Material for lecture #5

Usefulness of resource reservations

This document discusses the usefulness of resource reservations in packet networks. Let us start with a list of possible reasons for resource reservation (this is based mainly on discussion during the lecture 17.10):

1. Applications' need of minimum bandwidth
2. Virtual Private Network or Leased Lines
3. Pricing model
4. Need to know service availability beforehand
5. Emergency calls
6. Need to limit traffic sent into the network
7. Technological reasons (reservations may improve the use of network resources in some specific cases)
8. The need to favour existing connections over new ones

In general, the viewpoint of this document is that whether real reservations are used inside the network is operator's own concern. Thus even though a resource reservation may appear to be the best choice for the operator, if a better result is achieved by an alternative method without using any reservation, the operator is free to use that method. In other words, the goal (e.g., better service) and tools to achieve the goal (e.g., reservations) are separate matters, and should not be mixed up.

Here I mean by real reservation a system in which the network reserves the required capacity (essentially a constant bit rate capacity) in every node and on every link through the network in a way that the sum of the reserved capacities always remains below the capacity of the node or link. The main advantage of this simple system compared to more advanced reservation systems is that it can likely be automated quite well without any human intervention during normal operation conditions.

It is definitely possible to also use statistical reservations, but then we must be aware of the entailing problems, mainly related to the management of the whole system. The degree of statistical multiplexing that can be exploited in real environment is an intricate issue and the required control system will be extremely hard to automate in a way that human interactions can be totally avoided.

Applications' need of minimum bandwidth

Although minimum bandwidth may, indeed, be an appropriate reason for reservations it is not clear that reservation is only possible tool to achieve the goal - but what is the fundamental goal? This issue about goals and tools is discussed further in the next section, but before that let us evaluate the other possible reasoning for resource reservations.

Virtual Private Network or Leased Lines

This indeed might be a valid justification for reservations, because the actual contract may include a mention of reserved capacity. However, this kind of wholesale service is not the actual topic of this document (note that the previous lecture #4 were partly related to this issue).

Pricing model

Once again, what is the point that may justify reservation by itself? The pricing model selected by an operator can, surely, be based on the idea of reserved capacity. The operator can, for instance, believe that the price of a voice call is directly determined by the amount of network resources needed to realise and provide the service. In that case, the operator may think that actual reservations are necessary because of that relationship. This appealing but somewhat peculiar line of thought is based on the vague idea that there is no resource without reservation.

However, a router has a certain packet transmission capacity independent of any reservation. Actually if there is a relationship between reservations and performance, the situation could be somehow opposite. The reservation process may use the same processing capacity as the actual packet forwarding, and might therefore deteriorate the system performance. As to the original pricing issue, the operator may think that this only strengthens the connection between pricing and reservations: because reservations are costly, the pricing should be based on the amount of reservations, and if that is true, reservations should be used, because that is exactly what customers are paying for!

Nonetheless, this loop leaves open the question whether the whole pricing model is reasonable from business point of view, since if a good enough service can be provided without reservation and corresponding pricing model, that alternative could be better for the operator. A less clear issue is whether reservation are necessary just because customer suppose that there are real reservations (once again, this belief may stem from the pricing model) or if customer are satisfied if the service is good enough independent of the implementation. I am tend to favour the latter option.

Need to know service availability beforehand

This issue is closely related to the minimum bandwidth item. The additional point is that in some cases it might be necessary to know the availability of the service long before the actual service event (as an analogy, you want to reserve a movie ticket beforehand in order to avoid unnecessary travelling). In this situation it is considerably better to get the call blocking signal as early as possible. On the other hand, the implementation of this feature in a packet network could be very laborious and expensive compared to the real need for the feature. Appropriate priorisation could be a better approach in real environment, although it cannot provide precisely the same characteristics.

Emergency calls

Because the need of emergency calls cannot be predicted, there is not much sense to make capacity reservations in advance. Thus these calls must be handled in the best possible way when they emerge. The question is what is the best way? Call blocking evidently is at least as bad a situation as a connection with intermittent interruptions. Thus it is somewhat hard to identify any reason why reservation could be considered better than appropriate priorisation. If packets belonging to emergency calls have higher priority than other packets, the result should be appropriate without any reservation. However, this is true only on the condition that the emergency traffic is under tight control (note that policy control is not the same thing as resource reservation).

Need to limit traffic sent into the network

In short, it is possible and often useful to limit the total traffic sent into a network, but that does not imply that reservation is a compulsory tool inside the network. A more valid approach is to analyse the needs to limit the traffic and they decide whether those needs require the use of reservations.

Technological reasons

It indeed is true that reservations may improve the use of network resources in some specific cases. If a radio channel is divided among a large number of terminals, the use of radio resource can be done more efficiently, if the network divides the radio capacity semi-permanently among the terminals. Particularly, this might be a practical solution, if the traffic process is constant enough, whereas reservations are always problematic with highly variable traffic process (most of the time, the reservation is either too small or too large).

The need to favour existing connections over new ones

This item is also related to the minimum bandwidth issue with the extra aspect that the conventional reservation guarantees that the service quality will remain the same until the customer finishes the call. It seems that this feature is quite difficult to implement with a simple priorisation system. However, it is somewhat questionable to assume the priorisation of existing calls is always a desirable result, in particular, if the pricing is based on flat rate. Still, this might remain relevant point and should be considered carefully.

Remarks

Most of these issues stem from the basic question about the fundamental nature of reservations in a system that is based on packet delivery. First of all, a capacity reservation for an individual packet does not make any sense, simply because of the huge overhead (the reservation probably consumes more resources than the original packet, and if the reservation is made by means of packets, the reservation should evidently concern those reservation packets, too). Thus the actual target of reservation always is an aggregate of packets rather than one packet. Yet a router handles one packet per time, and if the reservation has any effect, it should somehow be manifested during the handling process of individual packets. Apparently, the main issue is whether the packet is accepted or not; a packet with reservation should be accepted instead of any packet without appropriate reservation. But as discussed earlier, individual packets have no reservation so there must be another way to distinguish the packets belonging to an aggregate with reservation from other packets.

One essential question is how to determine which packets belong to the reservation of an aggregate if and when the packet arrival process is not totally deterministic. If all packets arrive through the same entrance point, it is relatively easy to build a traffic control system that marks packets in a way that packets complying the reservation are marked as highly important while non-complying packets are either immediately discarded or marked as less important (we would say that discarded packets are of the lowest importance, by definition).

But then we may ask what is the difference between a reservation and a priorisation system? The right answer is that the reservation checks the whole path of the traffic aggregate whether there is enough capacity, whereas priorisation *can be done* without any sufficiency check. The capacity check removes the need of marking packets accepted into the network with different importance, because the strong

assumption is that there is no need to discard any of those packets inside the network.

Utility of resource reservations

This section considers the usefulness of resource reservations from the viewpoint of an individual user or an application, and is based mainly on the assumption that the minimum bandwidth required by certain applications is the most important reason to use resource reservations. Regardless of the limited, one user viewpoint, it is necessary to make some reasonable assumptions about traffic in general. So let us use the following model:

- there are a (large) number of similar active users (N)
- there is a bottleneck link (or node) with a fixed capacity (C)
- each user generates similar traffic
- the unit of resource (e.g., bandwidth) is the average traffic generated by a user (a = 1)
- both the number of active users and the momentary traffic generated by a user are random variables with certain coefficient of variation
  - $\varepsilon(x) = \frac{stdev(x)}{E(x)}$
    - ε(N) depicts the variability in the number of active users
    - ε(a) depicts the variability of traffic generated by one user

Both ε(N) and ε(a) are basically unpredictable quantities in the meaning that when a reservation is made or the bottleneck capacity is determined the exact values are not known, only the probability distribution is assumed to known (even that assumption is quite optimistic). In this compact analysis both distributions are assumed to be log-normal.

Because the evaluation is made by means of utility, we have to fix the utility figures for the basic cases:

- utility (per time unit) for a successful connection = 1
- utility (per time unit) for an attempt that gets busy tone, u_B = -5
- utility (per time unit) for a situation in which there is no busy tone but still the available bandwidth is not sufficient, u_Q = -20

What do these figures actually mean? If the call blocking (the probability of busy tone) is 1/6, the average utility for a user is 0, and for an ordinary call blocking of 1% the average utility is 0.94. Then if there is no call blocking at all, but the available bandwidth is insufficient about 3 seconds per minute, the average utility is also 0.

As a result, if we know two factors, call blocking (B) and the probability that the available bandwidth is insufficient (Q), we can calculate the average utility for a user:

$$U = (1 - B) + B \cdot u_B + (1 - B) \cdot Q \cdot u_Q$$

One interesting, but hard, question is what could be the relationship between this average utility and other methods to measure user satisfaction, like Mean Opinion Score (MOS). Table 1 shows that the guess of the author of this document (here I have thought a VoIP, but it is up to the reader to consider the validity of the model in case of other applications). Note that from this evaluation point of view, the main matter is the relationship between 3rd and 4th columns, e.g., whether a 5% call blocking is as annoying as 2.7 s interruption within a 3 minute call.

Table 1.

MOS	U	B (%)	error seconds per 3 min call
5	0.98	0.33	0.18
4	0.92	1.33	0.72
3	0.7	5.00	2.7
2	0	16.67	9
1	-3	66.67	36

Now let us try to compare two fundamentally differing systems: a reservation and a sharing system.

1. In a *reservation system* each user reserves capacity (w) that is sufficient with high enough probability for the application user is willing to use. For instance if this probability is 0.1% it means that the user reserves bandwidth that satisfies the following equation

   $$\Pr(a > w) = 0.001$$

   This assumption limits the average utility to 1 - 0.001*20 = 0.98 independent of the total capacity (which is supposed to be high enough utility from end-user viewpoint if we rely on the table 1).

   In addition average utility is limited by the call blocking. Call blocking is a function of number of active users (N), the variability of ε(N) and the total capacity to reserved capacity ratio (C/w). The blocking formula used here is described in fig. 1. The original offered traffic distribution (blue line) is cut at the point C/w, and the excess part of the distribution (yellow area) is shift to the point of C/w (red pole). The call blocking here is assumed to be decrease of served users due to this shift. The author is aware that this model produces very optimistic (low) blocking values, for instance, compared with ordinary Erlang loss formula. Note that we cannot apply Erlang loss formula because it is applicable only if when $\varepsilon(N) = 1/\sqrt{N}$.

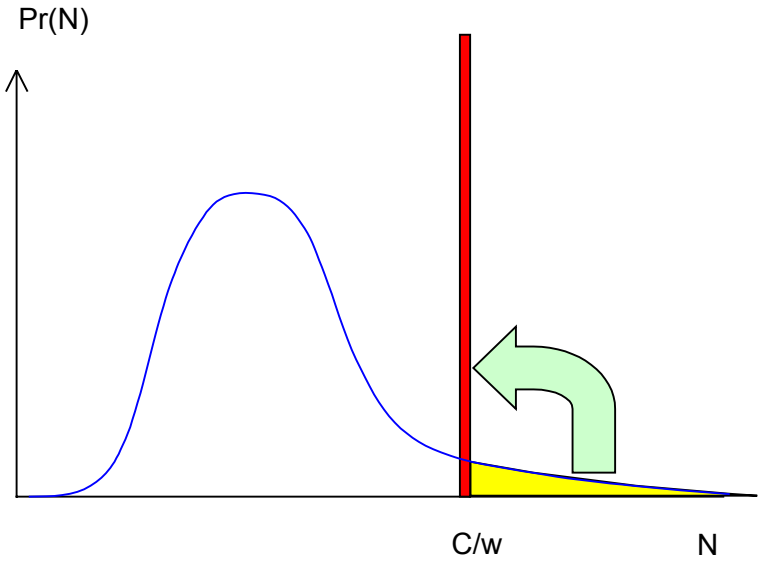

Fig. 1. Traffic distribution and a method for approximating call blocking

2. *A sharing system* in which the total capacity of every node and link is always divided evenly among all active users. In this system there is no call blocking but, instead, the available capacity can be insufficient for all users in the worst case. The probability that the capacity is insufficient for a given number of active users (N) can be simply calculated from the probability distribution:

$$Q(N) = \int_{C/N}^{\infty} \Pr(a = x)dx$$

and the average insufficiency probability by a weighted average:

$$Q = \sum_{i} Q(i) \cdot \Pr(N = i)$$

Note that this sharing system is supposed to be very crude in the sense that it does not exploit the unused capacity of those flows that need less capacity than the equal share, C/N. Moreover, it is assumed that applications are very loss sensitive and without any possibility to adapt into varying bandwidth.

Now let us evaluate how the average utilities of these two systems behave with different traffic parameters and capacity values. In general, it is quite apparent that the reservation system works better (that is, generates higher average utility) if there is a real lack of capacity, because with the sharing system all users suffer while in the reservation system some users still gets adequate service. On the contrast, if there is a lot of extra capacity, it is reasonable to assume that the sharing system is better because reservation system cannot exceed a certain utility limit (0.98 in our case). Therefore, it should be possible to find a threshold capacity, C_e, in a way that

if the capacity is less than C_e, reservation system is better, and

if the capacity is higher than C_e, sharing system is better

The results of a brief evaluation are presented in table 2. The main conclusion is that if $\varepsilon(a)$ is smaller than 10%, reservation seems to be reasonable because it performs better also with high utility (=low call blocking, high MOS). In contrast, if $\varepsilon(a)$ is larger

than 15%, it is quite difficult to identify any good reason for resource reservations, because its performance is superior compared with sharing system only with low average utility. The results shown in table # are calculated only for $\varepsilon(N) = 0.2$. However, the essential result - which one of the two systems is better - does not depend significantly on $\varepsilon(N)$. Only if $\varepsilon(N) << 0.1$, the results are more favourable for sharing system, apparently because then there is a smaller risk that the offered load exceeds the capacity of the system.

One point should be stressed when assess this results:

> All average utility values include also those cases in which the sharing system works poorly because all users suffer from the lack of resources. If the operator is afraid that the number of active user could occasionally be high, and it still wants to offer service with high quality, the only feasible way to achieve this is to increase capacity, because call blocking also is a serious deterioration of service.

Table 2. Capacity figure selected in a way that reservation and sharing system generates equal average utility, $\varepsilon(N) = 0.2$

$\varepsilon(a)$	C/E(N)	B %	U(res) = U(sha)	MOS - a guess
0.01	1.81	0.02	0.979	5
0.05	1.81	0.11	0.974	5
0.10	1.76	0.98	0.922	4
0.15	1.60	7.10	0.555	3
0.20	1.49	18.9	-0.15	2
0.30	1.42	40.1	-1.41	1.5
0.50	1.36	64.6	-2.88	1

A further illustration is provided in figures 2 and 3. Now we fix, in addition to $\varepsilon(N)$, also $\varepsilon(a)$ and look how the utility figures change when the capacity is changed. For $\varepsilon(a) = 0.1$ (fig. #) the message is that while sharing is somewhat better with high enough capacity, $C > 1.7\ E(N)$ in our example, reservation system is clearly better there is not enough capacity. This situation changes drastically when the traffic per user is known less accurately. For $\varepsilon(a) = 0.3$ sharing system works better even with unacceptably low utility levels.

From practical viewpoint the conclusion seems to be that the operator should know how accurately users could know beforehand their capacity needs. Only if that knowledge is accurate enough resource reservations seems to offer significant advantage from total utility viewpoint. Otherwise sharing system appears to be better, and then the operator have to identify some other significant reason in order to justify resource reservations.

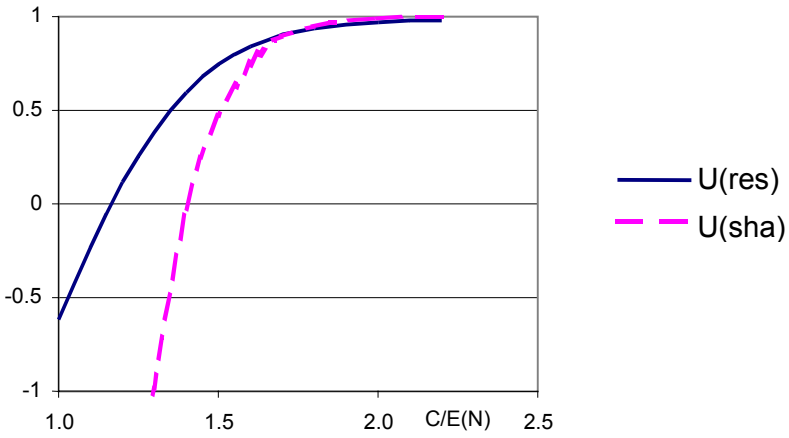

Fig. 2. Utility as a function of capacity fir ε(a) = 0.1 and ε(N) = 0.2.

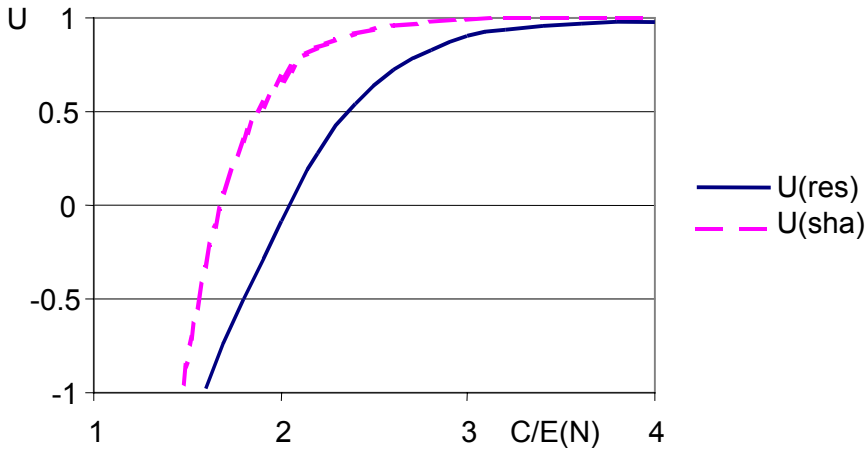

Fig. 3. Utility as a function of capacity fir ε(a) = 0.3 and ε(N) = 0.2.

Two critical issues related to this very preliminary evaluation are the following (the consideration of other issues is left for the reader):

- Users are not similar but they may have significant differences; some users can predict their needs better and the utility and bandwidth requirements depend on the application. However, we may still argue that differences are more about priorisation than reservation issue.

- If the reservation system can, in spite of the doubts we discussed earlier, exploit the network resources better than what a simple addition of requirements -approach can do, the result could be more favourable for the reservation system. However, then an appropriate evaluation have to consider very carefully the implementation and management issues related to that kind of system.

Note that both systems are equally resistant to any excessive traffic sent by a user. The only difference is that in the reservation system excessive packets are discarded immediately while in the sharing system excessive packets are discarded only if the traffic exceeds the momentary equal share on a link. In both cases, if the number of active users is given, the service level of one customer is independent of the traffic

sent by other users. On the other hand, the number of active users has a similar effect on the service in both systems, even though we may argue that the effect is somewhat smoother with the reservation system from the average utility viewpoint (while the reverse appears to be true from individual user viewpoint!).