Material for lecture #10 & some final considerations

1. RESERVATION OVER RADIO INTERFACE

Here we discuss some issues related to QoS in mobile networks. We will not discuss any particular technology but instead, we try to evaluate some QoS aspects by using an abstract model for mobile environment.

The main special property from QoS viewpoint is the control of up-link capacity

- the resource is essentially shared (though someone may argue that in communication networks resources are always shared in a manner or another)
- unreliable media, which means varying and hardly predictable capacity
- distributed queues on sender side
- limited information about queue states in the receiver side

Several technologies and access control methods can be used to manage traffic in this kind of situation. One possibility could be to use the principle used in Ethernet: Carrier Sense Multiple Access with Collision Detection (CSMA/CD). Unfortunately this type of solution is not applicable without modifications because the collision detection is not as simple task as in a fixed environment. First, two terminals with a valid connection to the base station do not necessarily hear each other, which means that they can try to send data at the same time to the base station without noticing the collision. Further, the maximum transmission delay is not controlled as well as in a cable. Even if some nice modification of the CSMA/CD could be used over radio interface, it seems likely that the efficiency of the system were not as high as desired.

Therefore, we assume here that the receiver side (base station or another entity within the network) strictly controls the data sent by each terminal. The way to do this might be depicted in a way that time is divided into limited periods (for instance 10 ms) and the traffic control unit divides the capacity during the period among all active terminals. Let us call the period here as *cycle*, and the smallest amount of capacity that can be given for a terminal as *cell*. The capacity of the system is basically defined by the amount of data that can be transmitted within a cycle. In addition, part of the total capacity must be allocated

- to transfer the necessary information about the states of the terminals to base station
- to transfer information about the permissions to sent data

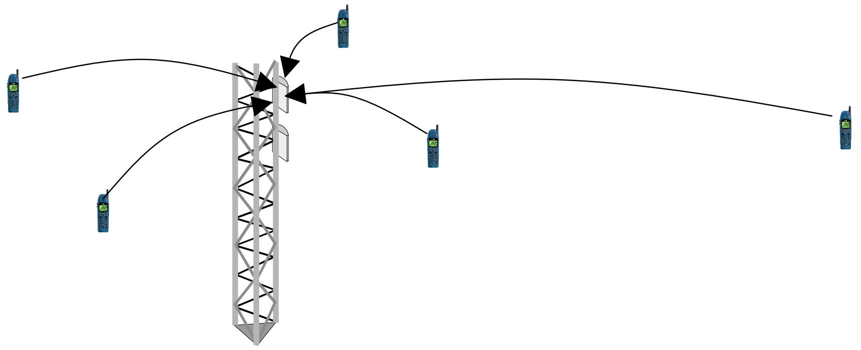

Fig. 1. Base station and mobile terminals

From the viewpoint of useful data these operations just consumes expensive radio capacity, and therefore the methods used to decide how much information is transferred and when should be optimized as well as possible.

The momentary capacity is not necessarily known exactly in advance, and it may depend on the location of terminals. In this paper these complications are mostly ignored so that we can concentrate on one of the problematic issues, that is, how the base station can optimally divide the radio capacity using the limited information that it has about the states of terminals. But what is the essence of the state of the terminal? From QoS viewpoint, the evident answer is the state of traffic process in each terminal. Basically, a simple model shown in fig. 2 can be used. The traffic process generated by a terminal is divided into two layers: flow and burst. The meaning of these terms is quite clear with certain applications, like voice, whereas with many applications, like web browsing, the situation is less clear. Anyway, we think may that the start of a flow (instant A in Fig. 2) is characterized by a so long idle period before instant A that the traffic control unit considers the terminal as silent. In practice, this may mean that the terminal has to in a special manner to inform the traffic control unit that it has become active and wants to send some data, for instance because the user wants to start a voice call, or starts a web-browsing session.

The other layer describes the activity variations during a call, for instance, during a voice call either of the two participants is usually idle. Similarly, during a web-session file transfers alternate with silent periods without transmission needs.

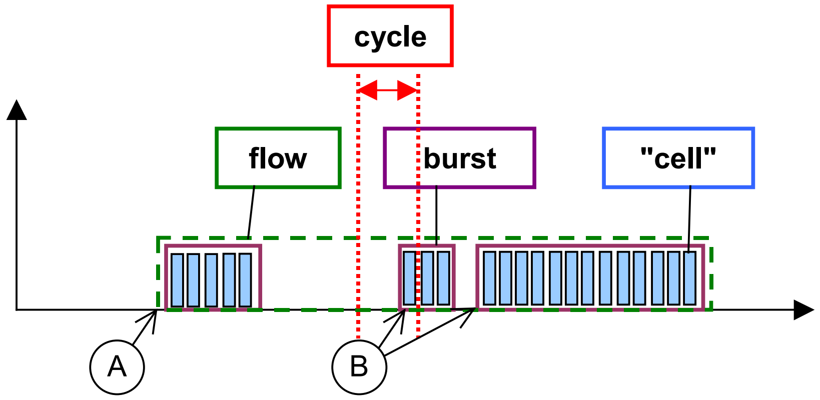

Figure. 2. Layers of activity

From technical perspective the whole issue of layers is essentially related to the capability to predict traffic process. A state "flow-on" means that the probability of starting a burst is significantly higher than if the state is "flow-off". This additional information can be used to optimize the use of resources. Similarly, the state "burst-on" shall mean a high probability that the user has something to send during next cycle.

This kind of system can perhaps be modeled by assuming some conditional probabilities:

Pr(flow k has traffic demand of x cells in cycle i | for a given past traffic pattern)

In principle the question is "What is the optimal strategy for a given probability". Or actually the optimal strategy should take into account a longer period than one cycle. Now we can approach the question what kind of reservations are used in the optimal strategy. As a result we may be inclined to say that reservation is necessary because of a technical reason if the optimal strategy includes the use of reservation. On the contrary, if an exhaustive analysis bring about a result that reservations do not improve the utility from service provider viewpoint, we should not claim that reservations are necessary although they, indeed, are used by most radio technologies.

The following pages provide a very tentative (far from exhaustive) analysis of the situation. Let us assume that there are basically 4 scenarios:

- CBR reservation for flows

  Apparently most efficient with true CBR traffic

- VBR reservation for flows & CBR reservation for burst

  More efficient with on/off traffic but who can know traffic parameters

  Burst reservation consumes some resources

- CBR reservation for bursts without flow reservation

  If burst are truly CBR, potentially efficient

  Risk of overload on flow level

- no call or burst reservations

  This is the most efficient multiplexing if these (per cycle) reservations were free of charge, but in practice this probably requires a lot of extra transmission of information.

Let us try to make a similar analysis as in the earlier cases and assume some traffic, system and utility characteristics:

Capacity	2 Mbps	
cell	50 bytes	(= 0.2 ms in a time-division system)
cycle	50 cells	= 10 ms

Note that the length of the cycle should not be too long because of the delay requirements of real-time applications

active time (burst)	0.6 s	(60 cells)
idle time	0.4 s	
peak bit rate	40 kbps	(bit rate during burst = 1 cell/cycle)

Note that the real useful bit rate for the application can be much smaller

average bit rate	24 kbps	(average bit rate of a flow)
U_S (transmitted cell)	1 per cell	
U_B (blocked flow)	-5 per cell	

U_L (arbitrarily lost cell)	-20 per cell

Once again, the only significant issue from the final result viewpoint is the ratio $(U_S - U_B)/(U_S-U_L)$

In addition to these basic parameters we have to make some assumptions concerning the overhead of different reservations:

CBR reservation for flow	4%	(2/50 cells)
VBR reservation for flow	9%	(3/50 cells + 2 cells/burst)
Burst reservation	8%	(5 cells/burst)
Reservation for each cycle	33%	(0.5 cells / transmitted cell)

The effects of different reservations on the total utility are

CBR reservation	Blocked flows based on peak rate Call blocking (B) based on Erlang model Utility = 1 - $B*U_B$
VBR reservation	Blocked flows based on effective bandwidth = 30 kbps Call blocking based on Erlang model (U_B) (burst blocking is so small that it can be ignored here)
Burst reservation	Lost bursts (U_L)
Reservation for each cycle	Lost cells (U_L)

This brief evaluation assumes that flows are coming according to Poisson process and that the burst and idle periods within a flow are exponentially distributed. Under these assumptions we get the result shown in Fig. 3.

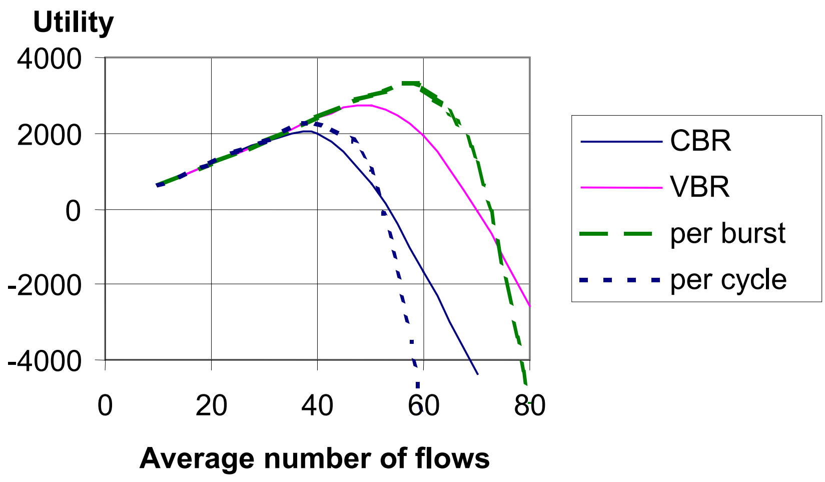

Fig. 3 Utility curves for different type of reservations

Theoretically the maximum capacity of the system is 83 flows but because of the statistical variations on flow and burst scales and all various overheads, the maximum utility is achieved by much lower average traffic. With CBR reservation 35 flows produces total utility of 2015 whereas a more advanced (=complex) VBR reservation is able to produce total utility of 2750 with average traffic of 50 flows. Although this difference seems very convincing, it is not at all clear that VBR reservations are as efficient in practice because most applications cannot offer any reliable information about the average traffic.

Then if we totally ignore the flow scale reservations and do reservations only on the burst scale, we appear to get the best result. With average traffic of 55 flows, the total utility is 3200, that is, 58% higher than the total utility of CBR reservations. However, we shall be very cautious with our conclusions because we do not know whether our assumptions related to the overheads are valid in real networks. Nevertheless, the difference in total utility is so large that it cannot be totally ignored.

Finally, if the reservations are made purely for each individual cycle without any information about the traffic on previous cycles, the result is naturally worse. Anyway, it is interesting to notice that even an overhead of 33% in this case produces essentially the same total utility as the CBR reservation on burst scale. So once again, the real merits of reservations for flows appear to be quite small even when for a VoIP type of application.

2. ONCE MORE ABOUT PRIORITIZATION VS. RESERVATION

As one of the main matters of the whole course has been the comparison between two systems, reservation on connection scale and prioritization on packet scale, I will provide one more example related to this issue. The special feature of this example is that it takes into account the high probability that an unsuccessful call attempt generates new attempts. From utility perspective the following assumptions are made:

- utility of a successful call = 1
- utility of an unsuccessful attempt = -0.1
- utility when the customer finally gives up = -5

The primary characteristics of the traffic model are:

- traffic variations are divided into two time scales
  - long term variations (mean value) are log-normally distributed, standard deviation per mean traffic = 0.5
  - short term variations that occur around the mean value are Poisson distributed.
- an unsuccessful call attempt recurs with a probability of 0.8.
- the traffic generated by one user is exactly CBR in a way that there is no need to reserve any extra capacity due to the traffic variations within one connection.
- the capacity of the system is 100 times the bit rate needed by one connection.

Because of the long-term traffic variations, the average load level has to be quite low even with a perfect reservation system. For instance, if long term average traffic is 40 we obtain the following performance figures (these are average values over a long period):

- probability that the first call attempt is discarded = 2.39%
- probability that a customer gives up before getting service = 0.86%
- average number of re-attempts per first attempt = 0.034
- average utility per an original call attempt = 0.954

Because the traffic and system assumptions were very favorable for the reservation model, it is obvious that a pure packet based, best effort system cannot provide as good service as the reservation system. This assumption is strengthened by a bare calculation. For instance, if we make the rough assumption that the utility per time unit is -20 for *all connections* if the momentary traffic exceeds the system capacity, the average utility is only 0.007 even for average load of 40.

This, indeed, seems to offer valid reasoning for resource reservation. In fact, if the service were based purely on best effort packet forwarding, there should be 70% more capacity in order to attain the same level of utility as with the reservation model. We have discussed earlier the validity of the assumptions and noticed that very few real situations are as favorable for reservations as in the model used here - for instance, the bit rate of a packet based application seldom is exactly constant, and the establishment and maintaining of reservations tend to consume some resources. However, if we put aside these issues, there is still an additional issue that may be worth of consideration, namely, the utility differences between connections. For instance, let us evaluate the above case under the assumptions shown in table.

	share of the traffic	utility of successful call	utility per time unit for bad service
class 1	1/3	+2	-58
class 2	2/3	+0.5	-1
average		1	-20

Class 1 represents an application with high QoS requirements and users with high quality expectations, whereas class 2 represents users with much lower expectations and readiness to pay. We can assume that if the system takes into account the differences between the classes, the overall utility can be higher. So let as assume that the system simply classifies packets into two priority classes depending on the customer class, and serves the lower class (2) only if there is free capacity left by the higher class (1). This can be done by very simple mechanisms on the packet scale without any admission control on the connection scale. But is the result good enough compared to the reservation system?

With average traffic of 40, the average utility is 0.952, which is quite exactly the same as with the reservation system. Somewhat surprisingly, the situation remains similar if we increase the traffic load until the average traffic exceeds the system capacity by 25%. After that point the advantages of reservations are clear even compared with the prioritization approach. However, then the utility already is significantly below zero, which means that the user satisfaction is either poor or very poor. The utility curves are shown in Figure 4.

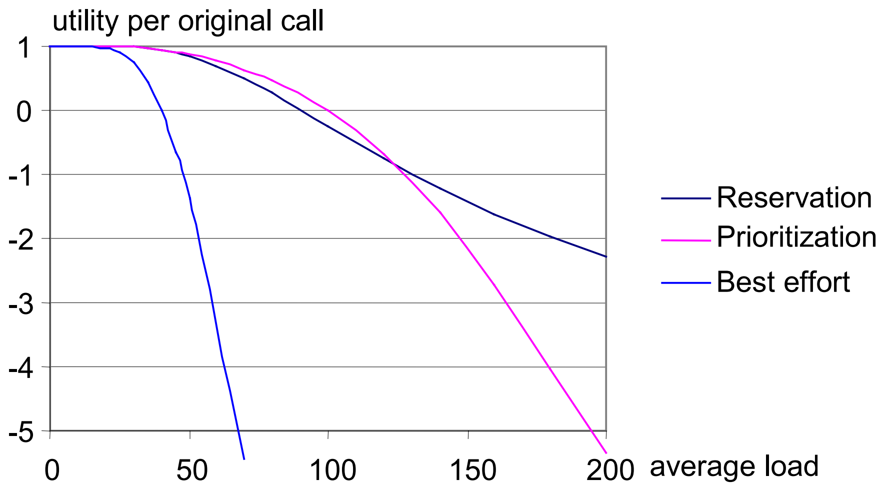

Figure 4. Reservation vs. prioritization on packet level

Finally, what was the effect of re-attempts on the general figure? The answer is, hardly noticeable. With low load and high quality, the effect of re-attempts is lightly positive (0.933 vs. 0.945 with A = 40) whereas with high load re-attempts actually have a negative effect on the utility (-2.11 vs. -2.28, on the condition that the number of re-attempts is not correlated with the utility of the call).

3. DOCOMOS COMMENT ON3G SERVICES

As a final comment to the whole course, let us consider the statement made by K. Enoki of DoCoMo.

http://news.ft.com/ft/gx.cgi/ftc?pagename=View&c=Article&cid=FT3H3CWOVFC&live=true&tagid=IXLC078IH7C&Collid=Any

"DoCoMo sounds alarm on 3G"
By Dan Roberts and Michiyo Nakamoto in Toyko
Published: November 22 2000 21:39GMT | Last Updated: November 23 2000

> Third-generation mobile networks may not provide the revenue growth many European telecoms companies are counting on, according to NTT DoCoMo, the Japanese phone operator that pioneered mobile internet access.
>
> Keiichi Enoki, who runs DoCoMo's successful i-mode internet service, says operators will struggle to justify the more than E100bn ($85bn) they have spent on 3G licences in Europe.

"I don't think the business model will fundamentally change from 2G to 3G. The essence of the cellular phone business will be the same," said Mr Enoki in an interview. DoCoMo, which is testing 3G technology, is finding that it is unsuitable for carrying large video or sound clips, one of the services that could provide new revenue streams for mobile operators.

The new technology provides faster data speeds than 2G, allowing colour video and high-quality music to be sent to mobile handsets. Sustained bursts of multimedia data consume large amounts of the radio spectrum and DoCoMo says it will be too costly to download large files, such as pop videos, to handsets.

> Mr Enoki said: "The conclusion is that we will perhaps offer short video clips of 10 to 15 seconds and previews of music that people can purchase to download at home through their PCs or TVs."
>
> Mr Enoki said operators could best hope to increase revenue by charging commission on these transactions as well as from advertising.
>
> Japan will be the first country to offer 3G services next year. Its experience contrasts with the approach of operators such as Vodafone in the UK, which plans to download songs over mobile networks. DoCoMo says phones will be used to order multimedia content, which will be delivered over other, more cost-effective, telecoms networks.
>
> DoCoMo's pessimism about the potential of 3G technology could damage confidence in Europe just as investors are questioning the huge levels of debt generated by set-up costs.

If Enoki is right when claiming that 3G will be able to transport relatively short video clips but not real-time video with long duration - and I am tend to agree with Enoki - the conclusion evidently is that great majority of the 3G traffic will be some kind of file transfer rather than real-time traffic. Nevertheless, ordinary voice calls have to work at least as well as in the current mobile networks, or preferable better.

The utility of a clip of 15 seconds is, of course, much smaller than that of the whole video. However, there are at least three issues that favor a video clip service for 3G terminals. First, the utility (measured as the readiness to pay) is much higher for mobile terminals than for fixed access (compare SMS and e-mails in fixed networks). Secondly, the utility per byte for a short clip is higher than that of the whole video, at least if the clip is designed appropriately. Thirdly, a clip may well function as advertisement for other services or products (e.g. music videos on DVDs) in a way that the short clips can even be offered free of charge. These issues may together mean that even though the offering of whole videos for mobile terminals is not a promising business as such, shorter clips may provide some potential income.

==

The End

Kalevi Kilkki

30.11.2000