7

Analytical Comparison of Quality of Service Systems*

In this chapter, we use two analytical approaches to compare different Quality of Service (QoS) systems. We compare two QoS systems:

1. A QoS system using admission control and a reservation mechanism that can guarantee bandwidth for flows (Section 7.1) offers service differentiation based on priority queueing for the two service classes (Section 7.2)
2. and a system with no admission control and a single best-effort service class.

We call the second model *Best-effort (BE) model/system* and the first one *QoS model/system*.

Important for the evaluation in this chapter is the type of traffic application assumed. We use different application and traffic models. *Inelastic traffic* represents multimedia applications that require a certain rate. We speak of *strictly inelastic traffic* if no loss or delay bound violations are tolerated. Most multimedia applications can tolerate a certain level of loss or delay bound violations. For example, a typical voice transmission is still understandable – albeit at reduced quality – if some packets are lost or arrive too late. Therefore, *normal inelastic traffic* tolerates a certain amount of loss or delay bound violations. *Adaptive traffic* is similar to normal inelastic traffic but can adapt its required rate to the network conditions and is thus assumed to be extremely flexible. *Elastic traffic* represents file transfer traffic like WWW, FTP or peer-to-peer traffic. The utility of the elastic traffic is a concave function of its throughput as the throughput determines when the transfer is finished; the loss probability does not directly influence the utility.

Because of the complexity of the models, the analysis is focused on a single bottleneck. The next chapter deals with larger topologies, more realistic traffic, and so on using simulations.

The first set of models (Section 7.1) used is based on Breslau and Shenker (1998); Shenker (1995). As is common and good practice in sciences, we first reproduce the results of Breslau and Shenker (1998); Shenker (1995); then we give some further insights.

*Lecture Notes in Computer Science, 3552, 2005, 151-163, Best-Effort Versus Reservations Revisited, Oliver Heckmann and Jens B. Schmitt, copyright 2005. With kind permission of Springer Science and Business Media.

The Competitive Internet Service Provider: Network Architecture, Interconnection, Traffic Engineering and Network Design
Oliver Heckmann © 2006 John Wiley & Sons, Ltd. ISBN: 0-470-01293-5

In these works, a single type of traffic (elastic or strict inelastic or adaptive inelastic) uses the bottleneck. The expected total utility is analysed by assuming a probability distribution for the number of arriving flows. The main issues investigated with these models are admission control and bandwidth guarantees.

The second set of models (Section 7.2) is a contribution of this book. Contrary to the other models, they analyse a given load situation and a traffic mix consisting of elastic and inelastic flows filling the link at the same time. By using the queueing theory and the TCP formula, more sophisticated utility functions and more realistic network behaviour than in the first set of models can be modelled. The main effects investigated with these models are scheduling and service differentiation.

When we compare the QoS and the BE system, it is quite obvious that for the same capacity (e.g. bandwidth) the QoS system will offer better QoS. But it also has a higher complexity that leads to higher costs. For judging which of the two systems is 'better', a way has to be found to put the QoS and the costs in a relationship. For the additional costs of the QoS system, more bandwidth could be bought for the BE system, improving its QoS. To compare the two systems, we have to make sure that either the costs of the two considered systems or the QoS are equal. The costs are hard to predict[1] while the QoS is measured in the models anyway. Therefore, we bring the QoS levels in line and use the *overprovisioning factor* as metric to compare the systems: A specific QoS system leads to a certain level of QoS; its overprovisioning factor is the factor with which the capacity (bandwidth) of the BE system has to be multiplied so that it offers the same level of QoS. A high overprovisioning factor indicates that QoS system is the preferable choice while an overprovisioning factor close to one indicates that the QoS system is not worth its additional complexity. The factor for which the QoS system becomes the preferable choice depends on the exact costs. With the knowledge of the overprovisioning factor and an estimation of costs for its network, an Internet Network Service Provider (INSP) can therefore make the correct decision.

7.1 On the Benefit of Admission Control

Breslau and Shenker (1998); Shenker (1995) analyse two fundamentally different QoS systems in their works:

1. A BE system without admission control where all flows admitted to the network receive the same share of the total bandwidth.
2. A reservation-based QoS system with admission control, where only the flows are admitted to the network that optimally (w.r.t. total utility) fills the network. Their bandwidth is guaranteed by the system. This system can be built using the Intserv/RSVP architecture and to a certain extent using a Diffserv/bandwidth broker architecture.

[1] The technical costs like memory usage or used CPU cycles could be predicted. However, networking has seen many technological breakthroughs in the last years, for example, for packet classification (see Section 6.3.1.1) and scheduling (see Section 6.2.1). The prediction could therefore become insignificant quickly. Furthermore, the finally relevant costs are monetary costs of the systems and they depend among many other things on business policies and marketing decisions which are – besides being almost impossible to predict – completely out of scope of this technical work.

We start with a fixed load model that assumes a given traffic load for the network; next, a variable load and finally variable load and capacity are analysed.

7.1.1 Fixed Load

The fixed load model from Shenker (1995), also published in Breslau and Shenker (1998), assumes that there are a number of identical flows requesting service from a link with capacity C. The utility function $u(b)$ of a flow is a function of the link bandwidth b assigned for that flow with:

$$\frac{du(b)}{db} \geq 0 \;\; \forall b > 0\,,\; u(0) = 0\,,\; u(\infty) = 1 \tag{7.1}$$

A flow rejected by the admission control is treated as receiving zero bandwidth, resulting in zero utility. The link capacity is split evenly among the flows so that the total utility U of k admitted flows is given by

$$U(k) = k \cdot u\left(\frac{C}{k}\right) \tag{7.2}$$

If there exists some $\epsilon > 0$ such that the function $u(b)$ is convex but not concave[2] in the neighbourhood $[0, \epsilon]$, then there exists some k_{max} such that

$$U(k_{max}) > U(k) \;\; \forall k > k_{max} \tag{7.3}$$

In this case, the network is overloaded whenever more than k_{max} flows enter the network; the system with admission control would yield the higher total utility for because it could restrict the number of flows to k_{max}.

If the utility function $u(b)$ is strictly concave, then $U(k)$ is a strictly monotonically increasing function of k. In that case, the total utility is maximised by always allowing flows to the network and not using admission control.

Elastic applications typically have a strictly concave utility function as additional bandwidth aids performance but the marginal improvement decreases with b. Therefore, if all flows are elastic, the BE system without admission control would be the optimal choice.

Looking at the other extreme of the spectrum, there are *strictly inelastic applications* like traditional telephony that require their data to arrive within a given delay bound. Their performance does not improve if data arrives earlier, they need a fixed bandwidth $\tilde{b}$ for the delay bound (see Section 6.2.2.4). Their utility function is given by

$$u(b) = \begin{cases} 0 & b < \tilde{b} \\ 1 & b \geq \tilde{b} \end{cases}, \tag{7.4}$$

which leads to a total utility of

$$U(k) = \begin{cases} 0 & k > C/\tilde{b} \\ k & k \leq C/\tilde{b} \end{cases} \tag{7.5}$$

[2] This rules out functions simple linear functions $u(b) = a_0 + a_1 \times b$ which would, by the way, also violate (7.1).

In this case, admission control is clearly necessary to maximise utility. If no admission control is used and the number of flows exceeds the threshold $C/\tilde{b}$, the total utility $U(k)$ drops to zero.

The two extreme cases of elastic and strictly inelastic applications show that the Internet and telephone network architectures were designed to meet the needs of their original class of applications.

Another type are the *adaptive applications*; they are designed to adapt their transmission rate to the currently available bandwidth and reduce to packet delay variations by buffering. Breslau/Shenker propose the S-shaped utility function with parameter κ

$$u(b) = 1 - e^{-\frac{b^2}{\kappa+b}} \tag{7.6}$$

to model these applications (see Figure 7.1). For small bandwidths, the utility increases quadratically $\left(u(b) \approx \frac{b^2}{\kappa}\right)$ and for larger bandwidths it slowly approaches one $\left(u(b) \approx 1 - e^{-b}\right)$. The exact shape is determined by κ.

For these flows, the total utility $U(k)$ has a peak at some finite k_{max} but the decrease in total utility for $k > k_{max}$ is much more gentle than for the strictly inelastic applications. The reservation based system thus has an advantage over the BE system, but two questions remain: The first is *whether that advantage is large enough to justify the additional complexity* of the reservation based QoS system and the second is, *how likely is the situation where* $k > k_{max}$. These questions are addressed in the next section with the variable load model.

7.1.2 Variable Load

7.1.2.1 Model

The previous section showed that in an overload situation where $k > k_{max}$, the reservation-based QoS system offers a certain advantage over the plain BE system for some utility functions. Breslau and Shenker (1998) analyse the likelihood of the overload situation

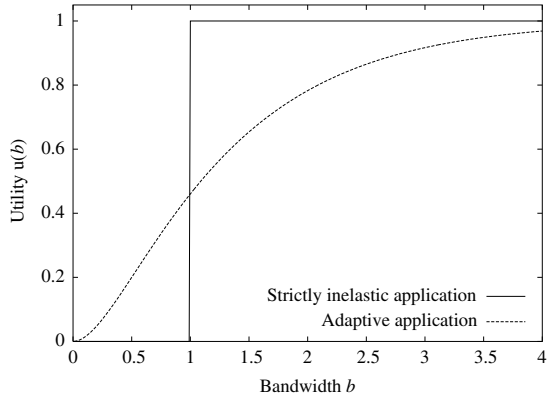


Figure 7.1 Utility Functions for $\tilde{b} = 1$, $\kappa = 0.62086$

for the strictly inelastic and adaptive applications (see Figure 7.1) by assuming a given probability distribution $P(k)$ of the number of flows k. They use two models, a model with a discrete and one with a continuous number of flows k. We base our following analysis on the discrete model[3], assuming three different load distributions (see Figure 7.2):

$$\text{Poisson:} \quad P(k) = \frac{\nu^k e^{-\nu}}{k!} \tag{7.7}$$

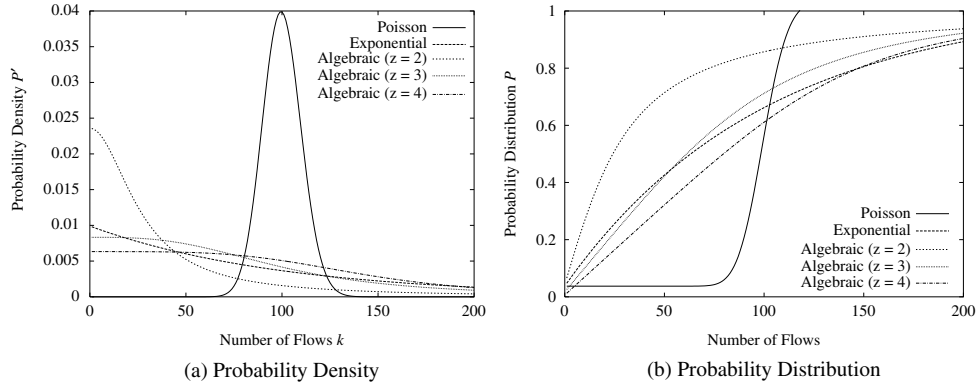
$$\text{Exponential:} \quad P(k) = \left(1 - e^{-\beta}\right) \cdot e^{-\beta k} \tag{7.8}$$

$$\text{Algebraic:} \quad P(k) = \frac{\nu}{\lambda + k^z} \tag{7.9}$$

The *Poisson load distribution* describes a scenario where the load is tightly controlled within the region around the average ν. Large or small loads are extremely rare. For the *exponential load distribution*, the load is not peaked around the average but instead decays at an exponential rate over a large range. The decay is determined by β; the expected number of flows for the exponential distribution is $E(k) = 1/\left(e^{\beta} - 1\right)$. The *algebraic load distribution* is similar but decreases slower than the exponential load distribution. It has three parameters ν, λ and z[4]. The algebraic distribution is normalised so that $\sum_{k=0}^{\infty} P(k) = 1$; we analyse $z \in \{2, 3, 4\}$.

Similar to Breslau and Shenker (1998), for the following analysis we choose the parameters of the probability distributions so that the expected number of flows $E(k) = \sum_{k=0}^{\infty} k \cdot P(k)$ is 100. Figure 7.2 depicts the probability density and distribution functions. For the utility functions, $\tilde{b} = 1$ in (7.4) and $\kappa = 0.62086$ in (7.6) this parameter setting yields $k_{max} = C$ for both utility functions.

The two utility functions analysed should be seen as the extremes of a spectrum. The strictly inelastic utility function does not tolerate any deviation from the requested minimum bandwidth $\tilde{b}$ at all, while the adaptive utility function embodies fairly large



(a) Probability Density (b) Probability Distribution

Figure 7.2 Load Distribution Functions (Continuous)

[3] That the number of flows increases in discrete steps seems more realistic. However, the continuous model is easier to solve in many cases and generally leads to similar results, see Breslau and Shenker (1998).

[4] λ is introduced so that the distribution can be normalised for a given asymptotic power law z.

changes in utility across a wide range of bandwidths above and below C/k_{max} (the level the reservation-based approach would assign to an adaptive flow).

The expected total utility $\overline{U}_{BE}$ of the BE system is

$$\overline{U}_{BE}(C) = \sum_{k=1}^{\infty} P(k) \cdot U(k) = \sum_{k=1}^{\infty} P(k) \cdot k \cdot u\left(\frac{C}{k}\right) \tag{7.10}$$

The QoS system can limit the number of flows to a k_{max}. The expected utility $\overline{U}_{QoS}$ of the QoS system is

$$\overline{U}_{QoS}(C) = \sum_{k=1}^{k_{max}(C)} P(k) \cdot k \cdot u\left(\frac{C}{k}\right) + \sum_{k=k_{max}(C)+1}^{\infty} P(k) \cdot k_{max} \cdot u\left(\frac{C}{k_{max}(C)}\right) \tag{7.11}$$

To compare the performance of the two QoS systems, Breslau and Shenker (1998) propose the bandwidth gap as a performance metric. The bandwidth gap is the additional bandwidth Δ_C necessary for the BE system so that the expected total utilities are equal:

$$\overline{U}_{QoS}(C) = \overline{U}_{BE}(C + \Delta_C) \tag{7.12}$$

As argued in the beginning of this chapter, we propose a different metric: the unitless *overprovisioning factor* OF. It puts the bandwidth gap in relation to the original bandwidth

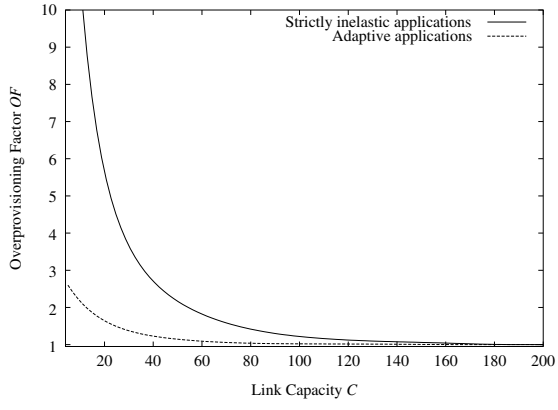
$$OF = \frac{C + \Delta_C}{C} \tag{7.13}$$

The overprovisioning factor expresses the bandwidth increase necessary for a BE based QoS system to offer the same expected total (and average) utility as the reservation based one. The higher the overprovisioning factor, the more attractive the reservation-based approach becomes; if the overprovisioning factor is close to unity, however, the additional complexity of the reservation-based approach is not justified.

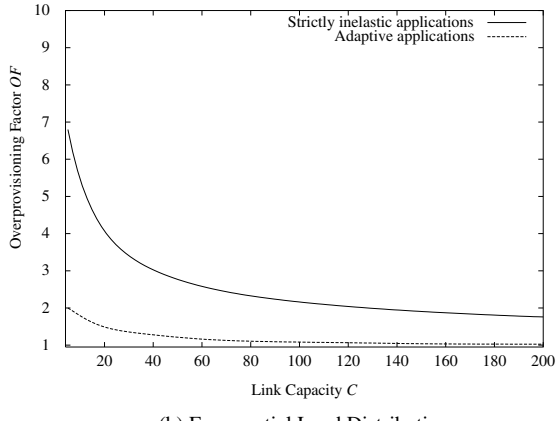
7.1.2.2 Evaluation

We now determine the overprovisioning factors. The results for the strictly inelastic and the adaptive utility function and for all three load distributions over a wide range of link bandwidths C are shown in Figure 7.3. The reader is reminded of the fact that the expected number of flows $E(k)$ is 100 in all cases.

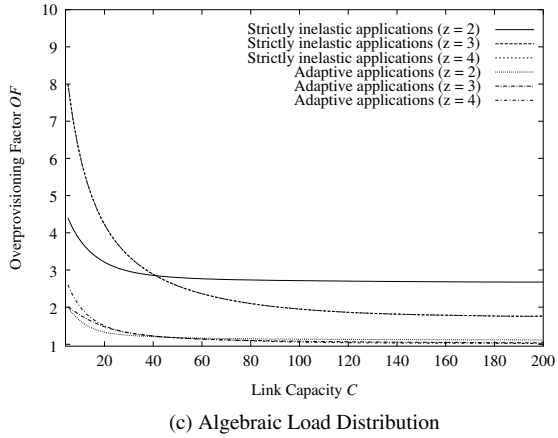
The *Poisson load distribution* (Figure 7.3 (a)) describes a situation where the load is fairly tightly controlled within a region around the average; excursions to large and small loads are extremely rare. If the link capacity is small compared to the bandwidth required by the average number of strictly inelastic flows, the overprovisioning factor is very high. It drops down to 1.2 if the link capacity equals the expected bandwidth demand and for higher bandwidths, it quickly approximates to 1.0.



(a) Poisson Load Distribution



(b) Exponential Load Distribution



(c) Algebraic Load Distribution

Figure 7.3 Results of the Variable Load Model

In contrast to the strictly inelastic application, the overprovisioning factor is much more controlled and smaller for the adaptive application. It is lower than 3.0 even if the link bandwidth is only 5% of the expected bandwidth demand and below 1.1 as soon as the link capacity exceeds 50% of the expected bandwidth demand. This demonstrates that the adaptive utility function (7.6) allows very large changes in utility across a wide range of bandwidths.

The results for the *exponential load distribution* (Figure 7.3 (b)) represent a situation where the load is not peaked around the average and decays over the whole range at exponential rate. For the strictly inelastic application, the overprovisioning factor for low capacities is lower and for higher capacities higher than the factor of the Poisson distribution. It is 2.2 if the capacity equals demand and 1.8 if the capacity is twice the demand.

For adaptive applications, the overprovisioning factor is again close to one (roughly 1.1 if capacity equals demand).

The *algebraic load distribution* also decays over the whole range but at a lower rate than the exponential distribution. The lower the z value, the slower the decay. The overprovisioning factor is quite similar to the exponential case but decreases more slowly for higher capacities. The very slow decay for $z = 2$ results in a significantly higher overprovisioning factor (2.70 if capacity equals demand and 2.67 if capacity equals twice the demand in the strictly inelastic case). For adaptive applications, the overprovisioning factor is again close to one (between 1.05 and 1.14 if capacity equals demand).

The results show that the overprovisioning factor is close to unity for adaptive applications and significantly higher than unity for the inelastic applications. The link capacity significantly influences the performance of both QoS systems and the overprovisioning factor. The capacity of the network is determined by the network design and the engineering process of the INSP. Therefore, these results are another indication that it is important to look at the QoS problem from a system-oriented point of view.

The reservation-based QoS system can provide significant advantages over the pure BE system in a well dimensioned network for strictly inelastic applications. For adaptive applications, the advantage is rather low in a well dimensioned network.

7.1.3 Variable Capacity

7.1.3.1 Model

The results above depended strongly on the relationship of the link capacity to the average number of flows and the flow/load distribution. One can further analyse the capacity level C_{opt} that maximises social welfare for both QoS systems. The social welfare W is the total utility minus the costs of the capacity C that are assumed as linear functions here:

$$W_{QoS}(C,\ p_R) = \overline{U}_{QoS}(C) - p_{QoS} \cdot C \tag{7.14}$$

$$W_{BE}(C,\ p_{BE}) = \overline{U}_{BE}(C) - p_{BE} \cdot C \tag{7.15}$$

If the provider uses a tariffing scheme that allows him to charge the users full utility, then the capacity maximising social welfare also maximises the provider's profit.

The bandwidth price of the reservation-based QoS system can be assumed to be a factor ρ higher than that of the plain BE system because of the additional complexity involved:

$$p_{QoS} = \rho \cdot p_{BE}\,,\ \rho \geq 1 \tag{7.16}$$

Now, the *equalising price factor* ρ' can be analysed as a function of the best-effort bandwidth price p_{BE} for the following situation: The reservation-based system is operated at the capacity C_{QoS}^{max} that maximises social welfare W_{QoS}. It yields the same social welfare as the BE system that is operated at the (different) capacity C_{BE}^{max}, which maximises social welfare W_{BE} in the BE case:

$$W_{QoS}\left(C_{QoS}^{max},\ \rho' \cdot p_B\right) = W_{BE}\left(C_{BE}^{max},\ p_{BE}\right) \tag{7.17}$$

If the real price factor for reservation-based capacity is higher than ρ', then the BE system offers higher social welfare (correspondingly, profit for the provider) than the reservation-based system and vice versa.

7.1.3.2 Evaluation

The equalising price factors for strictly inelastic and adaptive applications and the three different load distributions are depicted in Figure 7.4. If a certain BE price p_{BE} is exceeded, the social welfare profit becomes negative. In that case, not investing in network capacity is the optimal choice. The x-axis of Figure 7.4 only contains values of p_{BE} that lead to a positive profit.

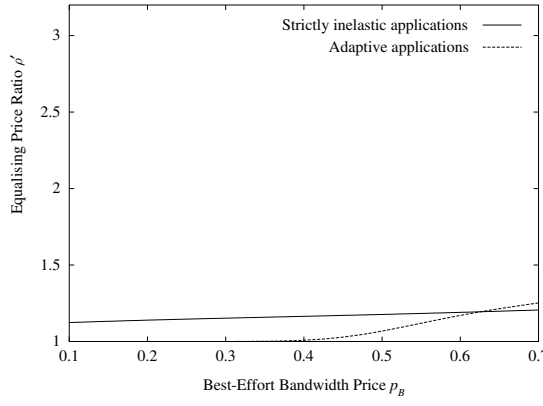
Similar to the overprovisioning factor, the equalising price factor is significantly higher for the strictly inelastic application than for the adaptive application. This holds true for all load distributions. For the adaptive applications, the equalising price ratio is below 1.25 for all distributions and p_{BE}. Thus, if the price for providing bandwidth with the reservation-based system is more than 25% higher than that of the BE system, it is in no case worth it.

The cheaper the bandwidth is (p_{BE}), the lower the equalising price factor for all load distributions. The conclusion is that the cheaper the bandwidth gets, the more attractive the BE system becomes.

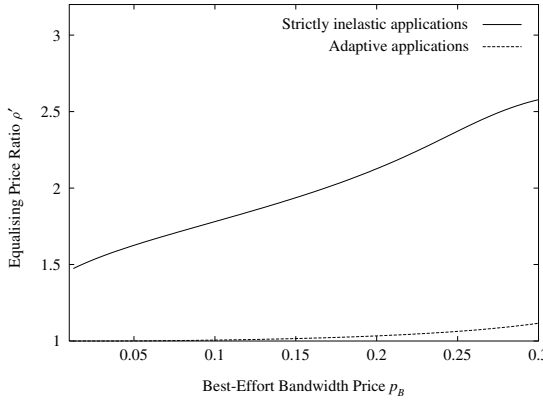
In the Poisson load distribution case, the equalising price factor is below 1.25 over a wide range of prices for both application types. For the strictly inelastic application and the exponential load distribution, the equalising price ratio is significantly higher than unity unless the BE price approaches zero. In the latter case, the equalising price ratio converges to one. For the algebraic load distribution, the equalising price ratio does not converge to one. This is shown analytically in Breslau and Shenker (1998). In these cases, the reservation-based system is preferable even if it is significantly more expensive than the BE system.

7.1.4 Summary and Conclusions

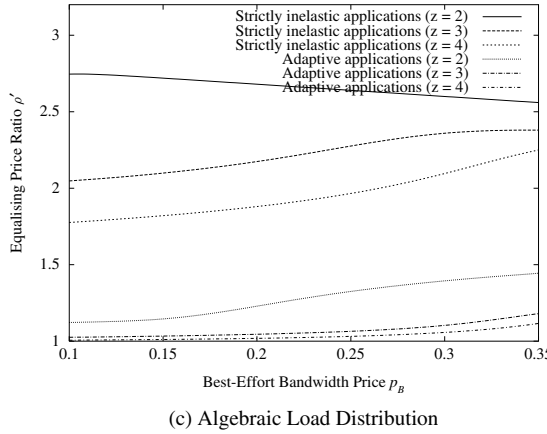
The models presented in this section help in understanding whether a reservation based or a pure BE QoS system is better. The overprovisioning factors express the amount of additional bandwidth necessary for the BE QoS system to offer the same utility as the reservation-based system. The costs of the additional bandwidth – expressed by the



(a) Poisson Load Distribution



(b) Exponential Load Distribution



(c) Algebraic Load Distribution

Figure 7.4 Equalising Price Factors

overprovisioning factor – have to be weighted against the costs of the additional complexity of the reservation based system. For linear bandwidth costs, we have seen that the bandwidth price of the reservation-based system can be twice the price of the BE system and still the reservation-based system would be the preferable choice for the strictly inelastic applications in many cases. However, as the price for bandwidth drops, the BE system generally becomes more attractive even for these types of applications.

The results indicate that for strictly inelastic applications, the reservation-based approach is probably more efficient while this is very doubtful for the discussed adaptive applications.

The above analysis in Breslau and Shenker (1998) gives valuable insights but can also be criticised in some points:

- It assumes that only a single type of application utilises the network. If different applications with different requirements utilise a network at the same time (Multiservice network), QoS systems that know the QoS requirements of the flows and can differentiate between them – for example, by protecting loss sensitive flows or by giving delay sensitive flows a higher scheduling priority – offer a further advantage over the BE system. This advantage is not included in the overprovisioning factors obtained with the models above.
- The load distributions (Poisson, exponential, algebraic) used in the models above to derive the expected utility for a given bandwidth are not based on empirical studies.
- In addition, it is doubtful whether this expected utility really represents the satisfaction of the customers with the network performance:
  If the network performance is very good most of the time but regularly bad at certain times (e.g. when important football games are transmitted), this might be unacceptable for customers despite a good *average* utility.
  Instead of assuming a load distribution and optimising for the whole range of the distribution, a provider would probably base its decision on the performance of the network in a high-load situation.

In the next section, we use a novel approach to avoid these drawbacks and shed more light on the comparison of the two QoS systems.

7.2 On the Benefit of Service Differentiation

When analysing a mix of different traffic types competing for bandwidth, it is not trivial to determine the amount of bandwidth the individual flows will receive and the delay it experiences. In this section, we present an analytical approach that – contrary to the previous approach – uses queueing theory and the TCP formula as a foundation to calculate the overprovisioning factor for a traffic mix of elastic TCP-like traffic flows and inelastic traffic flows.

7.2.1 Traffic Types

We assume that two types of traffic – elastic and inelastic – share a bottleneck link of capacity C. For *inelastic traffic*, we use index 1 and assume that there are a number

of inelastic flows sending with a total rate r_1. The strictly inelastic traffic analysed in Section 7.1 did not tolerate any loss. Most multimedia applications, however, can tolerate a certain level of loss. For example, a typical voice transmission is still understandable if some packets are lost – albeit at reduced quality. We model this behaviour here by making the utility of the inelastic traffic degrading with the packet loss[5] and with excessive delay.

For the *elastic traffic,* we use index 2; it represents file transfer traffic with the characteristic TCP 'sawtooth' behaviour: the rate is increased proportional to the round-trip time (RTT) and halved whenever a loss occurs. We use the TCP formula (4.2) to model this behaviour; the two main parameters that influence the TCP sending rate are the loss probability p_2 and the RTT delay q_2. We assume there are a number of greedy elastic flows sending as fast as the TCP congestion control is allowing them to send; their total rate is $r_2 = f(p_2, q_2)$. The utility of the elastic traffic is a function of its throughput.

7.2.2 Best-Effort Network Model

A BE network cannot differentiate between packets of the elastic and inelastic traffic flows and treats both types of packets the same way. The loss and the delay for the two traffic types is therefore equal:

$$p_{BE} = p_1 = p_2 \tag{7.18}$$

$$q_{BE} = q_1 = q_2 \tag{7.19}$$

Let μ_1 be the average service rate of the inelastic flows, μ_2 the one for elastic flows, λ_1 the arrival rate of the inelastic traffic and λ_2 the arrival rate of the elastic traffic. The total utilisation ρ is then given by

$$\rho = \rho_1 + \rho_2 = \frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2} \tag{7.20}$$

and the average service rate $\overline{\mu}$ by

$$\overline{\mu} = \frac{\rho_1\mu_1 + \rho_2\mu_2}{\rho_1 + \rho_2} = \frac{\lambda_1 + \lambda_2}{\rho_1 + \rho_2} \tag{7.21}$$

In the BE model, the loss probability p_{BE} is the same for both traffic types and can be estimated with the well-known *M/M/1/B* loss formula for a given maximal queue length of B packets assuming Markovian arrival and service processes:

$$p_{BE} = \frac{1-\rho}{1-\rho^{B+1}} \cdot \rho^B \tag{7.22}$$

For the queueing delay q_{BE} of the bottleneck link, the *M/M/1/B* delay formula is used:

$$q_{BE} = \frac{1/\overline{\mu}}{1-\rho} \cdot \frac{1 + B\rho^{B+1} - (B+1)\rho^B}{1-\rho^B} \tag{7.23}$$

[5] It can be seen as an intermediate application between the strictly inelastic and the adaptive traffic of Section 7.1.

The arrival rate λ_1 of the inelastic traffic is given by the sending rates r_1 of the inelastic flows (7.31) while the arrival rate λ_2 of the elastic traffic depends on the TCP algorithm and the network condition. As explained in Section 4.1.3, there are many works like Cardwell *et al.* (2000); Floyd (1991); Mathis *et al.* (1997); Padhye *et al.* (1998) that describe methods for predicting the average long-term TCP throughput, depending on the loss and delay properties of a flow. For our high-level analysis, we are not interested in details like the duration of the connection establishment and so on. Therefore, we use the plain square-root formula (4.2) for this analysis; it allows us to keep the complexity of the resulting model low:

$$\text{throughput} = \frac{\text{MSS}}{\text{RTT} \cdot \sqrt{2/3} \cdot \sqrt{p_2}} \tag{7.24}$$

with MSS as maximum segment size and RTT as the round-trip time. RTT is assumed to be dominated by the queueing delay q_2. The throughput of the queue can also be expressed as a function of the arrival process λ_2 and the loss probability p_2:

$$\text{throughput} = \lambda_2(1 - p_2) \tag{7.25}$$

Introducing parameter t that we call *flow size factor*, (7.24) and (7.25) can be simplified to

$$\lambda_2 = \frac{t}{q_{BE} \cdot \sqrt{p_{BE}}} \cdot \frac{1}{1 - p_{BE}} \tag{7.26}$$

t encompasses the $\text{MSS}/\sqrt{2/3}$ part of (7.24) and part of the RTT and is used to put the TCP flows in correct dimension to the inelastic flows, which are dimensioned by their fixed sending rate r_1.

The resulting best-effort network model is summarised in Model 7.1. As λ_2 is a function of p_{BE} and q_{BE} and at the same time influences p_{BE} and q_{BE}, the network model is a non-linear equation system. It can be solved with numerical methods. For individual equations, methods like the fixed point iteration method, the bisection or secant method, regula falsi, the Newton or the Newton–Raphson method can be used, see, for example, Press *et al.* (1992). For whole equation systems, the Gauss–Newton and the modified Newton–Raphson method can be used. Mathematical libraries like JMSL (Visual Numerics (2004)), MatLab (Mathworks (2004)) and Maple (Maplesoft (2004)) offer sophisticated non-linear equation solvers. We used the Maple 9 tool *fsolve* to solve the equation system.

7.2.3 QoS Network Model

To model a QoS system that differentiates between the inelastic and elastic traffic, we use priority queueing. The inelastic traffic receives strict non-preemptive priority in time and (buffer) space over the elastic traffic.

Using the *M/M/1* queueing model, the expected waiting time $E(W_1)$ for a packet of an inelastic flow depends on the expected number of packets waiting to be served $E(L_1)$ and the residual service time of the packet currently in the queue. Because non-preemptive queueing is used, the latter can be a type 1 (inelastic flow) or type 2 (elastic flow) packet;

Model 7.1 Best-effort Network Model

Parameters

r_1	Total sending rate of the inelastic flows [pkts/s]
t	Flow size factor of the elastic flows [pkts]
μ_1	Service rate of the inelastic traffic [pkts/s]
μ_2	Service rate of the elastic traffic [pkts/s]
B	Queue length [pkts]

Variables

p_{BE}	Loss probability
q_{BE}	Queueing delay [s]
λ_1	Arrival rate of the inelastic traffic at the bottleneck [pkts/s]
λ_2	Arrival rate of the elastic traffic at the bottleneck [pkts/s]
ρ	Utilisation of the queue
$\overline{\mu}$	Average service rate [pkts/s]

Equations

$$\overline{\mu} = \frac{\lambda_1 + \lambda_2}{\rho} \tag{7.27}$$

$$\rho = \frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2} \tag{7.28}$$

$$p_{BE} = \frac{1-\rho}{1-\rho^{B+1}} \cdot \rho^B \tag{7.29}$$

$$q_{BE} = \frac{1/\overline{\mu}}{1-\rho} \cdot \frac{1 + B\rho^{B+1} - (B+1)\rho^B}{1-\rho^B} \tag{7.30}$$

$$\lambda_1 = r_1 \tag{7.31}$$

$$\lambda_2 = \frac{t}{q_{BE} \cdot \sqrt{p_{BE}}} \cdot \frac{1}{1-p_{BE}} \tag{7.32}$$

because the exponential service time distribution is memoryless, the expected residual service time is $\sum_{i=1}^{2} \rho_i \frac{1}{\mu_i}$:

$$E(W_1) = E(L_1)\frac{1}{\mu_1} + \sum_{i=1}^{2} \rho_i \frac{1}{\mu_i} \tag{7.33}$$

By applying Little's Law (see Section 3.1.3)

$$E(L_i) = \lambda_i E(W_i) \tag{7.34}$$

we get

$$E(W_1) = \frac{\sum_{i=1}^{2} \rho_i \frac{1}{\mu_i}}{1 - \rho_1} \tag{7.35}$$

To determine the average queueing delay q_1, we need the expected sojourn time $E(S_1) = E(W_1) + 1/\mu_1$

$$q_1 = E(S_1) = \frac{1/\mu_1 + \rho_2/\mu_2}{1 - \rho_1} \tag{7.36}$$

For the second queue, the determination of the expected sojourn time is more complicated. The expected waiting time $E(W_2)$ and the sojourn time $E(S_2) = q_2$ for a packet of type 2 is the sum of

- the residual service time $T_0 = \sum_{i=1}^{2} \rho_i \frac{1}{\mu_i}$ of the packet currently in the queue because the queue is non-preemptive,
- the service times $T_1 = E(L_1)/\mu_1$ for all packets of priority 1
- and the service times $T_2 = E(L_2)/\mu_2$ for all packets of priority 2 that are already present waiting in the queue at the point of arrival of the new packet of type 2 and are therefore served before it
- plus the service times $T_3 = \rho_1(T_0 + T_1 + T_2)$ for all packets of priority 1 that arrive during $T_0 + T_1 + T_2$ and that are served before the packet of type 2 because they are of higher priority.

The waiting time is $E(W_2) = T_0 + T_1 + T_2 + T_3$, for the sojourn time; the queueing delay service time has to be added $q_2 = E(S_2) = E(W_2) + 1/\mu_2$. By applying (7.33) and (7.34), we get

$$q_2 = E(S_2) = \frac{(1 + \rho_1)\sum_{i=1}^{2} \rho_i \frac{1}{\mu_i}}{(1 - \rho_1 - \rho_1\rho_2)(1 - \rho_1)} + \frac{1}{\mu_2} \tag{7.37}$$

A packet of type 1 is not dropped as long as there are packets of type 2 waiting in the queue that could be dropped instead. With respect to loss, the arrival process 1 with arrival rate λ_1 thus experiences a normal *M/M/1/B* queue with a loss probability for a packet of type 1 of

$$p_1 = \frac{1 - \rho_1}{1 - \rho_1^{B+1}} \cdot \rho_1^B \tag{7.38}$$

We make the simplifying assumption that λ_1 is small enough so the loss for queue 1 is negligible $p_1 \approx 0$. For the low priority queue, the loss probability is then given by

$$p_2 = \frac{(1-\rho_1-\rho_2)}{1-(\rho_1+\rho_2)^{B+1}} \cdot (\rho_1+\rho_2)^B \cdot \frac{\lambda_1+\lambda_2}{\lambda_2} \tag{7.39}$$

The first part of (7.39) represents the total loss of the queueing system; the second part $\frac{\lambda_1+\lambda_2}{\lambda_2}$ is necessary because the packets of type 2 experience the complete loss.

The priority queueing based QoS network model is summarised in Model 7.2. Like the BE network model, it is a non-linear equation system.

7.2.4 Utility Functions

Before we compare the performance of the BE and QoS network models, we have to address the question as to which performance metrics is to be used. From the Models 7.1 and 7.2, it follows that the loss probability and queueing delay for inelastic flows are strictly smaller in the QoS model while for the elastic flows they are smaller in the BE model.

We now introduce utility functions for both types of traffic that transform the technical parameters loss and delay into a utility value.

7.2.4.1 Inelastic Traffic

The inelastic traffic represents multimedia or other real-time traffic that is sensitive to loss and delay. Therefore, the utility u_1 of the inelastic flows is modelled as strictly decreasing function of the loss probability p_1 and the deviation of the delay q_1 from a reference queueing delay q_{ref}:

$$u_1 = 1 - \alpha_p p_1 - \alpha_q \frac{q_1 - q_{ref}}{q_{ref}} \tag{7.40}$$

As a reference queueing delay q_{ref}, we use the queueing delay (7.44) of the QoS network model as that is the minimum queueing delay achievable for this traffic under the given circumstances (number of flows, link capacity, non-preemptive service discipline, etc.).

Please note that because $p_1 \approx 0$ for the QoS model, $u_1 = 1$ when the QoS model is used.

7.2.4.2 Elastic Traffic

The elastic traffic represents file transfer traffic. The utility of this traffic depends mostly on the throughput as that determines duration of the transfer. The utility u_2 is therefore modelled as a function of the throughput d_2:

$$u_2 = \beta \cdot d_2 = \beta \cdot \frac{t}{q_2 \cdot \sqrt{p_2}} \tag{7.41}$$

We determine the parameter β so that $u_2 = 1$ for the maximum throughput that can be reached if $\lambda_1 = 0$; both network models lead to the same β if there is no inelastic traffic.

Model 7.2 QoS Network Model

Parameters

r_1	Total sending rate of the inelastic flows [pkts/s]
t	Flow size factor for the elastic flows [pkts]
μ_1	Service rate for the inelastic traffic [pkts/s]
μ_2	Service rate for the elastic traffic [pkts/s]
B	Queue length [pkts]

Variables

p_1	Loss probability of the inelastic flows
q_1	Queueing delay of the inelastic flows [s]
λ_1	Arrival rate of the aggregate of inelastic flows [pkts/s]
p_2	Loss probability of the elastic flows
q_2	Queueing delay of the elastic flows [s]
λ_2	Arrival rate of the aggregate of elastic flows [pkts/s]
ρ_1	Utilisation of the queue with inelastic flows
ρ_2	Utilisation of the queue with elastic flows

Equations

$$\rho_1 = \lambda_1/\mu_1 \tag{7.42}$$

$$\rho_2 = \lambda_2/\mu_2 \tag{7.43}$$

$$q_1 = \frac{1/\mu_1 + \rho_2/\mu_2}{1-\rho_1} \tag{7.44}$$

$$q_2 = \frac{(1+\rho_1)\sum_{i=1}^{2}\rho_i\frac{1}{\mu_i}}{(1-\rho_1-\rho_1\rho_2)(1-\rho_1)} + \frac{1}{\mu_2} \tag{7.45}$$

$$p_1 = \frac{(1-\rho_1)}{1-\rho_1^{B+1}}\cdot\rho_1^B \approx 0 \tag{7.46}$$

$$p_2 = \frac{(1-\rho_1-\rho_2)}{1-(\rho_1+\rho_2)^{B+1}}\cdot(\rho_1+\rho_2)^B\cdot\frac{\lambda_1+\lambda_2}{\lambda_2} \tag{7.47}$$

$$\lambda_1 = r_1 \tag{7.48}$$

$$\lambda_2 = \frac{t}{q_2\cdot\sqrt{p_2}}\cdot\frac{1}{1-p_2} \tag{7.49}$$

7.2.5 Evaluation

The default parameter values we use for the following evaluation are depicted in Table 7.1. The effect of parameter variation is analysed later. The motivation behind the utility parameter α_p is that the utility of the inelastic flows should be zero for 10% losses (if there is no additional delay); for the parameter α_q the motivation is that the utility should be zero if the delay doubles compared to the minimal delay of the QoS system. β is chosen so that the utility of the elastic flow is 1 for the maximum throughput as explained in Section 7.2.4.2.

During the evaluation, we vary w_1, r_1 and t. For the choice of w_1, we assume that for the total utility evaluation, the inelastic flows are more important than the elastic flows because they are given priority over the elastic flows and it seems reasonable to expect users to also have a higher utility evaluation for one real-time multimedia flow (e.g. a phone call) than for a file transfer. An indication for that is the fact that the price per minute for a phone call nowadays is typically much higher than the price per minute for a dial-up Internet connection used for a file transfer.

To derive an anchor point for t, we arbitrarily determine a t_0 that leads to $\rho_1 = 20\%$ and to $\rho_2 = 60\%$ using the QoS network model. This represents a working point with $\lambda_1 = 0.2 \cdot \mu_1$ with a total utilisation of 80%. Every fourth packet is a multimedia packet, creating a typical situation where a QoS system would be considered. If t is increased to $t = 5t_0$ and λ_1 kept constant, then the proportion of multimedia packet to file transfer packet drops to 1:3.4 and for $t = 10t_0$ it drops to 1:3.8. At the same time, the aggressiveness of TCP against the inelastic flows increases in the BE network model as can be seen in the evaluation results below (e.g. Figure 7.5).

As evaluation metric we again use the *overprovisioning factor*; it is determined as follows:

- For a given r_1 and t, we determine the solution vector (p_1, q_1, p_2, q_2) of the QoS network Model 7.2.

Table 7.1 Default Parameter Values for the Evaluation

Parameter	Value
μ_1	1Mbps/(1500 bytes/pkt) = 83.3 pkts/s
μ_2	Same as μ_1
α_q	1
α_p	10
β	See Section 7.2.4.2
B	10 pkts
t	t_0, $5t_0$, $10t_0$
r_1	[0, . . . , 40] pkts/s
w_1	[1, 2, 5]
w_2	1

- The utility values $u_1 = f(p_1, q_1)$ and $u_2 = f(p_2, q_2)$ and the weighted average utility U_{ref} are derived from the solution vector with $w_1, w_2 > 0$

$$U_{ref} = \frac{w_1 u_1(p_1, q_1) + w_2 u_2(p_2, q_2)}{w_1 + w_2} \tag{7.50}$$

- For the best-effort Model 7.1, we can now also derive the solution vector (p_1, q_1, p_2, q_2) and calculate the weighted average utility U_{BE}. Unless the parameters α_p,α_q,w_1,w_2 are set to extreme values[6], the utility of the BE system is smaller than that of the QoS system ceteris paribus: $U_{BE} < U_{ref}$.
  - The BE system based on Model 7.1 is overprovisioned by a factor OF. The bandwidth respectively service rates μ_1 and μ_2 are increased by that factor OF. Additionally, the buffer space B is increased by the same factor:

$$\mu_i = OF \cdot \mu_i^{original} \tag{7.51}$$

$$B = OF \cdot B^{original} \tag{7.52}$$

  - U_{ref} is used as a reference value and OF is increased by a linear search algorithm until $U_{BE}(OF^*) = U_{ref}$.
  - OF^* is the overprovisioning factor and represents the resource increase in bandwidth and buffer space necessary for the BE system to perform as well as the QoS system w.r.t. the total utility U.

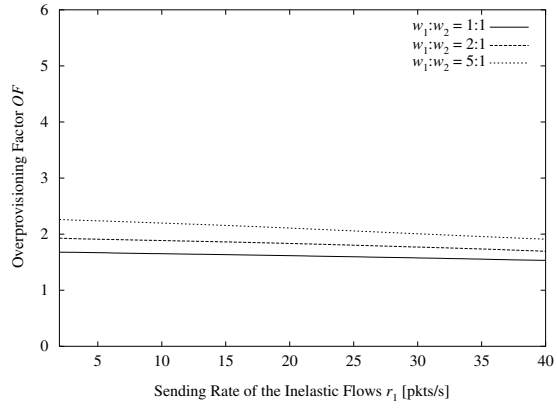
7.2.5.1 Basic Results

The overprovisioning factors OF for different flow size factors t and for different weight ratios $w_1 : w_2$ are depicted on the y-axis in the graphs of Figure 7.5. The total sending rate r_1 of the inelastic flows is shown on the x-axis.

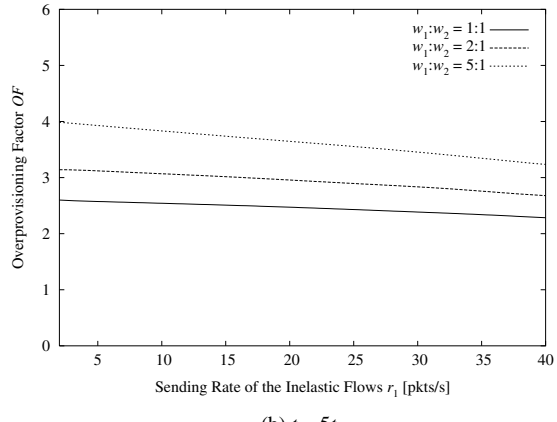
As can be seen from all three graphs, the higher the ratio $w_1 : w_2$ is – that is, the more important the inelastic flows are for the overall utility evaluation – the higher the overprovisioning factor becomes. This can be expected, because for small overprovisioning factors the utility u_1 of the inelastic flows is smaller in the BE system than the QoS system where they are protected from the elastic flows because they experience more loss and delay. Thus, the higher u_1 is weighted in the total utility function U, the more bandwidth is needed in the BE system to compensate this effect.

Comparing the three graphs, it can be seen that as the flow size factor is increased more overprovisioning is needed. Increasing the flow size factor represents increasing the number of elastic (TCP) senders and the aggressiveness of the elastic flows. In the BE system where the inelastic flows are not protected, a higher flow size factor increases the sending rate of the elastic flows on cost of additional loss and delay for the

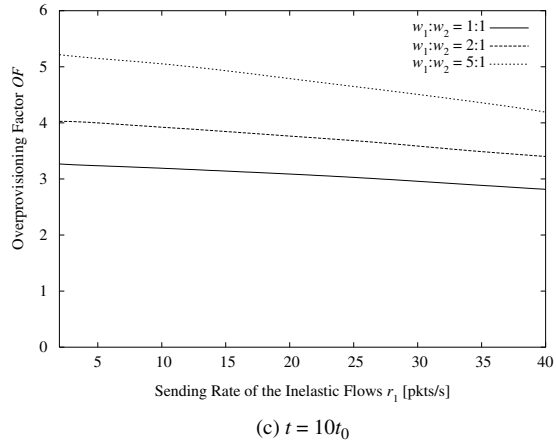
[6] Assuming $\lambda_1 = 10$, $U_{BE} < U_{ref}$ no longer holds true for example, if $w_2 > 4.58 \cdot w_1$ using the default α_i values or for $w_1 : w_2 = 2 : 1$ if the α_i are $\alpha_p < 0.05 \wedge \alpha_q < 0.005$. These values, however, are unrealistic and therefore not considered in our approach.



(a) Flow Size Factor $t = t_0$



(b) $t = 5t_0$



(c) $t = 10t_0$

Figure 7.5 Overprovisioning Factors for the Configuration of Table 7.1

inelastic flows that in return has to be compensated by more capacity leading to a higher overprovisioning factor.

Keeping the flow size factor constant, with an increase of the sending rate r_1 the overprovisioning factor decreases; the decrease is stronger when the flow size factor is higher. For a weight ratio of $w_1 : w_2 = 2 : 1$, for example, the overprovisioning factor drops from $r_1 = 2$ to 40 by 12.0% for $t = t_0$, 14.9% for $t = 5t_0$ and 15.6% for $t = 10t_0$. This phenomenon can be explained in the following way: When comparing the resulting utility values u_1 and u_2 of the QoS system with the BE system ($OF = 1$), the utility value of the inelastic flows u_1 drops because they are no longer protected. At the same time, the utility value of the elastic flows u_2 increases because they no longer suffer the full loss.

The increase of u_2 is stronger than the decrease of u_1 the higher r_1 is, therefore for higher r_1 less overprovisioning is needed.

7.2.5.2 Modification of the Utility Functions

The following graphs – unless stated otherwise – are based on a weight ratio $w_1 : w_2 = 2 : 1$ and a flow size factor of $t = 5t_0$.

If we increase or decrease the utility function parameters α_p and α_q of the inelastic traffic, the overprovisioning factor changes as shown in Figure 7.6.

A decrease of α_p and α_q represents more loss in delay tolerance of the inelastic flows as their utility is decreasing more slowly if the loss in delay increases. The lower the utility decrease is, the less additional bandwidth is needed for the BE system as compensation; therefore, the overprovisioning factor is lower.

Arguing vice versa, a higher α_i leads to a higher overprovisioning factor.

7.2.5.3 Different Bottleneck Resources

Figure 7.7 shows the overprovisioning factors if the reference buffer space B of the systems is increased from $B = 10$ to $B = 20$ while the bandwidth is kept constant ($w_1 : w_2 = 2 : 1$, $t = 5t_0$, and $\alpha_p = 10$ respectively $\alpha_q = 1$).

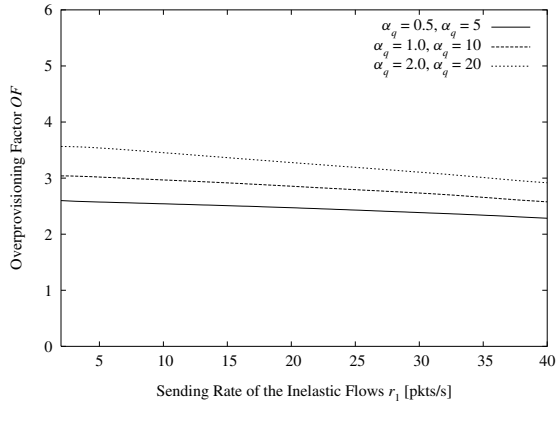


Figure 7.6 Overprovisioning Factors for Different Utility Parameters

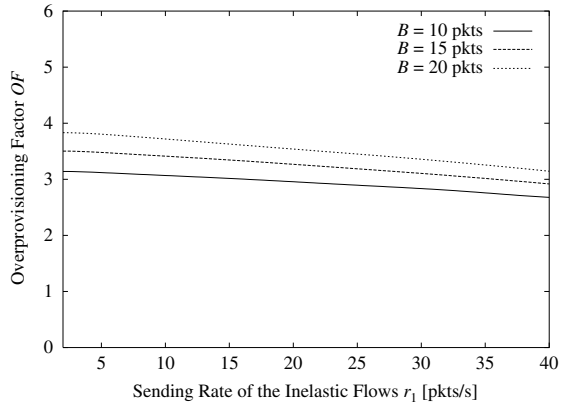


Figure 7.7 Overprovisioning Factors for Different Buffer Spaces

Increasing the buffer space B has two adverse effects; it decreases the loss rate and increases the potential queueing delay. As can be seen from the figure, an increase of B results in an increase of the overprovisioning factor OF. This is an indication that for the utility calculation, the queueing delay has a stronger effect than the loss rate. This is not surprising because for the *M/M/1/B* formulas, the loss becomes quickly negligible for larger B.

To confirm this, we reduced the queueing delay effects by setting $\alpha_q = 0.05$ and repeated the experiment. Now, with an increase of B from 10 over 15 to 20 the adverse effect can be observed: the overprovisioning factor drops from 1.76 over 1.68 to 1.66 for $r_1 = 10$.

To conclude, the effect of the buffer size depends on the ratio of α_p to α_q in the utility function.

Next, the reference buffer space B and at the same time the bandwidth (the service rates μ_1 and μ_2) are doubled; r_1 was increased accordingly. Figure 7.8 shows the results.

Compared to Figure 7.7, the overprovisioning factors only increased insignificantly for $t = 5t_0$. In the BE system – as can be seen from (7.30) – for large B, the queueing delay q_{BE} becomes inverse proportional to the service rate $\overline{\mu}$ and therefore the bandwidth. For large B, the loss p_{BE} exponentially approaches zero as can be seen from (7.29). Via (7.32), this leads to a massive increase in the elastic rate λ_2 and overall utilisation ρ. This explains why the buffer space has a larger influence than the service rate. Similar arguments hold true for the QoS system.

7.2.5.4 Different Packet Sizes

Real-time multimedia traffic like voice or video traffic usually has significantly smaller packet sizes than file transfer traffic that are mostly Maximum Transmission Unit MTU sized. The effect of the smaller packet size can be represented in the models by increasing the average service rate μ_1 of the inelastic flows. Figure 7.9 shows the results for an

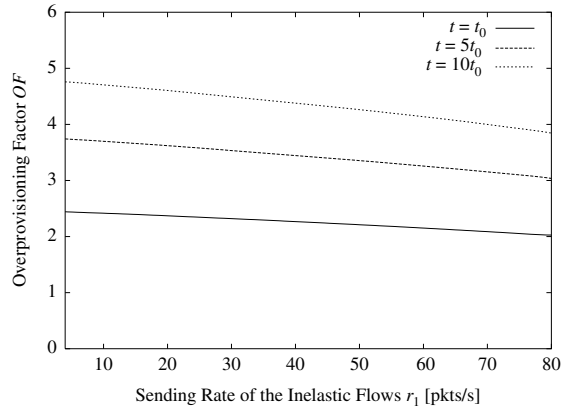


Figure 7.8 Overprovisioning Factors for an Increase in Bandwidth and Buffer Space

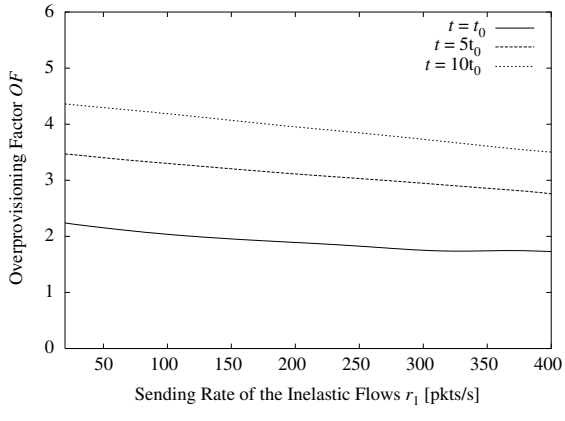


Figure 7.9 Overprovisioning Factors for Different Packet Sizes

decrease of a factor of 10 in the packet size for the inelastic flows compared to the default experiment of Figure 7.5. In this experiment, the sending rate r_1 was also increased by a factor of 10 to keep the average traffic volume constant.

As one can see, the difference in service rate increases the overprovisioning factors. This effect can be explained by the fact that the queueing theory based approach chosen in our models cannot handle different space requirements of the packets. The buffer space is limited to B packets irrespective of their type or size in our models. As the number of inelastic packets now significantly increases, the loss increases, too, and is compensated only by a further increase in bandwidth and buffer space that leads to higher overprovisioning factors. In the basic experiment of Section 7.5, the loss rate p_2 for $\lambda_1 = 10$ was 2.79%. In this experiment, for a comparable value of $\lambda_1 = 100$ the loss rate p_2 is 5.25% which confirms our explanation.

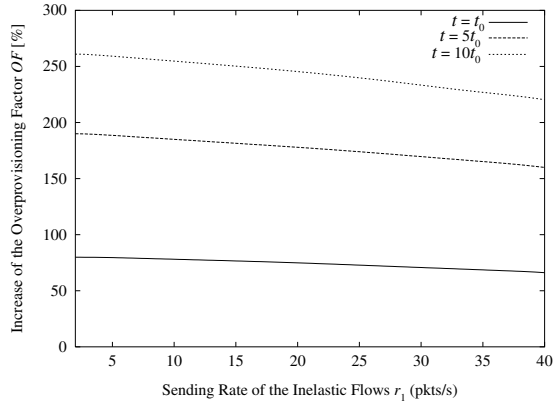


Figure 7.10 Isolation of the Service Rate Effect

7.2.5.5 Isolation of the Service Rate Effect

In the experiments so far, the bandwidth of the bottleneck link and the buffer space were overprovisioned equally. We now try to answer the question, what effect overprovisioning bandwidth alone has. Figure 7.10 depicts relative increase of the overprovisioning factor if for the BE system only the bandwidth – represented by the service rates μ_1 and μ_2 – but not the buffer space B is multiplied with the overprovisioning factor OF.

As we can see from the results, 60 to 200% additional bandwidth is needed to compensate the now missing buffer space. As a result, when overprovisioning a network the buffer space should be overprovisioned, too, unless it is significantly more expensive than additional bandwidth.

7.2.6 Summary and Conclusions

The experiments of this section evaluated the performance advantage of a priority based QoS system over plain BE system. The systems have two resources: buffer and bandwidth. We used two types of traffic – elastic and inelastic traffic – that share a bottleneck link. The evaluation is based on the aggregated utility function. Our results are overprovisioning factors. They show how much the resources of the BE system that cannot differentiate between the traffic classes have to be increased to offer the same total utility that the QoS system provides.

Compared to the approach in the previous Section 7.1, the overprovisioning factors of the models in this section are generally higher. This is explained by the fact that the models of Section 7.1 do not consider different traffic types sharing the bottleneck resources. Therefore, they miss one very important aspect of QoS systems: the service differentiation between traffic classes.

In today's Internet, the overwhelming part of the traffic is TCP based file transfer traffic, especially peer-to-peer and web traffic, see Chapter 5. In the beginning, when real-time multimedia applications spread, their initial share of traffic will be low. In our models this can be represented by rather low sending rates r_1 (few inelastic flows), and a high flow size factor t (many elastic flows). Unfortunately, our results show that especially for

this combination, the overprovisioning factors are the highest. Therefore, to support the *emerging* real-time traffic applications, QoS architectures have their greatest advantages.

The two approaches in this chapter have their limitations because they are based on analytical models that by nature only allow a certain degree of complexity to be still solvable. Our analysis is based on a single bottleneck link; the influence of the network topology has been neglected so far. We turn to simulations in the next chapter to shed more light on the question, how different QoS approaches perform. The simulations allow us to analyse more complex topologies and to employ more sophisticated traffic models.