Material for lecture #7

Reservation vs. priorities vs. best effort

Reducing risk and improving efficiency, these are the two main aspects for any traffic control or QoS model.

Improved efficiency can be considered either from service provider or from end-user perspective. It indeed appears reasonable to provide high quality video service for those customers that require it, for instance, by using appropriate resource reservation mechanisms. It is usually acknowledged that a new video call should not disturb any ongoing service in a way that the QoS of those calls is deteriorated. Unfortunately, this individual viewpoint totally misses the primary service provision issue, that is, how efficiently resources are used

Therefore, the position of this document is to stress the service provider viewpoint, because an approach to look only an individual user easily leads to situation where the negative effects caused by the service improvement of one customer to the service of other customers. These other customers include also those customers that are not active when a user is requesting the service. A principle that service requests are fulfilled *merely* based on the order of the requests, is not usually the optimal one in multiservice environment - I hope that this fact is clear at the latest in the end of this document.

The risks can be roughly divided into two categories:

- Risks related to aggregate streams
- Risks related to individual users

Aggregate risk is mainly related to the uncertainty of the number of active users while individual risk is mainly related to the uncertainty of the behaviour of individual users. These are, to some extent, two different thinks and should be considered separately.

Still the hypothesis used here is that the criteria for assessing the effects of both risks can be based on the notion of total utility. But then the evaluation has to very carefully consider the probabilities and weights of exceptional cases. Merely to minimise the risks is not a proper approach because if the costs related to the minimising effort are totally ignored, the final result could be totally impractical. In contrast, total utility is a very flexible framework, because all weights and probabilities can be set in a way that there is strong emphasis on the overload situations. By using a heavy-tailed probability distribution means that heavy-load situations can be quite common, even though the average load could be relatively low. Further, if we want to stress the negative effects of unfavourable situations, that can be done simply by defining a negative utility with large absolute value.

This document is based, once again, on an example. The basic facts of the example are presented in Table 1. The starting point is that the service provider's network is filled by voice traffic. The average number of voice users on a link during the busy hour is 1000 and standard deviation of this number is 200 (or $\varepsilon(N_{voice}) = 0.2$). The bit rate requirement is 1 [something, e.g., 30 kbps] and the utility generated by a successful call is 1 per time unit [e.g., 0.05 €/min].

Table 1.

	voice	video
Expected number of users, E[N]	1000	x
Uncertainty of N, ε(N)	0.2	0.5
Bit rate requirement	1	10
Utility of successful call per time unit	1	2
Utility of blocked call per time unit	-5	-5
Utility of bad quality per time unit	-20	-10

Now the service provider wants to introduce a new, video service that requires 10 times more capacity, and generates 2 times more utility per time unit. In addition to the lower utility per bit rate, the main problem related to this new service is that the service demand is much more difficult to predict than that of an established service. This fact is modelled by larger variation of probability distribution, ε(N) = 0.5 (actually this should perhaps be even larger - we return to this issue later).

The highest possible total utility is in this case:

$$U_{tot} = 1000 + 2 \cdot E[N_{video}]$$

In order to compare reservation and packet prioritization systems, we need to define how large reservation is made compared to the average bit rate needed by the user. The following calculations are based on a factor of 1.2, which means reservation is supposed to be 20% larger than the average bit used in reality. Note that this 20% overhead includes two main issues:

- The extra capacity that is needed to make certain that the packet loss ratio remains very small even when the traffic generated by the application is not totally constant.

- The extra capacity that is needed for signalling and maintaining the reservation information of the flow through the network.

In addition if the reservation process consumes other critical resources in the bottleneck point, for instance, processing power, these effects related to the reservation should be considered.

If all aspects are taken into account, it is quite evident that the 20% is a small rather than a large value. For instance if we a voice application uses silence detection, a constant bit rate reservation has to be almost twice as large than the average bit used by the application (otherwise the traffic control unit could discard a large amount of packets during active periods). A statistical multiplexing of voice calls can to some extent alleviate this problem, but it is totally unrealistic to assume that there is connection admission control that is able fill the network up to 100% load. According to my doctoral dissertation (about CAC in ATM networks), 80% load level is a very ambitious target with variable bit rate connections. And if we assume that the 20% margin is used to enable a realistic CAC for cases where the traffic parameters are accurately known, there is no additional margin for

a) the user to cope with the uncertainty of traffic patterns

b) signalling inside the network

Thus the following evaluation can be considered overly favourable for any reservation scheme. Even more so because the evaluation is based on an assumption that calls are blocked only if the average load (measured as reservations) exceeds the capacity of the system. In practice this is not possible due to the statistical variations in traffic process. For instance, Erlang blocking formula produces 1% call blocking if N = 1029 and A = 1000, and 0.1% for N = 1072. However, Erlang blocking formula is not realistic in our example because the dominant traffic variations are not due to the Poisson arrival process as supposed in the Erlang model, but due to the variations of the mean traffic. If we suppose that most of the traffic variations (or uncertainty of mean traffic) are caused by longer-term variations, a model in which traffic is discarded only if the mean traffic exceeds the capacity is more appropriate though not exact. Finally, this call blocking model totally ignores the fact that if the service demand exceeds the system capacity, very few customers get the service on the first attempt but needs to try several times.

Now let us first evaluate four possible scenarios for handling the situation

1. Best Effort, without any CAC or prioritization
2. Packet level prioritization, I(voice) > I(video)
3. Reservation system without any prioritization (call blocking is assumed to be the same for both voice and video calls)
4. Reservation system in which voice service has strict priority over video service. It is even assumed that new voice can break off a video call if necessary.

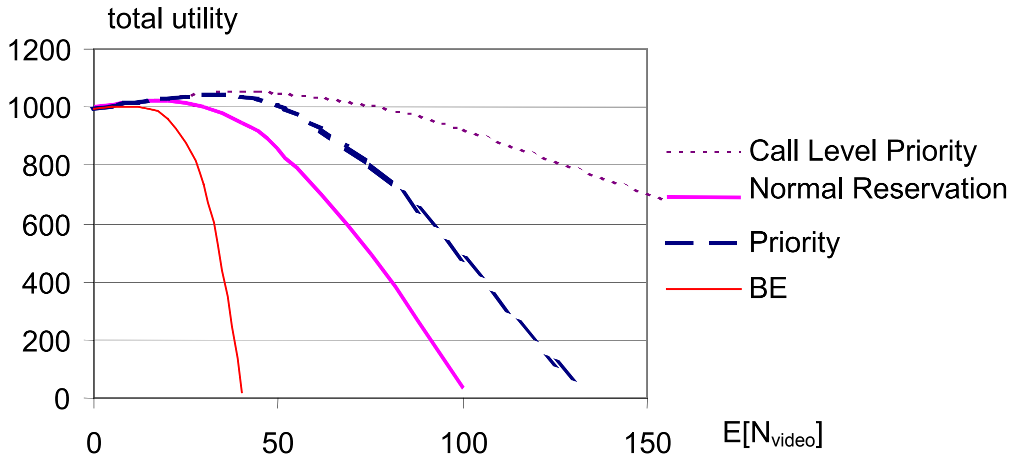

Fig. 1. Total utility as function of expecte number of video users

The main results are shown in Fig. 1. Best effort is the worst approach, as could be expected. The only, but not necessarily bad, solution with best effort is to increase capacity in order to cope with the new service class. For instance, if $E[N_{video}]$ = 50, the capacity should be increased from 2000 to 3350, in order to keep all users as satisfied as at the starting phase (U_{tot} = 1092).

Call level prioritization is basically the optimal approach, provided that a very efficient reservations can be made without any significant cost even when calls are prioritized based on the service and customer classes, or on whatsoever principle the service provider wants to use. Unfortunately this kind of system will be very hard to implement and what could be even more important is that call level prioritization is psychologically a very delicate issue. Packets can be prioritized because the prioritization system is mostly hidden behind various other effects related to the network service, applications and servers. In contrast, if you get a busy tone, or even worse, your call is terminated, because of a more important customer, your opinion about the service quality and service provider in general, could be seriously deteriorated. In general, I believe that call level prioritization will never be widely used in packet networks.

So we may assume that the operator rejects the idea of call level prioritization, but just accepts and discards new calls merely based on the load situation in the network in a way that generates equal call blocking for all service classes. The problem is that the video calls generate much less utility per bandwidth compared to voice calls, which brings about a kind of waste of resources if voice calls has to be rejected because of ongoing video calls. Even with the reservation system, the capacity required to keep all customers satisfied (U_{tot} = 1092) is as high as 2900. In this respect the saving of resources achieved by reservations is amazingly small compared to a pure best effort service; only 15% more resources are needed when the service is a pure best effort.

If someone has reservations with this outcome, I totally understand the situation. How could it be better to deliver voice and video packets just on best effort basis with a 10% of extra capacity than to have call level admission control when we are dealing with voice and video? This definitely requires more thorough studies, but still there is some grounding to believe that this indeed could be an appropriate conclusion under certain conditions.

Then what will happen if we add a simple prioritization scheme that gives precedence for voice packets because they can be seen as more important for the service provider than video packets. As fig. 1 shows the result is considerably better than that with best effort, and actually prioritization is better than a reservation system with equal call blocking. Once again, if we define that all users are satisfied with U_{tot} = 1092, the required capacity for prioritization system is 2670 (which means about 14% smaller capacity than with a best effort service).

As mentioned in the beginning, the accurate of prediction of a new service can be quite bad, and thus $\varepsilon(N_{video})$ = 0.5 can be quite optimistic. The operator may want to check the situation with larger values for $\varepsilon(N_{video})$. The results are shown in Table 2. With inaccurate knowledge about service demand, the required capacity is so huge that it is quite likely that the service provider rather decreases the target utilisation than just increases the capacity.

Table 2. *Required Capacity for U_{tot} = 1092, $E[N_{video}]$ = 50*

	$\varepsilon(N_{video})$=0.5	$\varepsilon(N_{video})$=1.0	$\varepsilon(N_{video})$=2.0
Best effort	3350	7400	25000
Packet level priority	2670	5500	20000
Reservation, equal call blocking	2900	4850	12500

The issue is further evaluated in Table 3. Now the service provider lowers the target utility down to 1000, and wants to know what are required capacities for different approaches. Now the figures are more realistic. For instance, a moderate increase of capacity from 2000 to 3000 might well be enough to make certain that reputation of the service provider is not lost. Note particularly, that with packet level prioritization, the additional video calls do not have any significant effect on the voice quality.

Table 3. *Required Capacity for U_{tot} = 1000, E[N_{video}] = 50*

	ε(N_{video})=0.5	ε(N_{video})=1.0	ε(N_{video})=2.0
Best effort	2700	4350	8650
Packet level priority	1980	2620	4850
Reservation, equal call blocking	2250	2740	3920

What could be the main conclusions of this example? First, if there are significant utility differences between calls, a reservation system with equal call blocking is not an efficient way to maximise total utility. Secondly, reservation may make sense if, and only if, the overall demand predictions are very inaccurate whereas the bit rate requirement of each individual call is known accurately. But even then it is not clear that reservation is better if all aspects, related implementation cost, realistic user behaviour etc. are taken into account.

Finally some words about data applications. The previous evaluation considered only applications, like voice and video, that are the most suitable ones for reservations, In contrast it is almost impossible to find any good reason to use reservation principle inside the network with any data application. Thus a large amount of data traffic probably decreases the already minor attraction of reservation principle compared to packet level prioritization. On the reverse, a combination of important, relatively constant traffic (e.g. voice) and highly variable data traffic is a good reason to use packet level prioritization instead of best effort service.

Even a small minority of users with high-speed access can fill a part of network by downloading, for instance, music or video files. The utility value per byte for downloading a video is really minimal compared to a voice connection; the ratio could easily be 1:100. Fortunately, this problem can be solved by a simple prioritizing system that treats every user in the same manner, but takes into account the bit rate used by each user. Now if a user wants to download a huge file, he is totally free to try. However, if the momentary bit rate is higher than with most of the other users, the network will start to drop his packets instead of voice, and other important, packets. There is no technical reason to use reservations to protect voice connections, or to limit the traffic of data flows.

===

Appendix

This appendix illustrates the effect of ε(N_{video}) on the probability of high traffic demand. The attached figure shows the probability that the number of video users exceeds certain value x for ε(N_{video}) = 0.5 and ε(N_{video}) = 2.0 when the average number of video calls, E(N_{video}) = 50. The tail of the distribution is actually what often determines the overall performance of the system, because most of the time there is enough capacity to handle all traffic demand. For instance, with ε(N_{video}) = 0.5, the number of video calls exceeds 187 by probability of 0.1% whereas with ε(N_{video}) = 2, the number of video calls exceeds 1055 by the same probability, though the average value for N_{video} is the same.

We may even argue that from the system performance viewpoint the most important area is in the region where the probabilities are from 0.1% to 1% as illustrated in the figure. On the one hand, if this region is handled appropriately it is very likely that the smaller traffic values can be handled without problems. On the other hand, it could be too costly to optimise the system for very rare cases with very high traffic demand.

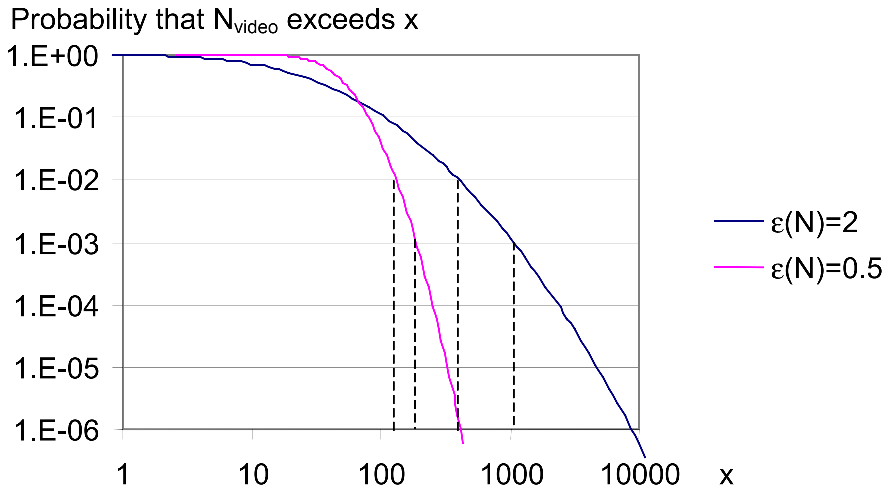

Fig. Log-normal distribution with mean of 50