6] F. Guillemin, C. Rosenberg and J. Mignault, "Guidelines for sustainable cell rate specification in ATN networks", submitted.

[7] W. Hahn, *Stability of Motion*, Springer-Verlag, Berlin, 1967.

[8] R. Jain, "ERICA+", Contribution to ATM Forum/95-1346.

[9] H. Ohsaki, M. Murata, H. Suzuki, C. Ikeda and H. Miyahara, "Rate-based congestion control for ATM networks", *Computer Communication Review, acm-sigcomm*, special issue in ATM, ed. R. Jain and K. Y. Siu, Vol 25, No. 2, pp. 60-71, 1995.

[10] M. Ritter "network buffer requirements of the rate-based control mechanism for ABR services", submitted, 1996.

[11] C. Rosenberg and G. Hébuterne, "Dimensioning a traffic control device in an ATM network", submitted.

[12] S. Shenker, C. Partridge and R. Guérin, "Specification of Guaranteed Quality of Service," Internet Draft, Internet Engineering Task Force, February 1997. ftp://ds.internic.net/internet-drafts/draft-ietf-intserv-guaranteed-svc-07.txt.

[13] M. Vidyasagar, *Nonlinear Systems Analysis*, 2nd ed., Prentice-Hall, NJ, 1993.