Congestion Control in 4G Mobile Networks
A Business Imperative

Author
Vivekanand Tiwary
Practice Head, Tech Mahindra
**Introduction**

With LTE rollout by operators across the globe, world is moving towards high bandwidth era. But if we look at the trend and forecast of the mobile data traffic, we should still be worried on disproportionate demand and supply aspect.

Data traffic globally is forecasted to reach 16 Exabyte by 2018. Please refer the figure below.

Therefore, it is safe to conclude from the above illustration is that it would be difficult to provide an “enough” network capacity for users. Hence, operator’s network capacity and demand need to go hand in hand. Operators typically plan their network capacity considering the business value. They want their average traffic to be as close to the capacity as possible. Congestion is unavoidable in optimally designed mobile networks.

In this paper, we are proposing an option to handle short duration spike of congestion which will help operators to maintain accessibility.
The challenges

The term 'congestion' in this document applies to a situation where the problems of congestion continue to persist even after all the congestion management mechanisms are applied upon – for example traffic shaping, policing, preferential treatment etc. In such a scenario, there is only option to decline new session request at the front door, as long as congestion persists. The idea is to provide good quality of service to smaller percentage of customer instead of degraded quality to all.

The diagram below depicts nature of congestion and its impact both in controlled or uncontrolled cases.

In the above case, quality and throughput both get degraded after “Congestive Collapse” point if not controlled. The diagram below shows the system behavior with congestion control in place. That means the system declines all the new service requests as long as the congestion persists.

With this approach, by declining new service request at the front door system we will be able to maintain the quality and throughput close to the threshold.
There are two types of congestion. The first one is practically persistent during peak hours, may be predictive and may result in loss of substantial amount of revenue if not resolved. This type of congestion need capacity enhancement.

The second type of congestion is unpredictable, persists for short duration and impact on revenue may be negligible. This type of congestion can't justify capacity enhancement. Some of the reasons of this type of congestion can be as follows:

1. Flash crowd
2. Arrival of trains at railway station. Depending upon the source location of the train a particular operator may face congestion
3. Arrival of a cruise to a tourist destination. Operator may face congestion depending upon the source location of the cruise
4. Public rally
5. Accident site
6. Local equipment failure
7. Others

A typical characteristic of transient congestion is depicted in the diagram below.

Let us assume that this type of congestion occurs on an average 30 minutes per month for the operator. That means that all new call/session requests during this period are declined. Although the call attempts during peak hours are higher, let us consider case of an average call attempts rate. One month equals to 30days x 24 hours x 60 minutes that is 43200 minutes. Thus declining 30 minutes per month translate to 0.069% thereby degrading accessibility to 99.931%. This is a substantial degradation when we talk about 'five nines' quality.
Option of last resort

In urban areas we generally have 5-10 mobile operators providing complete service coverage. It is obvious that such type of transient congestion is not likely to occur simultaneously in every operator's network.

In the above diagram the traffic shown as shaded portion can be catered by the other operators. When there are many operators in urban areas, it is always likely that one or more network is running at low traffic while the other is facing a spike in congestion.

In this scenario, we suggest that the operators should offload subscribers requesting new sessions during transient congestion to another operator who might be running much below their capacity. This can be done whether congestion is in the access network or in the core leveraging existing congestion detection mechanism. The concept is similar to roaming with the following differences:

1. In national roaming, a subscriber attaches to another PLMN only if home network is not available. In our case, it needs to attach to another network even if its home network coverage is available.
2. UE selects the PLMN based on priority order stored in its SIM. In our case, Home network decides the PLMN based on situation prevailing at that point of time.
3. Home network instructs a UE to attach to another PLMN per call basis or per congestion duration. One user can be catered by operator B while another UE by operator C depending upon the situation.

How

The following call flow depicts a typical MO session while its home network is congested. The steps used and the message names are only indicative in order to explain the concept and it is not an implementable call flow.

During congestion, if a user wants to initiate a session, the network sends an offload request to all of its partners (message 3 in the diagram). Depending upon their network condition the partner operators sends ACK or NACK. The operators who are running below their capacity should send ACK. By sending ACK the partner operators indicate that they have sufficient capacity to cater this service. The MME may get more than one operators who are willing to cater this service. Based on its preference and QoS indicator of the partners, MME selects one of them and sends cancel message to others.
It sends “Initiating Offload” to the chosen operator (message 7). It provides the TA (tracking area) of the UE in this message.

It sends handover command to UE through eNodeB, instructing it to detach from the home network and attach to another PLMN whose Id is provided in the message.

This point onward the procedure will be exactly same as national roaming.

After availing the service when UE goes to idle mode and congestion to the home network is over, the UE is detached from the partner’s network and it attaches back to home network.

Another scenario of MT session is depicted in the diagram below. In this case IMS is the front door system and it sends a trigger to MME which initiates offloading.

Once the UE attaches to the partner network, the visited network sends “Location Updated” to home MME in addition to location update. Once this message is received, MME requests IMS to redirect the call to the new location (partner’s network).

In most unlikely scenario if all the operators send NACK, the home network has no option but to reject the request.

### Advantages

1. Improved accessibility
2. Operators can avoid investment in capacity enhancement where it does not make business sense.
3. No hardware investment required.

### Issues

1. Call set up time. Since the UE has to attach to another network the call set up time will be of the order of CSFB. This should be acceptable as CSFB is.
2. Because of set up delay for MT call, we need to play suitable announcement to the caller to standby.
3. Operators need to evolve commercial models for the local offload. Users can’t be charged extra for offloading, unlike national roaming.
4. Possible misuse. An operator may not enhance capacity even if it is required and thus misusing other’s capacity. This can be controlled by commercial agreement.
5. User may be annoyed by seeing roaming indication. The indication needs to be suppressed when it is local roaming mode. User may not even be aware that he/she is using other operator’s network.
Conclusion

Given the fact that operators are under continuous pressure to provide better services at reduced cost, any investment in capacity enhancement needs to be carefully evaluated. By using local offloading they can avoid unnecessary investment. By entering into multi-operator agreement for local offloading they can utilize each other's idle capacity in a win-win manner while maintaining customer experience high from accessibility perspective.

Disclaimer

Intention of this paper is to describe a concept and the procedure/sequences shown are purely to explain the concept. It is not backed by any PoC or research. This document shouldn't be used for implementation directly. Once the operators agree that it is helpful and they want to have this functionality, detailed standards will evolve naturally.