PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

An approach to service level agreements for IP networks with differentiated services

R. J. Gibbens, S. K. Sargood, F. P. Kelly, H. Azmoodeh, R. Macfadyen and N. Macfadyen

Phil. Trans. R. Soc. Lond. A 2000 **358**, 2165-2182
doi: 10.1098/rsta.2000.0639

10.1098/rsta.2000.0639

An approach to service level agreements for IP networks with differentiated services

By R. J. Gibbens[1], S. K. Sargood[2], F. P. Kelly[1], H. Azmoodeh[2], R. Macfadyen[2] and N. Macfadyen[2]

[1]*Statistical Laboratory, Centre for Mathematical Sciences, University of Cambridge, Wilberforce Road, Cambridge CB3 0BW, UK*
[2]*Internet and Data Networks, Adastral Park, Martlesham Heath, Ipswich IP5 7RE, UK*

In this paper we report on a study of possible service level agreements in an IP network employing differentiated services. We discuss the nature of the quality of service guarantees given to network flows and relate this to the capacity provisioning processes of network operators.

A contribution of this paper is to address the way service level agreements might be determined from a coherent collection of models of network phenomena which themselves naturally operate on widely differing time-scales. The very fastest time-scales within IP packet networks are measured in microseconds to milliseconds, and are associated with buffer management and packet marking procedures inside IP routers. The next fastest time-scale relates to session level controls embedded within the end-system behaviour of the TCP/IP congestion avoidance algorithms, operating in the range of milliseconds to seconds. The per-packet routing and the management of aggregated traffic flows can take place over time-scales ranging from seconds to minutes to days. Provisioning of network resources takes place over intervals of weeks and months. All of these phenomena influence the overall structure of service level agreements.

This paper highlights the use of quantitative modelling methods which address fundamental concerns for network operators seeking to provide differentiated IP Quality of Service. The work described here is at a preliminary stage, but provides strong motivation for both further study and experimental validation. Our tentative conclusion is that the DiffServ Quality of Service mechanism is unlikely to be able to provide real measurable distinctions between classes on a pure IP network with no access restrictions, without *either* bandwidth partitioning at a lower layer *or* gratuitously damaging some traffic. It will, however, function as a back-stop *minimum guaranteed level* in times of congestion.

Keywords: Internet Protocol; Quality of Service; stochastic modelling; Transmission Control Protocol

1. Introduction

The differentiated services framework has been proposed within the Internet Engineering Task Force (IETF) to provide multiple Quality of Service (QoS) classes over

Internet Protocol (IP) networks. A field within the packet header is used to indicate the per-hop-behaviour (PHB) of the packet, and its forwarding treatment by routers. Traffic is aggregated according to the PHBs, without the need for per-flow state information.

Within the network, packets may, in practice, be given differentiated QoS using some form of priority queuing (scheduling), or using threshold dropping within the output buffer of the router (buffer management). In the former case, strict priority queuing or weighted fair queuing may be used to give packets in one queue priority over another queue. In the latter case, thresholds may be applied such that when a buffer occupancy reaches a threshold, packets with lower priority are dropped.

Since there is no signalled or per-flow control, performance guarantees rely on accurate dimensioning, and the use of policers at the edge of the network to ensure that users remain within their agreed profiles. The connectionless nature of IP networks means that the traffic matrix cannot be specified. However, based on a combination of network measurements, dimensioning and policing, the determination of statistical bounds on the end-to-end performance can be attempted.

In general, the differentiated services framework defines the components (such as edge policing and router forwarding) which will transport IP packets across multi-domain networks, and services are expected to be built using whatever components are available in a flexible and scalable manner. However, while the understanding of the performance and features of individual components is actively being researched and is relatively well understood, the overall behaviour of the network has attracted less attention. The IETF Differentiated Services Working Group is developing the building blocks for providing IP QoS, and it is the domain of Internet Service Providers (ISPs) and network operators to determine how and when to use them in building end-to-end services. While this maximizes flexibility and maintains openness in architectural developments, it leaves *end-to-end* issues to be addressed.

2. Problem definition

Explicit recognition of the different time-scales involved in modelling is essential. Events at the microsecond/millisecond time-scales (algorithms for packet forwarding, buffer management in routers) have to be related upwards progressively to higher layers through session control (millisecond to second), signalling (seconds to minutes), traffic engineering (minutes to hours to days) and then capacity planning (weeks to months). This approach has been used extensively in operations research (usually categorized as reactive, tactical and strategic) and applied in asynchronous transfer mode (ATM) networks for the development of effective bandwidths concepts and admission control. The time-scales and parameters of interest in this work are shown schematically in figure 1.

For the work reported in this paper, we simplified the modelling of the IETF defined PHBs for expedited forwarding (EF) and assured forwarding (AF). Three types of traffic class were assumed (note that strict definitions of EF and AF are as behaviours and traffic classes and are defined here to imply which behaviour is being considered). The following collection of classes was selected.

(i) EF/voice is high priority, needing bandwidth and delay assurances, and is non-adaptive (that is, it does not back-off its sending rate under congestion) and subjected to admission control.

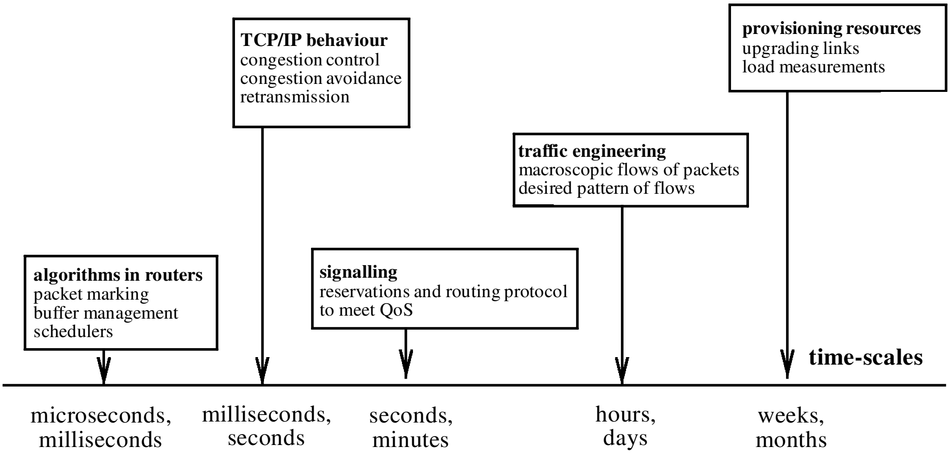

Figure 1. Controlling a data network.

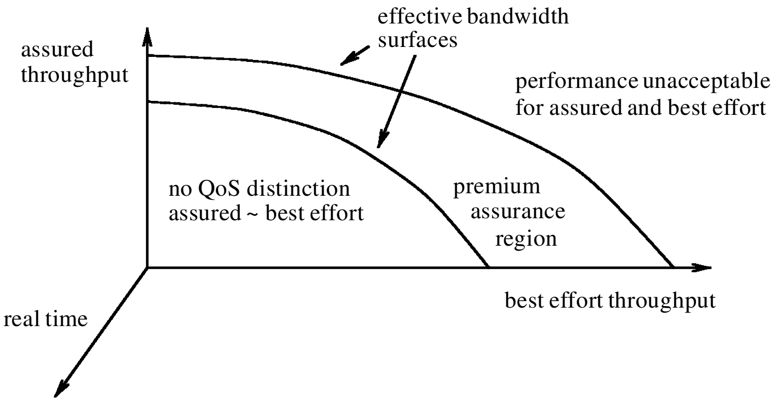

Figure 2. DiffServ QoS assurances.

(ii) AF1/TCP is a premium data class with defined QoS assurances (though relaxed compared with EF/voice), and is based on the Transmission Control Protocol (TCP).

(iii) AF2/TCP is a best effort data class with a minimal QoS assurance (more relaxed even than AF1/TCP) and is also based on TCP.

Figure 2 illustrates how the three classes under increasing network load may behave, with respect to a known QoS parameter. This could be delay or packet loss, but for classes based on AF it was decided that throughput of TCP traffic was the most suitable parameter, depending implicitly on both round-trip delay and packet loss. The key question is whether regions of QoS assurances could be defined (with hard/soft boundaries analogous to effective bandwidth surfaces (Hui 1988)), the shape and size of them and the key factors influencing them over the various time-scales.

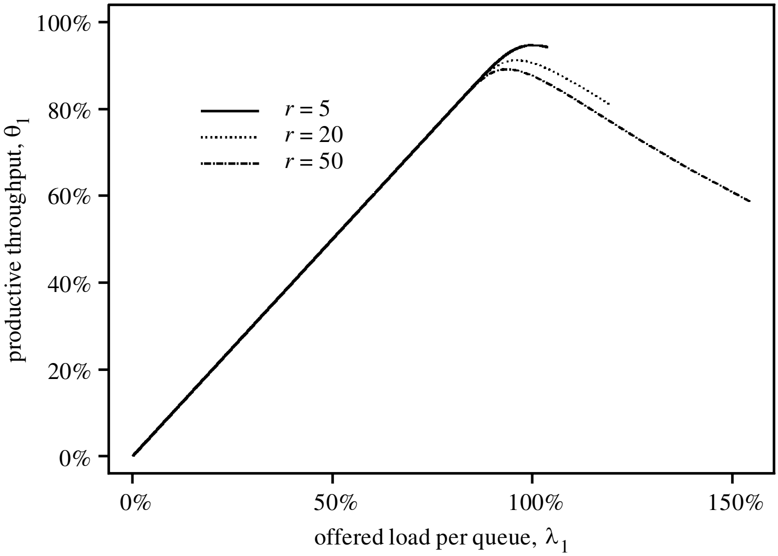

Figure 3. The productive throughput per queue θ_1 as a function of the offered load per queue λ_1, each expressed as a percentage of the capacity per queue. Buffer size $B = 50$, one class of traffic.

3. Packet level end-to-end models

Here we outline the packet level end-to-end models developed to examine the behaviour of multiple data traffic classes in a network. The traffic is assumed to be uniformly either non-adaptive (such as that based on the Universal Datagram Protocol (UDP)) or adaptive (such as that based on TCP). We use fixed-point (or reduced-load) approximations to generalize single resource models (May *et al*. 1999) and take into account traffic thinning from packet losses as route lengths, given in terms of the number of resources, increased. This approach allows end-to-end performance issues to be addressed.

We consider a very simple model, discussed in the appendix, where the network is assumed in the first instance to be homogeneous, with all routes of identical length and all links seeing an equal number of routes and traffic levels. Figure 3 shows an example of our results, for UDP-like traffic. Observe the occurrence of congestion collapse (Floyd & Fall 1998); increasing offered load eventually decreases overall network throughputs, as congested resources spend time forwarding packets which will be dropped later.

Figure 4 shows an example of our results for TCP traffic, where TCP sessions are assumed to operate in the congestion avoidance phase (Jacobson 1988) and the network is in a quasi-static state—that is, the number of TCP sessions changes relatively slowly. These assumptions allow the simple form for productive throughput η in relation to packet loss p and round-trip time RTT to be used, which is reported in studies by Floyd & Fall (1998) and Mathis *et al*. (1997), namely,

$$\eta(p, \mathrm{RTT}) = \frac{1}{\mathrm{RTT}\sqrt{\frac{2}{3}p}}. \tag{3.1}$$

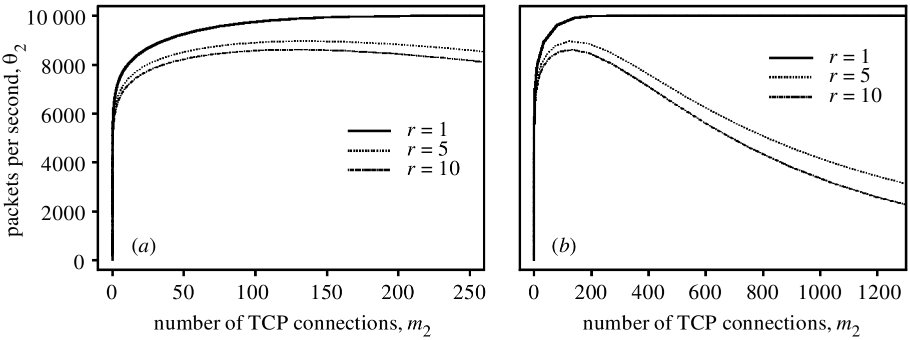

Figure 4. Productive throughput of TCP traffic, as a function of the number of TCP connections. The parameters are RTT = 50 ms, $B = 100$, $T = 50$, $r = 1,\ 5,\ 10$, $C = 20\,000$ packets per second and $\nu_1 = 10\,000$ packets per second.

This form addresses TCP in its natural operating state, rather than examining transient effects due to repeated slow-start. (It is recognized that many TCP sessions are of very short duration on the public Internet today, partly due to the HTTP protocol which requires a separate session for each object (text/graphics) downloaded.)

We note from figure 4*a* that there is a very mild form of congestion collapse, above *ca.* 150 connections for routes of length 5 and 10. Figure 4*a* suggests that, end-to-end along a route, the productive throughput shared between TCP connections does not depend heavily upon the *number* of TCP connections. This supports the connection-level representation of a network as a processor-sharing queue advocated by Heyman *et al.* (1999) and Massoulié & Roberts (1999). Observe that the total productive throughput is less than the capacity of the resource; in figure 4 the maximum productive throughput when $r = 5$ is *ca.* 80% of the capacity of the resource.

Figure 4*b* extends the horizontal axis, giving the number of TCP connections, by a factor of 5: note the extremely large number of TCP connections necessary to produce significant congestion collapse. Processor-sharing models assess the probability that n or more TCP connections are in progress to be about ρ^n, where ρ is the traffic intensity: a traffic intensity of 0.9 would imply a probability of more than 100 connections of *ca.* 3×10^{-5}.

Note that the models leading to the form (3.1), and presented in the appendix, assume that the packet loss probability p is constant. This is a reasonable assumption for packet-level models, on a time-scale where the number of connections does not change substantially, and the form (3.1) then implies that the throughput of a connection is inversely proportional to its round-trip time. However, on longer time-scales, if the resource is not fully utilized, then a different conclusion is reached. If, over longer time-scales, the number of connections fluctuates, then a connection's throughput over these longer time-scales is more heavily influenced by the utilization of the resource than by the round-trip time of a connection, a point we develop further in the next section.

4. Service level agreements

Our aim in this section is to determine quantitatively the boundaries of the QoS assurance regions in figure 2 in terms of the service level agreements (SLAs) for the different EF and AF traffic classes.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

Table 1. *Illustrative service level agreements*

class	QoS guarantees
EF	packet loss $\leqslant 10^{-6}$ connection blocking $\leqslant 10^{-3}$
AF1	mean throughput of connection $\geqslant$ 2000 kb s^{-1} measured over periods of 1 min, with probability 0.99
AF2	mean throughput of connection $\geqslant$ 128 kb s^{-1} measured over periods of 10 min, with probability 0.95

The form of SLAs studied for different traffic classes is illustrated in table 1. For EF/voice traffic, the principal QoS assurance is a limit on packet loss, which is achieved through a connection acceptance control (CAC), which results in connection-level blocking. The precise formulation of the AF QoS assurance is delicate, and is expressed in terms of throughput: over short periods, TCP throughput may be constrained by packet delay or loss, through the form (3.1), but over the longer periods used in the SLA we shall see that utilization is the key influence.

(*a*) *Modelling assumptions*

To investigate the form of SLAs, a single link was studied, under the following modelling assumptions.

(a) No access throttling (ingress or egress); that is, a demand has full access to as much of the link's bandwidth as is available.

(b) EF traffic is served with absolute priority; and AF1 has similar priority over AF2.

(c) No bandwidth partitioning of the network.

(d) No gratuitous damage to any class; that is, no attempt is made to achieve a QoS distinction by deliberately holding back any traffic.

The last two assumptions imply that we are studying a network where *all* service differentiation is to be done through varying service disciplines (priority) in a unified network where all streams have access to all resources. It is, of course, straightforward to produce different QoS levels in a network where streams are segregated, but segregation also necessarily implies running the network at lower overall efficiency, and is not considered here.

These assumptions are crucial to the implications of this work. Note, in particular, that the assumption of strict priorities between the classes has been chosen deliberately to be both simplistic and extreme, so as to give the maximum possible QoS separation between the classes.

The statistical characteristics of the traffic classes were as follows.

(i) EF/voice. Voice calls arrive as a Poisson process with fixed mean arrival rate and mean holding time. A call comprises 'on' and 'off' talk-spurts, which are similarly distributed.

(ii) AF1/TCP. Premium data TCP *sessions* arrive as a Poisson process, and the file sizes to be transferred have known mean and coefficient of variation.

(iii) AF2/TCP. As for AF1/TCP, but possibly with different values for the parameters.

(iv) All TCP sessions are in congestion-avoidance, and TCP sessions delay, rather than reduce, this workload on the network at times of congestion.

(v) Arrivals of voice calls and TCP sessions form independent processes.

The modelling throughout this stage was at the *connection level*; that is, it does not consider explicitly the detailed packet-level dynamics, which is subsumed in the assumption that all AF (TCP) connections are in congestion avoidance. The packet-level dynamics operate on a shorter time-scale and are taken into account by degrading the achievable link throughputs by a proportion, in line with the findings from § 3. Note, in particular, that it is not possible to specify packet *loss* rates for the AF classes, since these are determined by, and in their turn determine, the throughput.

A CAC for the EF traffic was assumed which limits the total number of such connections to a maximum value K, such that, at this number of connections, the probability of more talk-spurts being simultaneously active than the link bandwidth can support is 10^{-6}. The details are given in the appendix.

A simple fluid-flow-type model was used, which assumes that the AF1/TCP traffic sees the service capacity reduced by that required for the EF/voice (the usual simple approximation for a second-priority class). The details are covered in the appendix. From this, over a time-scale sufficiently large compared with that set by the talk-spurt variation of the EF/voice traffic, the distribution of the total volume of service effort available to the AF1/TCP traffic will be Normal (Gaussian), as is that of the AF1/TCP service demands arriving in any interval.

The AF2/TCP traffic was treated inductively. The offered load for both AF1/TCP and AF2/TCP traffic that just satisfies the throughput for reference connections in each of these traffic classes is determined, accounting for the presence of EF/voice traffic. Figure 5 illustrates the QoS assurance regions for EF and both AF traffic classes with different resource capacities.

The AF traffic is assumed to be TCP only (that is, it is adaptive under congestion); however, it can be generalized to mixes of UDP (non-adaptive) and TCP traffic by modifying the buffer management policy of the resource. If UDP traffic is not discarded preferentially over TCP traffic, its net effect will be to consume available spare capacity from EF/voice before AF1/TCP, and consume available spare capacity from AF1/TCP before AF2/TCP; that is, reduce the potential load offered to TCP traffic in these classes.

The approximation of a priority queue through the reduced-service-rate approach is, of course, classical and known to have deficiencies. Also, it is important to be clear on the region of applicability of the approximations. While it is reasonably clear in the AF1/TCP performance analysis that the EF/voice traffic satisfies the assumptions of the central limit theorem, for the AF it is not so clear; it is important to check in any specific instance that a substantial number of AF1/TCP demands can be expected to be processed during the SLA period, both for AF1/TCP and AF2/TCP.

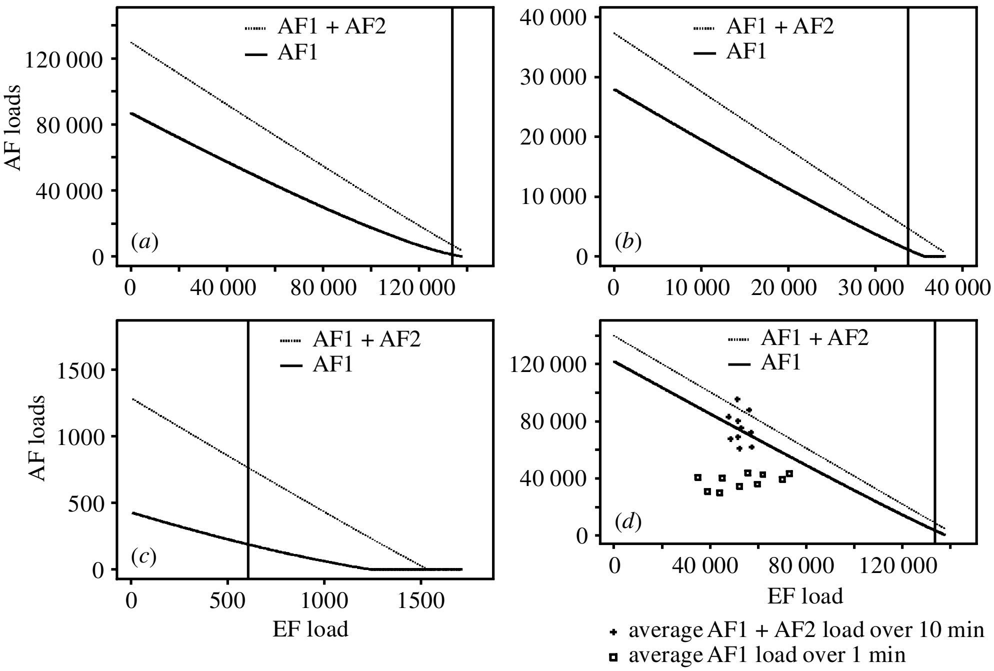

Figure 5. QoS assurance regions for EF and AF classes. The parameters of the SLAs are given in table 2. The solid line shows the maximum AF1 load such that the AF1 SLA is not violated. The dotted line shows the maximum possible total AF load (AF1 + AF2) such that the AF2 SLA is not violated.

Table 2. *Service level agreement parameters*

figure part	capacity (kb s^{-1})	throughput (kb s^{-1}) AF1	AF2	measurement period (s) AF1	AF2	mean file size (kb)
figure 5*a*	145×10^3	2000	128	60	600	80 000
figure 5*b*	40×10^3	2000	128	60	600	8000
figure 5*c*	1.8×10^3	128	32	60	600	8000
figure 5*d*	145×10^3	2000	128	60	600	8000

Similar remarks apply to the details of the EF/voice traffic. The parameters of this are set solely for illustrative purposes, and assume a mean rate of 32 kb s^{-1} and duration of 312 s. Any realistic application would naturally use the actual data rate, inclusive of packet overheads.

(*b*) *Key results*

Figure 5 plots the maximum allowed AF traffic against the EF traffic carried for the putative SLA just to hold. In each case, the upper curve (shown as a dotted line) is the maximum total AF traffic possible, such that the AF2/TCP SLA still holds; the lower curve (shown as a solid line) shows the maximum AF1 traffic such that its

SLA is not broken. Provided the AF1/TCP load is below this limit, it will receive adequate service whatever the AF2/TCP value; conversely, if it exceeds this limit (and hence does not meet its SLA), the AF2/TCP will still be satisfactory, provided the total traffic is below the upper limit.

In all cases, the vertical line shows the maximum allowable EF traffic load for that traffic to meet its own (very different) SLA. Because of the modelling assumption of strict priority, the AF load has no effect at all upon meeting the SLA for EF traffic.

Inspection of figure 5 shows several features, which we now discuss.

The assurance region for clearly differentiating QoS of all three traffic classes can be rather narrow. In fact, the traffic classes will receive very similar QoS assurances for a wide range of traffic loads, and only when the resource starts to become congested (typically at loads in excess of around 0.8) will the classes clearly differentiate themselves, and this region rapidly becomes one where no QoS assurances can be met because the resource is experiencing severe congestion. It may be observed intuitively that aggregation of traffic in coarse classes combined with measurements of a mean parameter (throughput here) over a sufficiently long time-scale will wash out second-order effects, and provide little differentiation of classes under low load, but allow them to be distinguished at high loads. Thus quantitative differentiation is considerably harder to achieve than relative differentiation.

The relative size of the QoS assurance region increases as the resource capacity decreases from 155 to 2 Mb s^{-1}. This suggests that access provisioning is the critical factor to QoS, and that core network performance will have little effect on traffic classes unless it is in a congested state. This case may occur during periods of rapid customer growth which network expansion through resource provisioning fails to match, or at inter-domain boundaries, or between ISPs where bandwidth is either expensive or scarce (for example, over trans-oceanic or leased terrestrial routes).

Reducing the throughput guarantees to 128 (AF1/TCP) and 32 kb s^{-1} (AF2/TCP) has a minimal effect on the assurance boundaries shown in figure 5*a*, because the overall link capacity is large. In figure 5*c*, by contrast, because the link size is much smaller, the EF boundary is much more restrictive and the illustrated throughput targets for AF TCP traffic have had to be reduced, as indicated, to lower values in order to obtain any meaningful boundaries.

(*c*) *Discussion of model*

The M/G/1 model does not rely on an assumption that connections share the same approximate round-trip time and it is perhaps surprising that SLAs can nonetheless be assured with large admissible regions. The intuition is as follows. If resources are operating a margin within their capacity, then they will tend to be idle sufficiently often for all connections, even those with long round-trip times, to be satisfied. At overloaded queues where the number of TCP connections is fixed, the throughput of a connection is inversely proportional to its round-trip time. But intuition developed from the overloaded case is misleading. If SLAs are to be met, then queues must operate a margin within their capacity; if, as a consequence, busy periods are short, then the impact of different round-trip times will be mitigated.

One circumstance in which connections with a long round-trip time might suffer service degradation occurs when AF/TCP file sizes have a long tail, producing a large coefficient of variation. Our models predict that higher coefficients of variation will reduce the allowable region. We expect that our current models overestimate this effect in the case where round-trip times are of similar magnitudes. If the loads are controlled, then even heavy-tailed distributions will have little effect on AF1 throughputs (see Zwart & Boxma 1998), when connections share the same approximate round-trip time.† We note that there exist proposals for TCP that eliminate the round-trip time bias (Floyd & Jacobson 1992); if the bias is a problem in DiffServ networks, then these proposals provide a means to remove it at source. Alternatively, the 'small print' of the SLAs might define throughputs as guaranteed for *reference* round-trip times chosen on the basis of historical data on the mix of round-trip times.

5. Provisioning and service differentiation

(*a*) *Provisioning*

Suppose that in a previous period (day, week or month) traffic has been measured to give the points illustrated in figure 5*d*. If any of the measurements are close to the solid line, then the SLAs for AF1/TCP traffic are in danger of violation. If any of the measurements are close to the dotted line, then the SLAs for AF2/TCP traffic are in danger of violation. Observe that the mean value of EF load must be the same for the two clouds, but the first cloud has larger variance, since its measurement interval is smaller. On the other hand, the vertical variance of the second cloud may be larger because of the inclusion of AF2/TCP traffic, especially if this is a substantial proportion of traffic.

(*b*) *Service differentiation*

In this section we consider whether there is a discernible or substantial distinction between the end-to-end performance of the AF1/TCP and AF2/TCP classes.

First, consider a link of given capacity. Figure 6*a* depicts a situation with a fixed EF/voice load. Suppose that the SLA is for AF1/TCP connections to receive a certain throughput measured over some nominated time interval. Then, if the AF1/TCP load is too high, the SLA will be violated, as shown by the right-hand vertical strip. If the AF1/TCP load is sufficiently low (say, within the left-hand vertical strip), then connections experience a throughput of at least a certain quantity and so the link is effectively transparent to that extent. Whether the SLA is violated or the link is effectively transparent does not depend on the levels of lower priority traffic.

Figure 6*b* shows the violated and transparent regions for AF2/TCP traffic. Observe that the regions now depend on the levels of both the AF1/TCP and the AF2/TCP load.

Figure 6*c* uses a combination of the two previous figure parts to show where the SLA of at least one class of connections is violated or where both classes are effectively transparent (and hence not effectively differentiated from the perspective of a SLA based on throughput).

The tentative conclusion we draw from this discussion is that there may be a very narrow operating region for the EF, AF1 and AF2 loads where the resource is not

† Long-tailed distributions for AF1/TCP file sizes will produce very-long-tailed distributions for the busy period of the AF1/TCP queue: these are starvation periods for AF2/TCP connections whose throughputs would be affected.

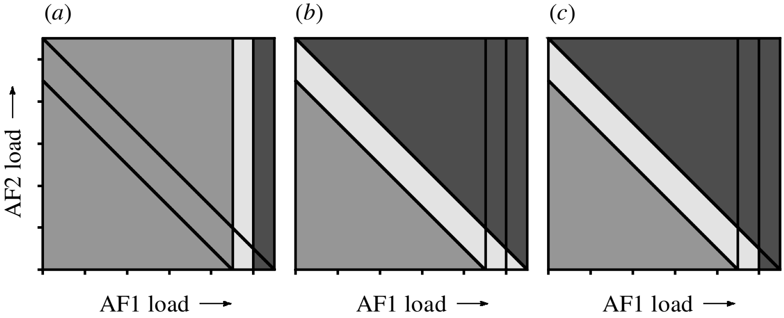

Figure 6. Differentiation between service classes AF1 and AF2. The parts show the load regions corresponding to either satisfaction or violation of an SLA and also where the network becomes effectively transparent for connections of a particular service class.

loaded sufficiently to violate any SLAs, and yet is loaded sufficiently to produce service differentiation between AF1/TCP and AF2/TCP traffic.

In fact, the traffic classes will receive very similar QoS assurances for a wide range of traffic loads, and only when the resources starts to become congested (typically at loads in excess of 0.8) will the classes clearly differentiate themselves, and this region rapidly becomes one where no QoS assurances can be met because the resource is experiencing severe congestion.

(c) *Sensitivity to traffic models*

A critical model sensitivity is the assumption that the arrivals of EF, AF1 and AF2 loads are *independent* processes. To indicate the importance of this assumption, consider figure 7. Suppose that initially the AF1 and AF2 loads are described by point A, so that both classes receive transparent service. Now suppose the AF2 load increases, and the operating point moves to point B. Now AF2 traffic is constrained, while the AF1 class still receives transparent service. If there are any mechanisms whereby users or end-systems can transform their AF2 load to AF1 load, then we might expect a movement of the operating point towards point C, where both service classes are constrained, or perhaps even to point D, where the SLA is violated for the AF1 class.

A proper consideration of this important issue seems likely to require a discussion of pricing for differentiated services (Courcoubetis & Siris 1999; Gibbens & Kelly 1999; Key & McAuley 1999). An alternative framework for SLAs, which places more emphasis on the revenues generated by flows, is provided by Bouillet *et al.* (2000).

6. Conclusions

The assumptions, network scenarios and approximations in this paper have all been tailored to *maximize* the distinction between traffic classes. This implies that in any real network the distinctions will be less than this investigation suggests.

Traffic engineering and access mechanisms can both be effective approaches to ensure that differentiated services provide relative QoS assurances to a range of traffic classes; however, considerable work is required in this area to quantify the benefits

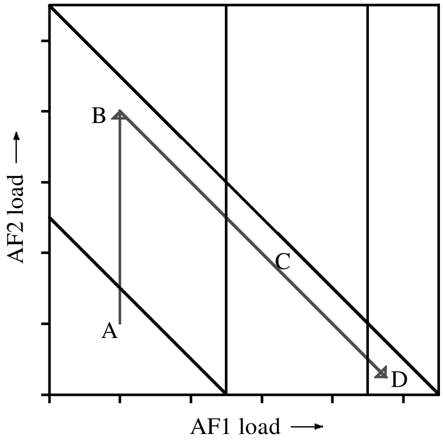

Figure 7. Illustration of possible behaviour when arrival processes are *dependent*. Increase in AF2 traffic, causing poorer performance for AF2 traffic; some AF2 traffic becomes AF1 traffic to improve its performance; SLA for AF1 traffic violated.

and match these to SLAs. Traffic engineering in general could be used to *hard-partition* the available bandwidth so that specific traffic classes could be guaranteed a given fraction of this; however, this requires that the network utilization be lower to allow for the statistical nature of the arrival process.

The work described here is at a preliminary stage, but provides strong motivation for both further study and experimental validation.

This work results from a collaborative study conducted with British Telecommunications plc.

Appendix A.

(*a*) *Packet level end-to-end models*

Fixed-point models generalize single-link models by taking into account traffic thinning from packet losses, and they provide a framework within which the adaptive nature of TCP can be represented. In part (iii) of this appendix we describe how end-systems using TCP may be represented within the model.

(i) *A symmetric network model*

Consider a symmetric network of n resources, and suppose that routes involve exactly r resources. There are $n(n-1)\cdots(n-r+1)$ such routes. Let the offered load per route be α_1 for high-priority traffic and α_2 for low-priority traffic. Let L_1, L_2 be the probabilities that a high- and low-priority packet is lost at a resource, respectively. Then, under an independent loss approximation (likely to be valid in a network with diversity of routing), the reduced load ν_1 of high-priority packets at a resource is

$$\begin{aligned}\nu_1 &= \alpha_1(n-1)(n-2)\cdots(n-r+1)(1+(1-L_1)+\cdots+(1-L_1)^{r-1}) \\ &= \alpha_1(n-1)(n-2)\cdots(n-r+1)\sum_{k=1}^{r}\binom{r}{k}(-L_1)^{k-1}. \qquad \text{(A 1)}\end{aligned}$$

Similarly, the reduced load of low-priority packets at a resource is

$$\nu_2 = \alpha_2(n-1)(n-2)\cdots(n-r+1)\sum_{k=1}^{r}\binom{r}{k}(-L_2)^{k-1}. \tag{A 2}$$

If there were no loss in the network, then the *offered load* per queue would be

$$\lambda_1 = \alpha_1(n-1)(n-2)\cdots(n-r+1)r, \tag{A 3}$$
$$\lambda_2 = \alpha_2(n-1)(n-2)\cdots(n-r+1)r, \tag{A 4}$$

respectively, for high- and low-priority traffic. The loss probabilities L_1, L_2 are, in fact, functions of ν_1, ν_2, as follows

$$L_1 = L_1(\nu_1,\nu_2) \quad \text{and} \quad L_2 = L_2(\nu_1,\nu_2), \tag{A 5}$$

where the precise functional form depends on the priority mechanism used at the resources.

From (A 1), (A 3) and (A 5),

$$\lambda_1 = \nu_1 r \Bigg/ \sum_{k=1}^{r}\binom{r}{k}(-L_1(\nu_1,\nu_2))^{k-1}, \tag{A 6}$$

while from (A 2), (A 4) and (A 5),

$$\lambda_2 = \nu_2 r \Bigg/ \sum_{k=1}^{r}\binom{r}{k}(-L_2(\nu_1,\nu_2))^{k-1}. \tag{A 7}$$

Define the *productive throughput* per queue of high-priority packets to be

$$\theta_1 = \lambda_1(1-L_1(\nu_1,\nu_2))^r, \tag{A 8}$$

and of low-priority packets to be

$$\theta_2 = \lambda_2(1-L_2(\nu_1,\nu_2))^r; \tag{A 9}$$

θ_1 or θ_2 is just the throughput per queue of high- or low-priority packets, respectively, that will not be lost at later stages.

(ii) *Resource models*

A model for the behaviour of a resource may be defined as follows (May *et al.* 1999). Suppose that the resource has a buffer of size of B packets but rejects low-priority packets if there are T or more packets already in the buffer. Let j be the occupancy of the buffer and suppose that the arrival rates for high- and low-priority streams are ν_1 and ν_2, respectively. Suppose that the resource serves packets at the rate of C packets per second. We model the state j by a Markov chain with transition rates

$$q(j,j+1) = \begin{cases} \nu_1, & T \leqslant j < B, \\ \nu_1+\nu_2, & 0 \leqslant j < T, \end{cases} \tag{A 10}$$
$$q(j,j-1) = C, \quad 1 \leqslant j \leqslant B. \tag{A 11}$$

The equilibrium distribution for the state j is given by

$$\pi_j = \pi_0 \prod_{k=1}^{j} \frac{q(k-1,k)}{q(k,k-1)}, \tag{A 12}$$

where π_0 is chosen to normalize the distribution. The loss probabilities for high- and low-priority traffic streams are then given by

$$L_1(\nu_1,\nu_2) = \pi_B \quad \text{and} \quad L_2(\nu_1,\nu_2) = \sum_{j=T}^{B} \pi_j. \tag{A 13}$$

Suppose the queuing discipline at the server is first-in first-out. Then the mean delay of a packet accepted by the server when there are n packets already in the queue is the sum of $(1+n)$ independent exponential random variables, each of mean duration $1/C$. Thus the expected delays at a single resource for accepted high- and low-priority packets are given by

$$\mathbb{E}(D_1(\nu_1,\nu_2)) = \sum_{n=0}^{B-1}(1+n)\pi_n \Big/ C \sum_{n=0}^{B-1} \pi_n, \tag{A 14}$$

$$\mathbb{E}(D_2(\nu_1,\nu_2)) = \sum_{n=0}^{T-1}(1+n)\pi_n \Big/ C \sum_{n=0}^{T-1} \pi_n. \tag{A 15}$$

(We have amended the formulae in § 3 of May *et al.* (1999) so as to omit lost packets from the delay calculation.) Note that

$$L_2(\nu_1,\nu_2) \geqslant L_1(\nu_1,\nu_2) \quad \text{and} \quad \mathbb{E}(D_2(\nu_1,\nu_2)) \leqslant \mathbb{E}(D_1(\nu_1,\nu_2)), \tag{A 16}$$

while low-priority packets are less likely to be accepted than high-priority packets, accepted low-priority packets see lower mean delays than accepted high-priority packets.

(iii) *Incorporating TCP in the end-to-end model*

In this section we describe how the behaviour of end-systems using TCP may be incorporated in fixed-point models. Major assumptions underlying the numerical illustrations are that TCP is operating in the congestion avoidance phase, and that the network resources are homogeneously loaded.

TCP is a window-based protocol that ensures reliable delivery using retransmission and a congestion avoidance algorithm. Models of TCP leading to (3.1) have been developed (Floyd & Fall 1998; Mathis *et al*. 1997). Padhye *et al*. (1998) developed a more sophisticated model, showing that the flow rate $\eta(p)$ out of a TCP source, in packets per second, including retransmission, is approximately

$$\eta(p) = \min\left\{ \frac{W_{\max}}{\mathrm{RTT}}, \frac{1}{\mathrm{RTT}\sqrt{(\frac{2}{3}p)} + T_0 \min\{1, 3\sqrt{(\frac{3}{8}p)}\}p(1+32p^2)} \right\}, \tag{A 17}$$

where p is the packet loss probability, $W_{\max}$ is the receive window size, RTT is the round-trip time and T_0 is the retransmission time-out value. (We assume measures are in place at resources, such as Random Early Discard (Floyd & Jacobson 1993), to lessen packet loss correlation, and thus to lessen the chance of multiple packet losses within one round-trip time. Without this assumption, the parameter p has a slightly different interpretation (see Padhye *et al.* 1998).) Using (A 17) rather than (3.1) produces similar qualitative behaviour to that shown in figure 4, with the flat behaviour shown in part (*a*) extending to even higher levels for the number of connections.

Assume TCP traffic is low priority and let m_2 be the number of TCP connections per resource. Then the productive throughput per resource may be written as

$$\theta_2 = m_2(1-p)\eta(p), \quad \text{where } 1-p = (1-L_2(\nu_1,\nu_2))^r. \tag{A 18}$$

Hence, using (3.1),

$$m_2 = \theta_2 \times \text{RTT} \times \sqrt{\frac{2}{3}}\frac{(1-(1-L_2(\nu_1,\nu_2))^r)^{1/2}}{(1-L_2(\nu_1,\nu_2))^r}, \tag{A 19}$$

and, from (A 7) and (A 9),

$$\theta_2 = \nu_2 r(1-L_2(\nu_1,\nu_2))^r \Bigg/ \sum_{k=1}^{r}\binom{r}{k}(-L_2(\nu_1,\nu_2))^{k-1}. \tag{A 20}$$

(*b*) *Time-scale analysis*

(i) *EF traffic*

Let $n(t)$ be the number of calls present in an $M/M/\infty$ queue with mean holding time τ_1 seconds, and arrival rate ν_1 calls per second. Thus, in equilibrium, $n(t)$ has a Poisson distribution with mean $\nu_1\tau_1$. If this mean is large, then

$$x(t) = \frac{n(t\tau_1)-\nu_1\tau_1}{(\nu_1\tau_1)^{1/2}} \tag{A 21}$$

will approximate an Ornstein–Uhlenbeck process with covariance

$$\mathbb{E}(x(s)x(s+t)) = \text{Cov}(x(s),x(s+t)) = \mathrm{e}^{-|t|} \tag{A 22}$$

and stationary distribution $x(t) \sim N(0,1)$. Thus

$$\begin{aligned}\text{Var}\left(\int_0^t x(s)\,\mathrm{d}s\right) &= \mathbb{E}\left(\int_0^t\int_0^t x(s_1)x(s_2)\,\mathrm{d}s_1\,\mathrm{d}s_2\right)\\ &= \int_0^t\int_0^t \mathrm{e}^{-|s_1-s_2|}\,\mathrm{d}s_1\,\mathrm{d}s_2\\ &= 2(t+\mathrm{e}^{-t}-1)\end{aligned} \tag{A 23}$$

(see § 5.9 of Cox & Miller 1965). Note that we have not modelled the bursty nature of voice calls; over the time-scales of interest, the major variability will be caused by fluctuations in $n(t)$.

The service effort consumed over the period $[0,T]$ is

$$\int_0^T n(t)\,\mathrm{d}t = \nu_1\tau_1 T + \tau_1(\nu_1\tau_1)^{1/2}\int_0^{T/\tau_1} x(s)\,\mathrm{d}s, \tag{A 24}$$

and so may be approximated as

$$\int_0^T n(t)\,\mathrm{d}t \sim N(\nu_1\tau_1 T, 2(\nu_1\tau_1)\tau_1^2(T/\tau_1 + \mathrm{e}^{-T/\tau_1} - 1)). \quad \text{(A 25)}$$

For a link of capacity $C\,\mathrm{kb\,s^{-1}}$ and calls of mean rate $\delta\,\mathrm{kb\,s^{-1}}$, the spare capacity available over the period $[0, T]$ is thus approximately

$$N(CT - \delta\nu_1\tau_1 T, \delta^2 \times 2(\nu_1\tau_1)\tau_1^2(T/\tau_1 + \mathrm{e}^{-T/\tau_1} - 1)). \quad \text{(A 26)}$$

If calls are not exponentially distributed, a more involved analysis is possible (Whitt 1982).

(ii) *AF1 traffic*

Assume connections arrive as a Poisson process of rate ν_2 per second, and files to be transferred have mean μ_2 kb and variance σ_2^2 $(\mathrm{kb})^2$, and hence coefficient of variation $c_2 = \sigma_2/\mu_2$. The workload arriving in the interval $[0, T_2]$ is then approximately

$$N(\nu_2\mu_2 T_2, \nu_2\mu_2^2(1 + c_2^2)T_2). \quad \text{(A 27)}$$

Total spare capacity at the resource is then, from (A 26), approximately

$$N(CT_2 - \delta\nu_1\tau_1 T_2 - \nu_2\mu_2 T_2, \delta^2 \times 2(\nu_1\tau_1)\tau_1^2(T_2/\tau_1 + \mathrm{e}^{-T_2/\tau_1} - 1) + \nu_2\mu_2^2(1 + c_2^2)T_2). \quad \text{(A 28)}$$

To offer a mean throughput guarantee of θ over a time interval T, with probability 0.99, thus requires

$$\theta < C - \rho_1 - \rho_2 - (2.33/T_2)[\delta \times 2\rho_1\tau_1^2(T_2/\tau_1 + \mathrm{e}^{-T_2/\tau_1} - 1) + \rho_2\mu_2(1 + c_2^2)T_2]^{1/2}. \quad \text{(A 29)}$$

This constraint is used to determine the upper limit for AF1 load in figure 5.

(iii) *AF2 traffic*

Assume AF2 connections arrive as a Poisson process of rate ν_3 per second, and files to be transferred have mean μ_3 kb and variance σ_3^2 $(\mathrm{kb})^2$, and hence coefficient of variation $c_3 = \sigma_3/\mu_3$. The workload arriving in the interval $[0, T_3]$ is then approximately

$$N(\nu_3\mu_3 T_3, \nu_3\mu_3^2(1 + c_3^2)T_3). \quad \text{(A 30)}$$

To offer a mean throughput guarantee of θ over a time interval T_3, with probability 0.95, thus requires

$$\theta < C - \rho_1 - \rho_2 - \rho_3 - (1.64/T_3)[\delta \times 2\rho_1\tau_1^2(T_3/\tau_1 + \mathrm{e}^{-T_3/\tau_1} - 1) + \rho_2\mu_2(1 + c_2^2)T_3 + \rho_3\mu_3(1 + c_3^2)T_3]^{1/2}. \quad \text{(A 31)}$$

If $\mu_2 = \mu_3$ and $c_2 = c_3$, then this constraint can be used to determine an upper limit of (AF1+AF2) load, as illustrated in figure 5. More generally, a third dimension would be needed to illustrate the allowable region.

(iv) *System model*

We suppose the resource has rate C kb s^{-1}, and gives strict priority to EF/voice packets. These packets use a short buffer adequate to cope with packet-scale fluctuations, and there is a connection acceptance control mechanism that limits the number of EF/voice calls in progress. The spare capacity of the resource is allocated next to AF1/TCP traffic, and then to AF2/TCP traffic. We do not model here the fine detail of buffer mechanisms.

We suppose the resource accepts EF/voice calls as long as the number already in progress, n_1, satisfies $n_1 < K$, where K is chosen to be the largest integer such that

$$\sum_{k=\lceil pC/\delta \rceil}^{K} \binom{K}{k} p^k (1-p)^{K-k} (\delta k/p - C)/\delta K \leqslant 10^{-6}. \tag{A 32}$$

When K calls are in progress, the number of *on* bursts has a binomial distribution with parameters K and p. Hence the rate while in the *on* state is δ/p. Thus the numerator of expression (A 32) gives the expected excess bit rate over the capacity C, while the denominator gives the expected bit rate when K calls are in progress. The connection acceptance threshold K is chosen so that while K calls are in progress, the packet-drop probability is just less than 10^{-6}. The values $p = \frac{1}{2}$ and $\delta = 32$ kb s^{-1} were chosen for the numerical examples of § 4.

Finally, the blocking probability for arriving EF/voice calls is given by Erlang's formula $E(\nu_1\tau_1, K)$.

References

Bouillet, E., Mitra, D. & Ramakrishnan, K. G. 2000 Design assisted, real-time, measurement-based network controls for management of service level agreements. In *Proc. INFOCOM2000.*

Courcoubetis, C. & Siris, V. 1999 Managing and pricing service level agreements (SLAs) for differentiated services. In *Proc. of 7th IEEE/IFIP Int. Workshop on Quality of Service (IWQoS99), London, UK.*

Cox, D. R. & Miller, H. D. 1965 *The theory of stochastic processes.* London: Chapman & Hall.

Floyd, S & Fall, K. 1998 Promoting the use of end-to-end congestion control in the Internet. (http://www-nrg.ee.lbl.gov/floyd/end2end-paper.html).

Floyd, S. & Jacobson, V. 1992 On traffic phase effects in packet-switched gateways. *Internetworking: Research and Experience* **3**, 115–156 (ftp://ftp.ee.lbl.gov/papers/phase.ps.z).

Floyd, S. & Jacobson, V. 1993 Random Early Detection gateways for congestion avoidance. *IEEE/ACM Trans. Networking* **1**, 397–413 (ftp://ftp.ee.lbl.gov/papers/early.pdf).

Gibbens, R. J. & Kelly, F. P. 1999 Resource pricing and the evolution of congestion control. *Automatica* **35**, 1965–1985.

Heyman, D. P., Lakshman, T. V. & Neidhardt, A. L. 1999 A new method for analysing feedback-based protocols with applications to engineering Web traffic over the Internet. Research Report, AT&T Labs.

Hui, J. Y. 1988 Resource allocation for broadband networks. *IEEE J. Sel. Areas Commun.* **6**, 1598–1608.

Jacobson, V. 1988 Congestion avoidance and control. In *Proc. ACM SIGCOMM88* ftp://ftp.ee.lbl.gov/papers/congavoid.ps.Z.

Key, P. & McAuley, D. 1999 Differential QoS and pricing in networks: where flow control meets game theory. In *IEE Proc. Software* **146**, 39–43.

Massoulié, L. & Roberts, J. W. 1999 Arguments in favour of admission control for TCP flows. In *Proc. ITC16*, pp. 33–44.

Mathis, M., Semke, J., Mahdavi, J. & Ott, T. 1997 The macroscopic behaviour of the TCP congestion avoidance algorithm. *Computer Comm. Rev.* **27**(3).

May, M., Bolot, J.-C., Jean-Marie, A. & Diot, C. 1999 Simple performance models of differentiated service schemes for the Internet. In *Proc. INFOCOM99*.

Padhye, J., Firoiu, V., Towsley, D. & Kurose, J. 1998 Modeling TCP throughput: a simple model and its empirical validation. In *ACM SIGCOMM98*.

Whitt, W. 1982 On the heavy-traffic limit theorem for GI/G/∞ queues. *Adv. Appl. Prob.* **14**, 171–190.

Zwart, A. P. & Boxma, O. J. 1998 Sojourn time asymptotics in the M/G/1 processor sharing queue. Technical report, CWI http://www.cwi.nl/static/publications/reports/PNA-R9802.html.