Stateful vs. Stateless Admission Control: which can be the gap in utilization efficiency?

N. Blefari-Melazzi, M. Femminella

D.I.E.I. Department, Via G. Duranti 93, 06125 Perugia – Italy
Ph. +39 075 585 3630, e-mail: {blefari, femminella}@diei.unipg.it

ABSTRACT

IETF developed two main approaches to provide QoS aware services in the Internet: Integrated Services (IntServ) and Differentiated Services (DiffServ). Both have well known pros and cons (e.g., [1][2]). The stateful IntServ has a greater level of accuracy and a finer level of granularity. The stateless DiffServ possesses excellent scaling properties, but lacks a standardized admission control scheme and, upon overload in a given service class, degradation of service can occur. To provide QoS in DiffServ, three possible strategies are: i) plain and heavy over-dimensioning; ii) admission control at the borders of the DiffServ region, coupled with suitable assumptions on the distribution of the traffic within the region, which can lead to over-dimensioning, even if less severe than the previous one; iii) per-node (i.e., within the region) admission control.

Following RFC2990, we recently proposed an "admission control function which can determine whether to admit a service differentiated flow along the nominated network path [1]", i.e., the third of the above strategies. This function, named GRIP (Gauge & Gate Reservation with Independent Probing), can provide strict QoS guarantees by means of stateless DiffServ-compliant procedures. This feature is paid with a potential loss of efficiency, with respect to an ideal, stateful admission control.

The goal of this paper is to evaluate analytically such loss of efficiency, in a specific scenario. In other words, we want to estimate how much resources we can waste if we go stateless and avoid state maintenance functions. The comparison between stateless and stateful approaches is performed under the constraint of strictly offering the same performance levels, in terms of, e.g., loss probability and delay.
I. INTRODUCTION

The IntServ QoS framework can provide strict QoS guarantees, due to its stateful approach [23], [22]. On the other side, the DiffServ paradigm is stateless and can offer different treatments to different traffic classes, without strict QoS guarantees within each class [24], [6]. In fact, the DiffServ way of operation of inner routers, by itself, does not intrinsically solve the problem of controlling congestion. Upon overload in a given service class, all flows in that class may suffer a degradation of service. To guarantee pre-defined levels of performance to single flows, it is necessary to deploy an admission control function, in a DiffServ domain, to admit or reject flow setup requests [1].

Following this line of reasoning, we recently proposed an admission control scheme, named GRIP (Gauge & Gate Reservation with Independent Probing), which provides strict QoS guarantees by means of stateless DiffServ-compliant procedures.

The detailed description of GRIP and a first simple performance evaluation can be found in [3], [5]. In [8], we show that GRIP is semantically compatible with the AF PHB (defined in [26]). GRIP has been also adapted to a mobile environment, in [9], and it is being upgraded to cope with multicast communications. Finally, we implemented GRIP in a test-bed, where it is running fine and in agreement with our theoretical findings.

As regards the description of the main components of an IP QoS framework and of the IntServ and DiffServ approaches, we refer the reader to the interesting survey [20].

As for the organization of this paper, in Section II, we recall previous work on admission control over stateless networks (while we assume the reader to be familiar with stateful admission control, such as that defined in the framework of the Integrated Services paradigm). In Section III, we present the basic concepts of GRIP. Section IV is devote to the comparison of the efficiency achievable with an ideal stateful approach with that resulting from our stateless procedure. In Section V, we present numerical results. In Section VI, we draw our conclusions and directions for future work.
II. RELATED WORK

Recent literature (see [11] and therein contained references) has shown that an admission control function can be provided over stateless networks by means of the so-called Endpoint Admission Control (EAC). EAC builds upon the idea that admission control can be managed by pure end-to-end operation, involving only source and destination hosts. At connection set-up, each sender-receiver pair starts a probing phase whose goal is to determine whether the considered connection can be admitted to the network. In some EAC proposals ([11], [12], [13]), during the probing phase, the source node sends packets that reproduce the characteristics (or a subset of them) of the traffic that the source wants to emit through the network. Upon reception of the first probing packet, the destination host starts monitoring probing packets statistics (e.g., loss ratio, inter-arrival times of probing packets, etc.) for a given period of time. At the end of the measurement period and on the basis of suitable criteria, the receiver takes the decision whether to admit or reject the connection and notifies this decision to the source node.

Although the described scheme looks elegant and promising (it is scalable, it does not involve inner routers), a number of issues arise when we look for QoS performance. A scheme purely based on endpoint measurements suffers from performance drawbacks mostly related to the necessarily limited (few hundreds of ms, for reasonably bounded call setup times) measurement time spent at the destination. Measurements taken over such a short time, and on an end-to-end basis, cannot capture stationary network states, and thus the decision whether to admit or reject a call is taken over a snapshot of the network status, which can be quite an unrealistic picture of the network congestion level.

The simplest solution to the above issue is to attempt to convey more reliable network state information to the edge of the network. To this end, several solutions have been proposed in the literature. Paper [14] proposes to drive EAC decisions from measurements performed on a longer time scale among each ingress/egress pair of nodes within a domain. Papers [15], [16] use packet marking to convey explicit congestion information to the network nodes, which are in
charge of taking admission control decisions. Paper [17] performs admission control at layers above IP (i.e., TCP), by requiring each core router to parse and capture TCP SYN and SYN/ACK segments, and forward such packets only if local congestion conditions allow admission of a new TCP flow. Paper [18] proposes a lightweight signaling protocol, with explicit reservation messages, which requires network routers to actively manage packets (via remarking of signaling packets when congestion occurs), and thus it does not fit within a DiffServ framework, where the core routers duty is strictly limited to forwarding packets at the greatest possible speed (see e.g., what stated in [11]).

To summarize the above discussion, and to proceed further, we can state that an abstract and general EAC can be defined as the combination of three logically distinct components, even if in some specific solutions this is not apparent at a first glance:

- edge nodes in charge of taking explicit per flow accept/reject decisions;
- physical principles and measures on which such decisions are based;
- specific mechanisms adopted to convey internal network information to edge nodes.

In such a view, and with reference to each of the above points, we argue that:

- to allow edge nodes to take learned accept/reject decisions, the congestion status of the network can not be inferred only on an end-to-end basis; inner routers must be actively involved, but without adding functionality other than that of the DiffServ paradigm, in the basic IP forwarding scheme;
- inner routers can determine whether a new call can be locally admitted (i.e., as far as the local router is concerned) by means of suitable Measurement Based Admission Controls (MBAC) techniques [19]. These schemes do not exploit per-flow state information and related traffic specifications. Instead, they operate on the basis of per-node aggregate traffic measurements carried out at the packet level. Their robustness stays in the fact that, in suitable conditions (e.g., when the peak rates of the flows are small with respect to link capacities), they are barely sensitive to uncertainties on traffic profile parameters.
As a consequence, it seems that routers can independently carry out scalable estimations, as far as local decisions are concerned. We propose one such scheme in [3], [5];

- an important problem is then how to convey the status of inner routers (evaluated by means of aggregate measurements) to the end points so that the latter devices can take learned admission control decisions, without violating the DiffServ paradigm. For obvious reasons, we cannot use explicit per flow signaling. Similarly, we do not want to modify the basic router operation, by introducing packet marking schemes or forcing routers to parse and interpret higher layer information. What we want to do is to implicitly convey the status of core routers to the end points, so that the latter devices can take learned admission control decisions, by means of scalable, DiffServ compliant procedures.

Our admission control paradigm is briefly recalled in the next Section. GRIP is DiffServ compliant since all traffic is managed according to the DS Code Point field only.

III. THE GRIP PARADIGM

GRIP combines an admission control operation, driven by end-points, with run-time traffic measurements, performed within each router to detect congestion. Thus, GRIP is completely described by two logical components: i) GRIP end-nodes operation, and ii) GRIP router operation.

GRIP’s end nodes operation is extremely simple. When a user terminal requests a connection with a destination terminal, the source node transmits a probing packet. Meanwhile, the source node activates a probing phase timeout, lasting for a reasonable time. If no response is received from the destination node before the timeout expiration, the source node enforces rejection of the connection setup attempt. Otherwise, if a feedback packet is received in time, the connection is accepted, and control is given back to the user application, which starts a data phase, simply consisting in the transmission of information packets. In DiffServ, packets belonging to different “traffic classes” are tagged with different values of the DSCP field, in the IP packet header, as
stated in RFC 2475. In GRIP, we assume that packets belonging to a given class are further
differentiated between probing and information packets. Thus in our view, the DSCP field codes
both the class and the role of the packets. These different tags given to probing and information
packets allow internal routers to apply them different forwarding behaviors and enforce probing
packet dropping (and thus block the setup attempt) when congestion arises. The role of the
destination node simply consists in monitoring the incoming IP packets, intercepting the ones
labeled as probing, reading their source address, and, for each incoming probing packet, just
relaying with the transmission of a feedback packet, if the destination is willing to accept the set-
up request \(^1\).

As regards the GRIP router operation, shown in Fig. 1, packets incoming to a DifServ router
output port are dispatched to suitable queues according to their DSCP tag. Several GRIP
modules, devised to support different traffic classes, may co-exist within a router. Each of such
GRIP modules is in charge of handling both probing packets and information packets belonging
to a given traffic class. In Section IV, we propose an approach to handle heterogeneous traffic; in
this way, it is possible to limit the overall number of classes to be implemented. Information
packets are generated by traffic sources which have already passed an admission control test.
Within each GRIP module, a *measurement module* is in charge of measuring the load offered by
information packets, i.e., the overall aggregate accepted traffic. On the basis of these running
traffic measurements, and according to a suitable *Decision Criterion*, the measurement module
drives an *Accept/Reject switch*. When the switch is in the ACCEPT state, incoming probing
packets are forwarded to the output queue. Conversely, probing packets are dropped when the
switch is the REJECT state. In other words, the router acts as a gate for the probing flow, where
the gate is opened or closed on the basis of the traffic estimates (hence the Gauge&Gate in the
acronym GRIP).

---

\(^1\) For clarity of presentation, we identified the source and destination user terminals as the network end-nodes, taking admission control
decisions. However, for obvious security reasons, such end-nodes should be the ingress and egress nodes, under the control of the network
operator(s). If GRIP is seen as a domain specific mechanism, then such end-nodes are suitable nodes, located at the edge of the considered
domain.
This operation, in conjunction with the end-nodes operation described above, provides a per-flow admission control function over a stateless network, via an implicit signaling pipe of which the network remains unaware. In fact, in order for a flow setup procedure to succeed, the probing packet needs to find all the routers along the path in the ACCEPT state.

It is worth noting that the network operator has to properly define a Decision Criterion, to assure a desired performance. This can be made according to proprietary criteria, theoretical and/or off-line measurements procedures. In [3], [5] we defined a Decision Criterion, running on each router's output link, and based on the runtime estimation of the number of active flows, \( N \) (i.e., flows that have passed the admission control test and are currently emitting information).

In this scenario, target performance levels (e.g., loss/delay) can be hard-guaranteed by imposing that the maximum number of admitted flows does not overflow a suitable threshold \( K \). For our aims, \( K \) is a "tuning knob", which allows the domain operator to set target performance levels. The detailed relationship between \( K \) and such guaranteed performance levels results from an off-line computation (including evaluating \( K \) by means of trials in a specific network scenario).

In other words, we assume that the operator chooses target performance levels; the latter are mapped in a value of \( K \) by means of suitable procedures (the value of \( K \) is the acceptance threshold); then, the role of GRIP is to enforce such threshold, assuring that the number of active flows, \( N \), never exceeds \( K \). Note also that GRIP is completely independent by the specific algorithm chosen to evaluate \( K \).

In state-based admission control schemes, the enforcement of the threshold \( K \) is very simple: it is sufficient to keep track, by means of signaling exchanges, of the number of admitted connections \( N \), and accept a new connection setup request as long as \( N \leq K - 1 \). This way, the number of active flows can be as high as \( K \).

In our, stateless, procedure we estimate at runtime the number of flows, \( N \), avoiding signaling and state maintenance. This implies that the estimated value of \( N \), named \( N_{est} \), can be different
than $N$, due to possible measurement errors. In addition, since we want to guarantee performance (i.e., we want to strictly enforce the threshold $K$), to be on the safe side, we overestimate $N$. This means that, typically, $N_{ea} > N$. As a consequence, GRIP can accept a number of connections, $N_{acc}$, lower than $K$. In other words, GRIP has the advantage of avoiding explicit signaling and state maintenance. The price to pay is a potential system under-utilization with respect to the “ideal” value of $K$. The aim of this paper is to evaluate this loss in efficiency, i.e., the difference between $K$ and $N_{acc}$.

We stress that we are aware that GRIP can not be considered, up to now, a fully-fledged solution for the Internet nor a full protocol specification. Indeed, a number of detailed issues are not addressed in its definition, such as security, inter-working between different domains and policies, assumptions on traffic sources behavior, relationship between the proposed solution, which operates at the network layer, and upper layers, implementation details, etc. Rather, our goal is to present ideas which may contribute to the on-going discussions in the international arena, and may be integrated into the IETF vision.

In our opinion, GRIP could operate end-to-end over the whole Internet, but, more realistically, it could also be one of many possible different control mechanisms adopted to control traffic in different specific domains. In this vision, the degree of QoS support provided within each domain would depend on the tightness of control that the edge-to-edge mechanism will be capable to support. Schemes ranging from explicit per-flow resource reservation mechanisms (such as RSVP), down to aggregate forms of traffic control (e.g., via measurement based mechanisms, such as the one of GRIP) should be allowed to exist in different domains and interoperate, as suggested in [2]. The ultimate goal is that each domain should be placed in the ideal conditions of determining the suitable throughput/QoS support tradeoff within that domain. Note that, in this vision, GRIP could be an important building block to define and enforce a specific Per-Domain Behavior (PDB\textsuperscript{2}).

\textsuperscript{2} A PDB describes the performance experienced by a particular set of packets (that is those labeled with the same DS codepoint) as they cross a
IV. EFFICIENCY LOSS EVALUATION

A. Homogeneous Traffic Case

In defining the Admission Control (AC) criterion, our aim is to provide strict QoS guarantees in all operational conditions. To this purpose, information about traffic flows characteristics is needed, to correctly reach ACCEPT/REJECT decisions. Thus, we assume that traffic sources are regulated at network edges by standard Dual Leaky Buckets (see, e.g., [7]). We point out that the assumption of DLB regulated traffic sources is a common one in the literature and that without it we can not offer QoS guarantees. In addition, the DLBs are probably needed anyway, for policing and charging purpose, in advanced scenarios. This assumption implies that each source is characterized by its Traffic Descriptors, given in terms of three DLB parameters, namely Peak rate, $P_S$, (b/s), Sustainable Rate, $r_S$, (b/s), and Token Bucket Size, $B_{TS}$, (bytes). In addition, we specifically require the DLB to enforce that traffic does not underflow the sustainable rate specification [5]. All flows with the same DLB parameters are assumed as belonging to the same traffic class, thus they share a part of the DSCP tag (probing and information packets are further differentiated by a sub-field of the DSCP).

Thanks to this characterization, the GRIP modules running in the domain routers can estimate the traffic loading each port in terms of number of flows of a given service class. Let us consider the flows loading a generic port of a generic router and belonging to a given traffic class. GRIP estimates such number of flows as:

$$N_{est+STACK} = \left\lfloor N_{est} + STACK \right\rfloor = \left\lfloor \frac{A(T)}{r_S T - B_{TS}} + STACK \right\rfloor \quad (1)$$

where $A(T)$ is the number of bytes of the considered class transmitted by the given port and measured by means of a sliding window of size $T$. GRIP accepts a new flow setup request if
\(N_{est+STACK}\) is strictly less than a suitable number, \(K\), defined above as the threshold not to be overcome if suitable levels of QoS have to be provided. The number of flows estimated by (1) is an overestimate of the number of flows actually loading the considered port, \(N_{real}\). \(STACK\) is a scalar variable used to protect the system against transient loads [3].

Let us denote \(\lambda\) the mean flow arrival rate, \(\mu\) the mean departure rate and \(T_{OFF}\) the ratio \(B_{TS}/r_s\). The relationship between the average values of \(N_{est+STACK}\) and \(N_{real}\) (in the following \(\overline{N}_{est+STACK}\) and \(\overline{N}_{real}\), respectively) depends on the offered load. To this end, we assume that the contribution of each flow to the \(STACK\) be a random variable uniformly distributed in \([0,1]\), and thus its mean value is \(\frac{1}{2}\). In this way, we obtain:

1. when \(N_{est+STACK} < K\), assuming that all offered flows are accepted in the system, then \(N_{real}\) is equal to \(\lambda/\mu\). The mean \(STACK\) value is determined by the frequency of arrivals in the measurement window, and we get:

\[
\overline{N}_{est+STACK} = \frac{(\lambda / \mu) r_s T + \lambda T}{2} = \overline{N}_{real} \left( \frac{T}{T - T_{OFF}} + \frac{\mu T}{2} \right)
\]

2. when \(N_{est+STACK} \geq K\), (the system is estimated as full):

\[
\overline{N}_{est+STACK} = \overline{N}_{real} \frac{T}{r_s T - B_{TS}} + \frac{\mu T \overline{N}_{real}}{2} = \overline{N}_{real} \left( \frac{T}{T - T_{OFF}} + \frac{\mu T}{2} \right)
\]

In fact, in this case, a new flow can be admitted only if another one has just left the system, so the mean value of the \(STACK\) variable is determined by the overall mean number of departures, \(\overline{N}_{real} \mu T\), multiplied by \(\frac{1}{2}\).

Both equations are indicated as approximated due to rounding operations performed to take into account the fact that the values of flow numbers have to be integer values. In addition, (3) is

---

1 If a large number of new flows are offered in a short time frame, the measurement mechanism is not sufficient to protect the system from over-allocation, i.e., all the newly incoming flows may be accepted. In fact, when a router "accepts" the probing packet of a given flow (i.e., the probing packet finds the gate opened), the relevant information packets are not yet emitted by the source. In other words, it exists a transient time during which the router is loaded with a new flow, but \(A(T)\) only partially accounts for the new traffic contribution. To overcome this problem, and to provide strict guarantees in any operational condition, we introduced a protection mechanism based on a "stack" variable, which keeps memory of the amount of "transient" flows. Whenever a probing packet is accepted, the stack is incremented by one. A timer equal to the duration of the measurement window \(T\) is then started, and the stack is linearly decremented at a rate \(1/T\) until the timer expires.
evaluated thanks to two simplifying hypotheses: i) it does not take account flows whose overall duration is less than the measurement window, $T$; ii) it assumes so-called impulse load conditions, i.e., new flows enter the system as soon as other flows leave the system itself. This implies an overestimate of the number of flows in the system and thus an overestimate of the utilization factor of GRIP. We will remove this assumption in the following.

By looking at (2) and (3), it is easy to see that the ratio between the estimated and the real number of flows loading a given port, $R_{over}$, is given by (independently by the load):

$$ R_{over} = \frac{N_{est+STACK}}{N_{real}} \approx \left( \frac{T}{T - T_{off}} + \frac{\mu T}{2} \right), \quad (4) $$

while the maximum utilization coefficient that can be reached by using GRIP is:

$$ \rho_{\text{GRIP}} = \frac{\overline{N_{real}r_s}}{C} = \begin{cases} (\lambda / \mu) r_s / C, & \lambda / \mu < K / R_{over} \\ (K / R_{over}) r_s / C, & \lambda / \mu \geq K / R_{over} \end{cases} \quad (5) $$

When the number of estimated flows is less than $K$, GRIP does not cause an efficiency loss, since it accepts all offered flows. Instead, when the contrary happens (i.e., when $N_{est+STACK} \geq K$), the system is considered full by the GRIP admission control logic, even if $N_{real}$ can be less than $K$. Thus, when GRIP considers the system in overload, that is for $\lambda / \mu \geq K / R_{over}$, the reciprocal of $R_{over}$, denoted as $\varepsilon = 1 / R_{over}$, can be seen as the efficiency of a stateless procedure, able to allocate $\overline{N_{real}}$ flows, normalized with respect to a stateful one, able to allocate $K$ flows. In other words, $\varepsilon \leq 1$; besides, $\varepsilon = 1$ when the stateless AC procedure has the same utilization coefficient and allocates the same number of flows of the stateful AC procedure.

The value of $\varepsilon$ is, in general, a function of the GRIP measurement window size, $T$. When the system is in overload, according to GRIP, i.e., for $\lambda / \mu \geq K \varepsilon$, it is possible to evaluate the maximum value of $\varepsilon$, which can be defined as $\varepsilon_{\text{max}} = \max(\rho_{\text{GRIP}} / \rho_{\text{stateful}}$, with $\rho_{\text{stateful}} = Kr_s / C$ being the utilization coefficient of an ideal stateful admission control function, able to allocate up to $K$ flows. To evaluate the maximum of $\rho_{\text{GRIP}}$, we observe that it exists a value of the measurement window size, $T_{opt}$, which maximizes the GRIP efficiency and thus $\varepsilon$. This value is given by
\[ T_{\text{opt}} = T_{\text{OFF}} + \sqrt{2T_{\text{OFF}}/\mu} \]  

[3]. As a result, we have

\[
\varepsilon_{\text{max}} = \frac{\max(\rho_{\text{GRIP}})}{\rho_{\text{stateful}}} = \frac{\rho_{\text{stateless}}}{\rho_{\text{stateful}}} = \left(1 + \sqrt{\frac{\mu T_{\text{OFF}}}{2}}\right)^{-2}.
\]  

(6)

The corresponding number of flows accepted by GRIP is given by \( K_{\text{GRIP}} = K\varepsilon \) and its maximum value is \( K_{\text{GRIP,max}} = K\varepsilon_{\text{max}} \).

We stress that the maximum value of the ratio between the utilization of a stateless approach and the utilization of a stateful approach, \( \varepsilon_{\text{max}} \), given by (6), depends only on two parameters: the mean flow duration \((1/\mu)\) and the maximum silence period of the DLB, \( T_{\text{OFF}} \).

Parameters related to capacity and buffer of network nodes are not of importance. Thus (6) can be used to evaluate \textit{a priori} the under-utilization resulting by the adoption of our stateless AC as compared to a stateful approach, in a whole domain. This comparison is made under the constraint that each node of the domain provides the same \textit{minimum} performance levels to accepted flows, both in the stateless and in the stateful case, since the threshold \( K \) is the same.

Equation (6), which evaluates \( \varepsilon_{\text{max}} \), is very useful, since it depends on two parameters only. However, (6) is an overestimate of the real value, since \( \rho_{\text{GRIP}} \) has been evaluated with a simplifying hypothesis (see (3) and the following comment). When the traffic offered to the system is much greater than the system capacity, (i.e., \( \lambda/\mu \gg K \)), this approximation is reasonable. Instead, when \( \lambda/\mu \) is near to \( K_{\text{GRIP}} \) the impulse load condition is less verified.

Thus, we now remove this assumption and derive a very simple model, which allows evaluating the GRIP utilization coefficient as a function of the offered traffic load.

When we apply the GRIP AC to a given port, that port can support, \textit{in average}, a maximum number of flows equal to \( K_{\text{GRIP}} \). Thus, at call level, we can easily model this port with a queuing system with \( K_{\text{GRIP}} \) servers and no waiting room. For the sake of extreme simplicity, we assume a Poisson offered load and exponential call durations and we end up with a \( M/M/K_{\text{GRIP}} \) system.

In this way, defined \( A_0 = \lambda/\mu \), the number of flows really loading the considered port can be
better estimated as:

$$N_{\text{real,refin}} = A_0(1 - B(K_{\text{GRIP}} \cdot A_0))$$

(7)

where $B(K_{\text{GRIP}} \cdot A_0)$ is the classical Erlang’s B formula. As a consequence, we get:

$$\rho_{\text{GRIP,refin}} = \frac{N_{\text{real,refin}}r_s}{C}$$

(8)

and

$$\varepsilon_{\text{refin}} = \frac{\rho_{\text{GRIP,refin}}}{\rho_{\text{stateful}}} = \frac{A_0}{K} \left[1 - B \left(\frac{K(1-T_{\text{OFF}}/T)}{1 + \mu(T - T_{\text{OFF}})/2}, A_0\right)\right]$$

(9)

If, as above, we choose the size of the measurement window, $T$, so as to maximizes the GRIP efficiency, setting $T=T_{\text{opt}}$, we get:

$$\varepsilon_{\text{refin,max}} = \frac{A_0}{K} \left(1 - B \left(\frac{K}{1 + \frac{\mu T_{\text{OFF}}}{2}}, A_0\right)\right)$$

(10)

This refined evaluation of the efficiency of a stateless procedure normalized with respect to a stateful one, $\varepsilon_{\text{refin}}$, depends on the traffic descriptors, on the mean flow duration, and also on the overall traffic load and on nodes and performance parameters. In the following Section, we will show, by means of numerical examples, that this refined model is very close to simulation results, and that even the simple and basic evaluation of $\varepsilon$ presented above is still a good approximation.

B. Heterogeneous traffic case

To extend GRIP to a heterogeneous traffic scenario, we need some introductory considerations. Let us assume that the sources are divided in $I$ traffic classes, each comprising independent and homogeneous sources (i.e., with the same DLB parameters). To go further, we note that, in the homogeneous case, the evaluation of the number of admissible flows $K$ is equivalent to the evaluation of the so-called effective bandwidth of each flow, denoted as $e$. The effective bandwidth concept is well known in the literature and represents the amount of bandwidth that must be assigned to each flow, in a statistical multiplexing framework, so as to
reach some performance levels. The effective bandwidth is typically comprised between \( r_s \) and \( P_s \). In our case, \( e = C/K \), where \( C \) is the link capacity (and where \( K \) is evaluated as in [7], once the buffer size \( B \) and the performance parameters of interest, e.g., \( P_{\text{loss}} \), have been fixed). The ideal admission control function “accept new set-up requests as long as the number of admitted flows \( N \) is less than \( K \)”, can then be rewritten as “accept new set-up request as long as the sum of the effective bandwidth of accepted flows \( (N \cdot e) \) is less than \( C \)”. In GRIP, due to the lack of signaling, we do not know \( N \), so we estimate it; then we implicitly use the effective bandwidth concept by comparing the estimated \( N \) to \( K \). Finally, we add the stack mechanism. In other words GRIP exploits the DLB characterization two times. The first time to estimate the number of admitted flows and the second time to decide if a new set-up request can be accepted. Note that each router must be implicitly aware of the DLB parameters.

The effective bandwidth concept can be in principle easily extended to the heterogeneous case. In [7], the Authors propose an efficient algorithm to evaluate this quantity, which, in the assumption of DLB regulated sources, is additive even in the heterogeneous case (see the discussion in [21] about heterogeneous traffic handling). The equivalent bandwidth \( e_i \) of the \( i \)-th traffic class, \( \{1 \leq i \leq I\} \) is evaluated as \( e_i = C/K_i \), where \( K_i \) is evaluated for each class in isolation, i.e., in a homogeneous system. In other words, the limit value \( K_i \) is the number of flows such that a link with capacity \( C \) and buffer \( B \), loaded only with traffic belonging to the \( i \)-th class, offers pre-defined performance levels. With this approach (which is shown to be conservative, even if it leads to a loss of efficiency), the ideal admission rule remains a simple sum: new set-up requests are accepted as long as

\[
\sum_{i=1}^{I} N_i e_i \leq C
\]  

(11)

where \( N_i \) is the number of admitted flows belonging to class \( i \).

To extend GRIP to handle heterogeneous traffic, we have to estimate the number of admitted flows of each class and then apply the above admission rule. To this end, we can identify two
architectural alternatives. The former (trivial) consists of separating the overall available capacity in different portions, and treat each traffic class in isolation. Here, we assume that each class is handled in a separated way and recognized by routers by assigning different pairs of DS codepoints to different traffic classes (i.e., each class has a DSCP for probing packets and another for information packets). Thus, we need $2\cdot l$ different DCSPs and $l$ different logical queues. Packets belonging to class $i$ are classified on the basis of their DSCP tag and dispatched to the relevant module (Measurer or Dropper). However, we do not require to signal explicit information about the traffic mix composition, that is how many active flows per each class are present in each node.

The architecture is complex but the extension of GRIP is straightforward, since the heterogeneous case AC is reduced to a combination of homogeneous ones:

$$\left| \frac{A_i(T)}{r_{s,i}T-B_{TS,i}} + \text{STACK}_{i} \right| \leq K_i - 1, \text{ with } \sum_{i=1}^{l} K_i e_i \leq C, \quad (12)$$

In (12), $A_i(T)$ is the number of bytes emitted during the window $T$ by traffic sources belonging to the $i$-th class (i.e., at the $i$-th queue) and $K_i$ is evaluated off-line as discussed above. Note that this alternative implies that each node is aware of the DLB parameters of all possible classes. This architectural alternative is somehow acceptable only in presence of a small number of traffic classes (e.g., a class could be IP telephony), even if it is still compliant with the DiffServ approach. An advantage of this architecture is that it allows implementing procedures to fairly divide the overall capacity among the traffic classes.

In the second alternative, partially described also in [22], we assume that both probing and information packets belonging to different classes are handled in a joined way, with only one measure module and one dropper module, exactly as in the homogeneous case. All the traffic classes share the same pair of DSCP tag: one for probing packets and one for information packets. Thus we need only two DSCPs. As in the previous ones, we do not require to signal explicit information about the traffic mix composition. However, the Decision Criterion has to be
suitable modified with respect to the homogeneous case, in order to take into account the
presence of different traffic profiles, while still guaranteeing performance. To define a suitable
allocation rule, able to assure to not exceed to allocate more than $C$ effective bandwidth, we have
to evaluate the specific traffic mix, compliant with the measurement output made in the router,
which maximizes the overall effective bandwidth allocated:

$$e_{TOT} \sum_{i=1}^{I} N_i e_i$$

(13)

under the constraints expressed by the emitted bits in the sliding window:

$$\left\{ \sum_{i=1}^{I} N_i a_{i}^{\min} \leq A(T) \text{ with } a_{i}^{\min} = r_{s,i} T_B T_s,i \right\}$$

(14)

$$0 \leq N_i \leq K_i$$

Then, the solution of (13) has to be compared with the maximum admissible value $C$, in order
to identify a suitable rule able to make the system to respect the constraint in (11).

In other words, we have to find the values of $N_i \ (i \in [1, I])$ in the range $\left( N_i \in [0, K_i] \right)$ under the
constraint that the admitted flows must be able to emit $A(T)$ bytes during $T$ seconds. In addition, since we are dealing with the worst case, we require also that each flow is emitting according to
its minimum DLB emission profile (this is the meaning of the apex “$min$” to quantity “$a$”).

The main difficulty in resolving the problem in (13) is the need to find an integer solution,
since we are dealing with numbers of flows. We start by resolving the associated continuous
problem (i.e. by assuming that $N_i$ is a continuous variable) and then we apply a floor operator to
the solution (which is a slightly conservative operation), obtaining:

$$\begin{align*}
N_x &= \left\lfloor \frac{A}{a_x^{\min}} \right\rfloor \\
N_i &= 0 \quad \text{for } i \neq x
\end{align*}$$

(15)

where “$x$” is the index of the class with the greatest value of the ratio $e_i / a_i^{\min} \ (i \in [1, I])$. This
means that, due to the lack of any information about the composition of the traffic mix, the
measurement procedure reduces the heterogeneous case to an homogenous one, in which, for
estimation purposes, only the traffic class endowed with the worst estimation of resource
utilization is considered as present in the mix. Such class is labeled with the index “x”. The allocation rule relevant to all the traffic handled in the queue becomes (see (11): gate opened):
\[
\frac{A(T)}{T - r_{\text{S},x}} + \text{STACK} + e_{\text{max}} \leq K_x,
\]
where the STACK contribution is increased, for each probing packet forwarded, by \( \gamma = e_{\text{max}} / e_x \).

In fact, since the overall process is oblivious of the consumption of effective bandwidth by the newly admitted flow, we are forced to interpret each set-up request as belonging to the class with the greatest equivalent bandwidth, if we want to provide strict QoS guarantees in every operational condition.

In conclusion, this approach implies that each node has to know the DLB parameters of only one traffic class (the “x” one) and the value \( \gamma \).

The prices to pay are: i) a potential smaller system efficiency with respect to the previous cases; ii) the impossibility of implementing procedures to fairly divide the overall capacity among the traffic classes.

Let us analyze of the heterogeneous traffic case in order to evaluate the maximum utilization coefficient (condition of system overload. We say that the system is in overload condition when the total offered load, in terms of effective bandwidth, is higher than the system capacity \( C \). We make the hypothesis that the different traffic profile have the same process describing the call duration, described by its mean value \( 1/\mu \). This assumption is not so unrealistic: e.g., a number of different traffic profiles could be relevant to the same type of source (e.g., voice or video) encoded and/or policed in different ways due to, e.g., different charging profiles. However, we will remove this simplifying assumption in a further analysis. Even if the following analysis has been carried out with only two traffic classes for the sake of simplicity, it is general. We defined a parameter, \( \alpha \), in order to identify the composition of the mix. It represents the percentage of the overall utilization coefficient \( \rho \) provided by the flows of Class \( x \) using GRIP, i.e.,
\[ \alpha = \frac{r_{i,x} N_x}{\sum_{i=1}^{l} r_{i,y} N_i} \quad (17) \]

This implies that, in the two classes case, the ratio between \( N_{x,r_{S,x}} \) and the other (i.e., \( N_{y,r_{S,y}} \)) is fixed to \( \alpha (1 - \alpha) \). Following the approach developed in preceding section, it results that, in overload condition, the following results applies in average:

\[ \frac{N_{x,r_{S,x}} T + N_{y,r_{S,y}} T}{r_{S,x} T - B_{TS,x}} + \mu T \gamma \left( \frac{N_x + N_y r_{S,y}}{r_{S,x}} \right) \approx K_x. \quad (18) \]

In (18), the term representing the STACK contribution takes into account the fact that the traffic class “\( y \)” has to be scaled to “\( x \)” by means of the ratio of their sustainable rates, in order to be comparable to flows belonging to class “\( x \)”. By using (17) and (18), the overall utilization coefficient, independently of the composition of the traffic mix, can be written as (being \( T_{OFF,x} = B_{TS,x} / r_{S,x} \)):

\[ \rho_{TOT,GRIP} = \frac{1}{\mu} \sum_{i=1}^{l} \frac{N_i r_{S,i}}{C} = \frac{K_x r_{S,x}}{C} \left( \frac{T}{T - T_{OFF,x}} + \frac{\gamma T}{2} \right)^{-1} \quad (19) \]

As in [3], we can optimize the utilization coefficient with respect to the measurement window \( T \) by means of derivative, obtaining the expression \( T_{opt} = T_{OFF,x} + \sqrt{2 \rho_{OFF,x} / \mu} \). In this case, the maximum value (upper bound) of the utilization coefficient in the heterogeneous case is equal to:

\[ \rho_{TOT,GRIP,max} = \frac{K_x r_{S,x}}{C} \left( 1 + \sqrt{\frac{\rho_{OFF,x}}{2}} \right)^{-2} \quad (20) \]

It is interesting to note that the upper bound to the utilization coefficient reachable using the heterogeneous extension of GRIP is represented by the upper bound of the utilization coefficient for the class “\( x \)” in isolation, decreased only by the term \( \gamma \) which takes into account the (eventual) presence of traffic profiles with an effective bandwidth higher than that of class “\( x \)”. This result is reasonable, since, being the traffic aggregate estimated with the parameters of only
one class, the overall utilization coefficient is limited by the one of such class.

V. NUMERICAL RESULTS

GRIP does not need assumptions on the traffic source behavior, beyond the DLB characterization. However, to evaluate performance and generate traffic for a simulation study, we have to load the DLBs with specific sources. The choice of model parameters has been made in order to perform an analysis able to highlight technical features of the developed approach, without specific engineering purposes.

We loaded the DLB devices with on-off exponential voice sources, representative of classical voice encoding based on silence detection. The on and off state durations are exponentially distributed, with average values of 352 ms and 650 ms respectively; during the On state (talkspurt) the source emits packets periodically, with a bit rate equal to 32 kb/s. The DLB parameters are: \( P_S = 32 \text{ kb/s} \); \( r_S = 13.6 \text{ kb/s} \); \( B_T = 5300 \text{ bytes} \). We consider a generic router’s output link (with a FIFO queue) with parameters: link rate, \( C = 2048 \text{ kb/s} \); buffer size, \( B = 53000 \text{ bytes} \). We set, as target performance figure, a packet loss probability, \( P_{\text{loss}} \), equal to \( 10^{-5} \). According to the acceptance rule provided in [7], the corresponding maximum number of acceptable flows is \( K = 100 \). Thus both GRIP and an ideal stateful Admission Control must guarantee the enforcement of this threshold.

The call arrival rate, modeled as a Poisson arrival process, has been set in the range \( 0 \div 3 \) calls/s, while the call duration has been drawn from an exponential distribution with mean value 4 minutes, so the offered load is varying between 0 and 720 erls. The measurement window size has been set to \( T = 30 \) seconds (the optimal window size, \( T_{\text{opt}} \), which maximizes the GRIP efficiency is equal to 41.2 seconds, in this scenario).

Fig. 2 reports the utilization coefficient of the considered router port versus the offered load.

The curve labeled \( \rho_{\text{stateful}} \) refers to an ideal stateful system; the one labeled \( \rho_{\text{GRIP}} \) refers to the quantity evaluated in (5), while the curve labeled \( \rho_{\text{GRIP.refin}} \) is relevant to the refined evaluation in (8). The circles indicate simulations results (the relevant 95% confidence intervals are always
less than 5% and are not shown to improve the neatness of the figure). This figure shows that the refined model is really close to simulation results and that even the simple model depending only on the mean flow duration and on the maximum silence period of the DLB is close to simulations results. Fig. 3 reports the GRIP utilization coefficient, evaluated with the basic and the refined model as a function of the traffic load for three different values of the threshold \( K \) (100, 200 and 300). The corresponding values of \( C \) are 2048, 3686.4 and 5263.36 kb/s respectively, limiting the maximum queuing delay to a target value of 200 ms. The values of \( K \) are such that the target packet loss probability, \( P_{\text{loss}} \), is equal to \( 10^{-5} \). The measurement window size, \( T \), is equal to 30 seconds, as in Fig. 2. It is worth noting that, for increasing values of \( K \), the utilization coefficient increases, due to the effect of the statistical multiplexing gain, while the gap between our two models is almost the same. This hints that our models scale well with system size.

To better explore this effect, Fig. 4 shows the ratio between the utilization coefficient estimated with the refined model and the one of the basic model (\( \rho_{\text{GRIP,refin}} / \rho_{\text{GRIP}} \)) as a function of the ratio \( A_0/K_{\text{GRIP}} \), for three different values of \( K \) (100, 200, 300). This figure highlights the accuracy in the evaluation of \( \rho_{\text{GRIP}} \) when \( A_0=K_{\text{GRIP}} \), which is a critical point, since the maximum difference between these two approaches occurs exactly in this point. Note that, thanks to the multiplexing gain, the difference between the two models decreases when \( K \) increases. Finally, Fig. 5 depicts the maximum value of the efficiency of a stateless procedure normalized with respect to a stateful one, \( \varepsilon_{\text{max}} \), evaluated with the basic model in (6) as a function of the mean flow duration (\( 1/\mu \), in minutes) and of the maximum silence period of the DLB, \( T_{\text{OFF}} \) (in seconds). It is interesting to note that stateful and stateless systems tend to have the same utilization factor for long calls and for sources with small silence periods.

In order to test the effectiveness of the heterogeneous approach, we carried out simulations with two traffic classes.

We used on-off exponential sources with different DLB parameters, to represent two different
traffic classes. For the first kind of sources (labeled as Class 1), we used the one of the homogenous simulations. For the second kind of sources (labeled as Class 2), the on and off state durations are exponentially distributed, with average values of 400 ms and 1000 ms respectively. The bit rate during the on period has been set equal to 48 kb/s. We consider a generic router output link with parameters: link rate, \( C = 2.048 \text{ Mb/s} \); buffer size, \( B = 53000 \text{ bytes} \). We set, as target performance figure, a packet loss probability, \( P_{\text{loss}} \), equal to \( 10^{-5} \). Table 1 reports the DLB parameters of the sources and the maximum number of acceptable flows \( K_i \), according to the acceptance rules provided in [7]. The call duration for both the sources has been drawn from an exponential distribution with mean value 4 minutes. In order to stress the heterogeneous algorithm under different traffic conditions, we considered three different traffic mixes described by \( \alpha \), set to 0.25, 0.50, 0.75.

For each value of the parameter \( \alpha \), we evaluated the target numbers of flows for both the two classes as in [7]. Such values are reported in Table 2, with the values of utilization coefficient relevant to the reference system in heterogeneous case, indicated as \( \rho_{\text{TOT,REF}} \).

The arrival rate, modeled as a Poisson arrival process, has been set to different values, depending of the composition of the traffic mix. These values have been chosen in order to allow a slight system overload (110%) with respect to the target values in Table 2. It is simple to show that, in order to match the value of \( \alpha \), it is enough to set the ratio between the arrival rates \( \lambda_i \) (\( i = 1,2 \)) equal to the ratio of the value of \( K_i \) reported in Table 2.

In Fig. 6, we show simulation results relevant to the heterogeneous case, considering both the approaches developed in this paper. On the abscissa we report the measurement window size \( T \), while on the ordinate the value of the utilization coefficient, \( \rho_{\text{TOT,GRIP}} \). In the figure, the curve labeled Upper Bound is evaluated by means of (19), while the curve labeled Refined Estimation is analogous to (8) and relevant to a \( M/M/m \) queueing model, where \( m \) is given by (19).

As expected from theoretical results, the utilization coefficient reached when using only one queue is nearly always the same, independently of the traffic mix composition, while it is
different in the case of separated queue. In fact, in this second case, when the class with the highest value of the utilization coefficient is dominant, the overall utilization benefits from this condition. The overall utilization coefficient for Reference system result always higher than the one provided by GRIP algorithm (i.e., $\rho_{TOT,GRIP} < \rho_{TOT,REF}$). However, since our algorithm, also in the heterogeneous version, provides strict QoS guarantees, we expected such behavior, as in the homogeneous case.

For our choice of the DLB parameters (a variation of the 50% for the peak rate, a variation of the 20% for the sustainable rate and for the token burst size), the effectiveness of the approach with a unique queue for all the traffic classes is always higher than the approach with separated queues. This phenomenon is explainable in terms of statistical multiplexing gain, and is a good reason to move towards a solution simpler and more effective. However, this choice, even if simpler in terms of architectural design, could not be always the best in terms of utilization coefficient. In particular, when the DLB parameters of the sources loading the queues are very different, the solution with separated traffic and probe queues could be more efficient, even if it requires an higher complexity.

In order to explain better this concept, we developed the following analysis. Let us consider the possibility of varying with continuity the parameters of the regulator in a given range, e.g., $r_S \in [r_{S,\min}, r_{S,\max}]$, $P_S \in [P_{S,\min}, P_{S,\max}]$ and $B_{TS} \in [B_{TS,\min}, B_{TS,\max}]$. In this 3-dimensional space, we want to find out the coordinates of the point able to perform the estimation according to the criteria developed before, i.e., find the parameters of the class “x”. The solution will be the triple of DLB parameters $(r_{S,x}, P_{S,x}, B_{TS,x})$ able to maximize the ratio $e / a_{\min} = e / (r_S T - B_{TS})$. It is clear that, the higher the peak rate, the higher the effective bandwidth $e$, while $a_{\min}$ remains unchanged, so $P_{S,x} = P_{S,\max}$. For what concerning the burst tolerance, the higher $B_{TS}$, the higher $e$, while the lower the denominator, so it follows that $B_{TS,x} = B_{TS,\max}$. For what concerning the sustainable rate, both numerator and denominator are increasing with $r_S$. By taking the derivative
of the function to maximize with respect to the sustainable rate, it is quite straightforward to show that \( \frac{\partial e}{\partial r_s} \) (obtained by means of using implicit derivative of the non linear equation used in [7] to evaluate \( e \)) exists and is a continuous function. Since it is evaluated in a closed interval \([r_{S,\text{min}}, r_{S,\text{max}}]\), it has a finite maximum, i.e., its increasing rate is limited. Our conjecture is that, in general, the rate with which the numerator increases with \( r_s \) is inferior than that of the denominator, so that the solution to our mono-dimensional in \( r_s \) problem is given by \( r_{s,\ast} = r_{s,\text{min}} \).

It can be shown that this statement is equivalent to show that the following inequality is valid for each value of \( r_s \in [r_{s,\text{min}}, r_{s,\text{max}}] \):

\[
\frac{a_1(a_1 + a_3) + a_4}{a_5a_3} < 1,
\]

where the coefficients \( a_i \) are given by:

\[
\begin{align*}
a_1 &= eBr_s (P_s - r_s)(B_{r_s} C - Br_s) \\
a_2 &= \log(e / r_s) \\
a_3 &= \log \left( \frac{(P_s - r_s)(B_{r_s} C - Br_s)}{P_s(B_{r_s} C - Be) - e(B_{r_s} C - Br_s)} \right) \\
a_4 &= (e - r_s)P_sCB_{r_s} \left( (B_{r_s} C - Br_s) + B(P_s - r_s) \right) \\
a_5 &= P_s CB_{r_s} (P_s - r_s)(B_{r_s} C - Br_s)
\end{align*}
\]

(22)

Numerical search with different parameters have always confirmed this conjecture. In conclusion, it results \( \bar{x} = (r_{s,\text{min}}, P_{s,\text{max}}, B_{r_{s,\text{max}}}) \). Now, we need to move further, to find out the region in our domain where it is convenient, under the metric of the utilization coefficient, to use the approach with one queue in handling heterogeneous traffic. Before proceeding, we need to identify a description of the traffic mix: we chose to use always a two class description, in order to highlight more clearly the behavior of the system. Since the estimation point is fixed, when we consider another point in the domain, we evaluate the average utilization coefficient employing the approach with separated queues as follows:

\[
\rho_{SEP} = \int_0^1 \rho(\alpha)f(\alpha)d\alpha,
\]

(14)
where \( \rho(\alpha) \) represents the overall utilization coefficient obtainable when the ratio \( \alpha \) is fixed and \( f(\alpha) \) is the probability density function of \( \alpha \). For the case with one queue, we used the expression of \( \rho_{TOT,GRIP} \) given by (18). We chose the following search domain: \( r_s = 12-25 \text{ kb/s}, P_s = 32-64 \text{ kb/s} \) and \( B_{TS} = 3125-6250 \text{ bytes} \). In Fig. 7 we provide an example of \( \rho(\alpha) \), using the points with coordinates \( \vec{x} = (12 \text{ kb/s}, 64 \text{ kb/s}, 6250 \text{ bytes}) \) and \( \vec{y} = (15 \text{ kb/s}, 48 \text{ kb/s}, 4500 \text{ bytes}) \). Then, we proceeded to find out the region where \( \rho_{TOT,GRIP} \geq \rho_{SEP} \). Fig. 8 shows the result of this iterative process when a uniform distribution for \( \alpha \) is employed. The region of the domain where it is convenient to use a single queue approach is the one inside the surface. The estimation point \( \vec{x} \) has coordinates \( 12 \text{ kb/s}, 64 \text{ kb/s}, 6250 \text{ bytes} \). The shape is approximately triangular for each side of the region, so it should be easy to provide an analytical approximation. It is worth noting that assuming for \( \alpha \) a uniform distribution means that each combination of the mix is equally probable. However, in a more realistic scenario, the extreme situation in which one traffic profile is dominant on the others in terms of consumption of shared resource should not very likely, while it is more likely to have comparable weight in the system in terms of utilization coefficient from different traffic profiles. To model this different situation, we used a Bernoulli distribution for \( \alpha \) with parameter 0.5, so obtaining Fig. 9. Finally, in Fig. 10 we provide another example using the Beta distribution with both the parameters fixed to 2. In these two last cases, it is interesting to note that the region in which it is convenient to handle different traffic profiles with a unique queue is wider than uniform distribution case. In particular, in the case of uniform distribution for \( \alpha \), it results that the developed approach with one queue is convenient in a region covering the 37.4% of the domain, while choosing a Bernoulli or Beta distribution in the 70.4% and 59.9%, respectively.

VI. CONCLUSIONS

In this paper, we described an approach to provide strict QoS guarantees to heterogeneous traffic over DiffServ. The approach results in line with the basic concepts of DiffServ, it is
highly scalable, stateless and able to provide good performance when the differences in traffic profiles are not too emphasized. We developed a theoretical analysis of the performance of the proposed algorithm, highlighting pros and cons with respect to a stateful one. Our analysis allows the traffic engineer to make proper choices in term of QoS guarantees and utilization coefficient.

REFERENCES

Fig. 1: GRIP router operation.

Fig. 2: Utilization coefficient versus offered load: models and simulations.
Fig. 3: Utilization coefficient versus offered load: different values of $K$.

Fig. 4: Scaling factor between the two theoretical models ($\rho_{GRIP,ref}/\rho_{GRIP}$) as a function of the traffic load normalized with respect to $K_{GRIP}$. 
Fig. 5: $\epsilon_{\text{max}}$ as a function of the mean flow holding time ($1/\mu$) and of the maximum DLB silence period ($T_{\text{OFF}}$).

Fig. 6. Utilization coefficient ($\rho_{\text{TOT,GRIP}}$) versus the Measurement Window Size ($T$), for different values of the parameter $\alpha$. 
Fig. 7. Utilization coefficient $\rho(\alpha)$ for the case with separated queue with $\overline{X} = (12 \text{ kb/s, 64 kb/s, 6250 bytes})$

and $\overline{Y} = (15 \text{ kb/s, 48 kb/s, 4500 bytes})$.

Fig. 8. Region in the DLB parameters-domain where it is convenient to use the single queue approach to handle heterogeneous traffic with $\alpha$ uniform.
Fig. 9. Region in the DLB parameters-domain where it is convenient to use the single queue approach to handle heterogeneous traffic with Bernoulli distribution for $\alpha$ with parameter 0.5.

Fig. 10. Region in the DLB parameters-domain where it is convenient to use the single queue approach to handle heterogeneous traffic with Beta distribution for $\alpha$ with both parameters equal to 2.
<table>
<thead>
<tr>
<th></th>
<th>$P_s$</th>
<th>$r_s$</th>
<th>$B_{TS}$</th>
<th>$K$ (homogeneous)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Class 1</strong></td>
<td>32 kb/s</td>
<td>13.6 kb/s</td>
<td>5300 bytes</td>
<td>100</td>
</tr>
<tr>
<td><strong>Class 2 (x)</strong></td>
<td>48 kb/s</td>
<td>16.4 kb/s</td>
<td>6360 bytes</td>
<td>76</td>
</tr>
</tbody>
</table>

Table 1: DLB parameters of different traffic classes.

<table>
<thead>
<tr>
<th></th>
<th>$x=0.75$</th>
<th>$x=0.50$</th>
<th>$x=0.25$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$K_1$</td>
<td>23</td>
<td>48</td>
<td>74</td>
</tr>
<tr>
<td>$K_2$</td>
<td>58</td>
<td>40</td>
<td>20</td>
</tr>
<tr>
<td>$\lambda_1$</td>
<td>0.105 s$^{-1}$</td>
<td>0.216 s$^{-1}$</td>
<td>0.339 s$^{-1}$</td>
</tr>
<tr>
<td>$\lambda_2$</td>
<td>0.266 s$^{-1}$</td>
<td>0.18 s$^{-1}$</td>
<td>0.0917 s$^{-1}$</td>
</tr>
<tr>
<td>$\rho_{TOT,REF}$</td>
<td>0.617</td>
<td>0.639</td>
<td>0.652</td>
</tr>
</tbody>
</table>

Table 2: Configuration parameters for different values of the traffic mix.