Satellite Communications: Internet challenges and strategies for low-latency applications

Igor Bilogrevic
Royal Institute of Technology
Stockholm, Sweden

igorb@kth.se

March 1, 2008

Abstract

This paper analyzes several mitigation strategies and techniques that could enable delay sensitive applications to work properly over satellite IP networks. A study is conducted on different satellite classes (GEO, MEO and LEO) and the analysis of the current solutions for real-time interactive applications is performed. The great distance from Earth results in an important RTT, which is the main aspect to consider for real-time interactive applications. To date, the multi-layered architecture seems to be the best way to go, since it overcomes the drawbacks of the single layer approaches. This, in turn, creates traffic balancing and quality of service concerns that are also discussed according to the proposed solutions.

1 Introduction

During the last two decades satellites have been providing a very heterogeneous set of services that go from live television broadcasting, GPS navigation to disaster situation recovery.

In the 80's, little concern has been addressed to the possibility that a single user could access the internet through a satellite link and therefore most of the efforts in building satellite networks were aimed at diffusing the same information to as many users as possible, like the television broadcasting.

Then suddenly things changed. Internet is nowadays widespread and is used both for pleasure and business meaning that the satellite networks that have been built twenty years ago are no longer suitable for our current needs.

The main problem with satellites is their great distance from Earth, which results in long delays, increased noise and limited bandwidth [1]. Since there is not much that can be done in order to decrease the noise, the alternative is to redesign the satellite network architecture by, for instance, adjusting their position, number and altitude and by increasing their technological level. All of the aforementioned aspects have a direct impact on the satellite network performances which, in return, have influence on the user's satisfaction when using determined services.

In this paper we discuss the challenges that are involved with internet services and the strategies to enable a certain class of application to work properly. The latter are the delay-sensitive applications like VoIP phone calls compared to the delay-insensitive ones, like ftp file transfers, web browsing and e-mails.

The paper is organized as follows. In Section 2 we discuss the issues related to the integration of the satellite networks with the terrestrials while Section 3 covers more specifically the GEO satellites and their performance. Section 4 treats LEO satellites and the possible improvements regarding the TCP protocol in presence of large round-trip-times and error rates. In Section 5 we focus on a rather new approach that could significantly enhance performances by using different classes of satellites at the same time, namely the "multi-layered" network architectures [2]. Finally, Section 6 gives an insight on next-generation satellite networks architectures and we conclude in Section 7.

2 Satellite/Terrestrial Networks

The key for a successful inter-operation between the existing terrestrial communications infrastructure and the satellite networks is their dynamic and seamless inter-connectivity [3]. Each user, whether connected through a mobile device or a fixed landline, should be able to connect with other users at any given time and request the services that he needs without worrying about the medium that is being used. There are, however, distinct aspects that characterise terrestrial and satellite network performances.

According to Daoud [3], there are three categories for mobile multimedia services represented by *Asymmetric medium-band, asymmetric large-band* and *symmetric large-band* services. In this paper the first two types are referred as the "delay-insensitive" applications whereas the last one is "delay-sensitive". In the first group of applications, the delay-insensitive, we can account for all those services that do not require a real-time interaction between the user that is requesting the content and the content provider like, for instance, file downloading, email sending, ebanking, online shopping and multimedia streaming. The delay-sensitive applications require, on the other hand, a quick request/response between the involved parts and could be, for example, VoIP phone calls, video-conferencing and online gaming.

The advantage of the fixed or mobile terrestrial infrastructure, compared to satellite networks, is that it provides a relatively good quality of service because of its low delays, high bandwidth and adapted protocols [4]. The biggest shortcoming, however, is that it was not designed to provide total and global coverage. Indeed, satellites offer global coverage at the same cost for both high and low population density regions. That is not the case for landlines since the high initial investment has to be split among the customers in different ways (equally, per-region) [5], depending on social and economical policies.

In order to inter-connect and take advantage of both networks, there are some issues to consider mainly in the space segment of the network. The first choice is the orbit of the satellites since it will determine the coverage, the availability, the path delay and loss characteristics. There are three main possibilities: *Low-Earth (LEO)* at 500-1000 km from the surface of the Earth, *Medium-Earth (MEO)* at 3000-35'000 km and *Geo-stationary* orbits (GEO) at 35'786 km. Table 1 shows some aspects of each category as well as the required quantity of satellites for a global Earth coverage and call handover frequency. As it can be seen, the best choice for delay-sensitive applications are LEO and MEO satellites thanks to the low RTT rates.

Another aspect that has a huge impact on the overall performances of IP packet-switched network is the transport layer protocol *TCP* that is used to deliver connection-oriented information through many heterogeneous networks. According to [4], its design assumptions were based on *wired connections*, meaning that packet losses were supposed to be caused only by congestion, the network topology was fixed, bandwidth was constant and, most importantly, the propagation delay of a communication path was relatively stable [6]. Obviously none of the above assumptions holds for satellite links and therefore new adjustments are needed in order to achieve good performances in mixed terrestrial and satellite networks.

In the next two sections we will cover different strategies that could help achieve those performances in separate GEO and LEO network architectures.

3 GEO Strategies

Figure 1 describes a common scenario where the user is connected to the server through a GEO satellite link. In theory, if the user wanted to download a file of 100 MBytes and assuming the same Bit-Error-Rate (BER) for satellite and terrestrial links, everything would be fine except the fact that the file download would be completed 0.6 seconds later than the same file download if both user and server were connected through a wired link. As we said, this would be the case in theory.

In practice, however, FTP works on top of TCP which was designed for wired links that have a stable bandwidth, a short RTT (100 msec) and low error rates. In addition, TCP's biggest window size is only 64 KBytes [7] and the packet losses are assumed to be caused by network congestion, excluding the possibility that the bit-level changes are due to the characteristics of the channel [4].

Combining those results with the RTT (500 msec) for a GEO satellite link we have:

$$\begin{aligned}\text{Throughput} &= \frac{\text{Window size}}{RTT}\\ &= \frac{64\text{ KB}}{0.5\text{ sec}} = 128\text{ KB/s}\end{aligned}$$

which corresponds to a 1 Mbps ADSL or Cable connection. But those are not the only problems.

Actually, TCP's congestion and flow control mechanisms play an important role. In fact, packet losses and timeouts tell TCP that the network is congested and in the worst case they force it to cut its threshold value in half and reset the congestion window completely. This decrease in throughput leads to low bandwidth use and in the end to even worse achievable performances than the theoretical 1 Mbps.

In their work, Bharadway et al. [8] propose a solution to the above mentioned problems by making TCP "aware" of the satellite link characteristics. They manage to do so by installing two gateways (connection splitting proxies) between user and server and by using some enhancements of the original TCP protocol. In Figure 2 the connection between the user and the server is split by the two gateways that can now be configured to use any protocol to communicate between them. In the specific case, they use TCP with a larger congestion window [9] and initial window [10] together with selective ACK's [11].

Orbit	Required Qt.	Complexity	RTT	Cost per unit	Best suited application
GEO	3	Simple, no handover	500 ms	High	Broadcasting, navigation, *asymmetric multimedia applications*
MEO	≈ 10	Complex, some handovers	200 ms	High	Real-time interactive, *symmetric multimedia applications*
LEO	$\gg 10$	Complex, many handovers (every 10 min)	50ms	Medium	Real-time interactive, *symmetric multimedia applications*

Table 1: Satellite characteristics in different orbits (Adapted from [2]).

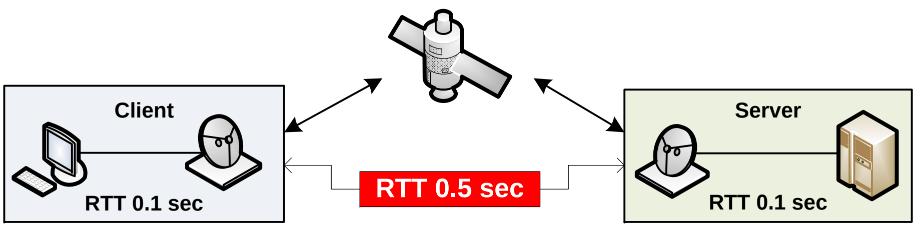

Figure 1: Internet connection over GEO satellite

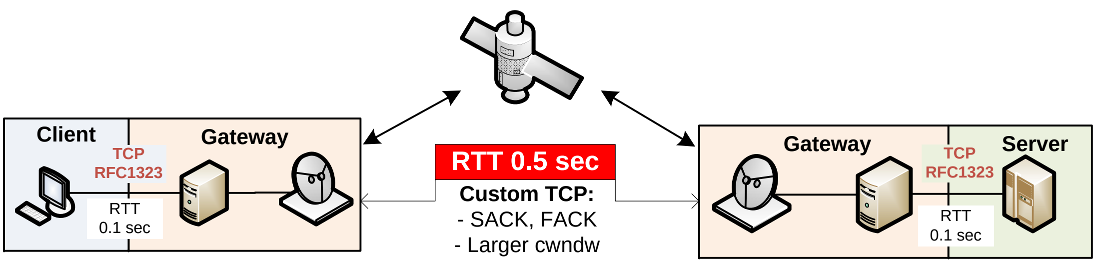

Figure 2: Internet connection through GEO satellite with two gateways (Adapted from [8]).

When there is a connection request from the user to the server, the user side gateway forwards this request and initiate a connection with the server side gateway by using the same ports and sequence numbers. Once the connection between user and server is established, the client starts sending packets to its gateway which acknowledges immediately those packets on behalf of the server eventhough the actual server has not yet received any of them. The gateway then forwards those packets to the other remote getaway which passes them to the final server. If there is an error between the two gateways, it is signaled to the host by telling it that the receive buffer is now full. The host would then adopt flow control policies and reduce the rate at which it is sending data, without the need for the very drastic congestion control policies to take actions.

By adopting those different techniques they have proven that in a noisy environment where the $BER = 10^{-6}$ and $RTT = 0.5$ sec, the average utilisation of a 1.5 Mbps link has passed from 5% to 55% [8], which corresponds to an increase of the average throughput from 75kbps to 820kbps. This holds, however, only for file sizes greater than one order of magnitude of the product $RTT \cdot bandwidth$ because of the TCP's "slow start" mode predominance for small files.

Nevertheless, this solution is not optimal. One of the assumptions for it to perform well is that the route from user to server does not change during the session. If that happens, the two gateways would not be able to synchronise their sequence numbers with the transiting packets anymore. Another problem is the buffer size at the gateways. There is no way of knowing in advance how many non acknowledged packets will be held in the gateway's buffer in case of a failure or a prolonged congestion at the other host.

It is therefore clear that efficient Internet over GEO satellites is a subject that needs to be developed and studied further.

4 LEO Strategies

In the previous section we have discussed about how to improve Internet performances on a GEO satellite system. By splitting the original client-server connection with two gateways that are aware of the different characteristics of the channel and by enhancing the communication protocol, the authors were able to show remarkable results for delay insensitive applications.

In the present section we study the LEO satellite performances for Internet applications by going deeply into the delay sensitive applications. The key for enabling that kind of uses is the relatively small distance from the satellite to the Earth, which is directly proportional to the RTT. Table 1 shows

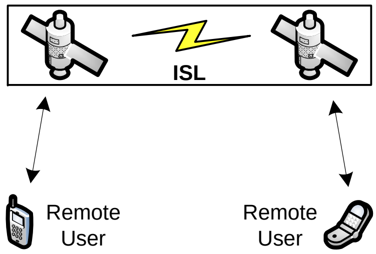

Figure 3: LEO topology with inter-satellite link between two remote users.

that LEO satellites have indeed RTT times that are comparable with the ones that we experience on the terrestrial network, i.e. smaller than 100 msec.

However, there are also some drawbacks. The small distance from the Earth makes the coverage area of the satellite, i.e. the footprint, much smaller than the GEO satellites and that translates into many more LEO satellites needed for a global coverage. There is one more aspect to consider: since their orbit is lower, they have to travel at higher speeds that the rotation of the Earth and therefore connection handovers are inevitable.

The *inter-satellite link*, or *ISL*, is a characteristic that enables satellites to transfer the current connections (to hand it over) directly to another satellite without having to relay to the ground station [12]. This represents an evolution of the processing done on board by the satellite that has not been possible before. Thus, the level of complexity of LEO and MEO units supporting ISL is greater since they have to provide packet switching and routing capabilities instead of only relaying the received signal to the ground gateways like some sort of "bent pipes".

4.1 Inter-satellite Links

In the study done by P. Loreti et al. [13], ISLs have been compared to the traditional satellite-ground-satellite scenarios for voice-over-IP performances in the case the two users who were not under the coverage area of the same satellite. They analyzed the delays in the polar orbit and inclined orbit LEOs assuming ISL links for the former and no ISL for the latter. The actual complete simulation setup and parameters can be found in [13].

The authors have achieved different results depending on specific scenarios. In the case where the two remote hosts were under the footprint of the same satellite, the inclined orbit provided better results for delay insensitive applications thanks to the possibility of combining different copies of the same signal through independent channels, i.e. the diversity technique. Since two or more satellites were covering one specific area in the inclined orbit case, the probability that the received SNR was below a certain treshold was lower than the polar orbit's case. On the other hand, delay sensitive applications are better supported by polar orbit constellations using inter satellite links because they suppress the need for a signal to be bounced back to Earth before reaching the final host and therefore help achieving lower RTT values.

4.2 TTL for TCP packet reordering

From the previous study it is straightforward that the inter-satellite links are very important for delay sensitive applications. Loreti et al. [13] have showed that VoIP traffic can be routed through LEO satellites without big time gaps that would annoy two users during a telephone conversation. Since VoIP streams are transferred on top of UDP [14], packet losses do not cause the sender to reduce its transfer rate as it would be for TCP. Eventhough less important, packet loss in real-time application could be disturbing in a case of an MPEG2 video stream, where even a single packet loss could cause significant quality degradation over several fractions of a second [15].

In a bursty environment like the satellite communication channel, packet losses are even more frequent. For instance, assume that two remote hosts want to establish a communication link through LEO satellites as in Figure 4. Packets 1 to 5 are ordered TCP segments that are forwarded by Sat1 and Sat2 to the receiver. At the same time, both satellites are moving and their coverage area is constantly changing. When the receiver switches to Sat1 because it has a stronger SNR than Sat2, packet 1 is already "on the fly" through Sat2. Since the distance between the receiver and Sat1 is shorter than the distance between the receiver and Sat2, packets 2 to 5 arrive at destination before packet 1 does. The congestion control mechanism of TCP reacts by sending a triple ACK back to the sender, which in turn recognizes that as a packet loss and sends packet 1 again after halving its congestion window, before the timout (fast retransmit mechanism). This in turn leads to poor bandwidth utilization since packet 1, in reality, has not been lost.

Tsunoda et al. [6] analyzed the fast retransmission mechanism of TCP and came up with an interesting algorithm that is fully compatible with the existing TCP and IP protocols and prevents TCP from sending duplicate ACKs if the packet arrived out of order has a lower TTL than the last in-order packet.

More precisely, they base the decision of sending the duplicate ACKs on an internal timer, in addition to the timout timer already present in TCP. The new treshold timer, T_{rsv}, is defined as

$$T_{rsv} = (TTL_{out-order} - TTL_{in-order}) \cdot \alpha$$

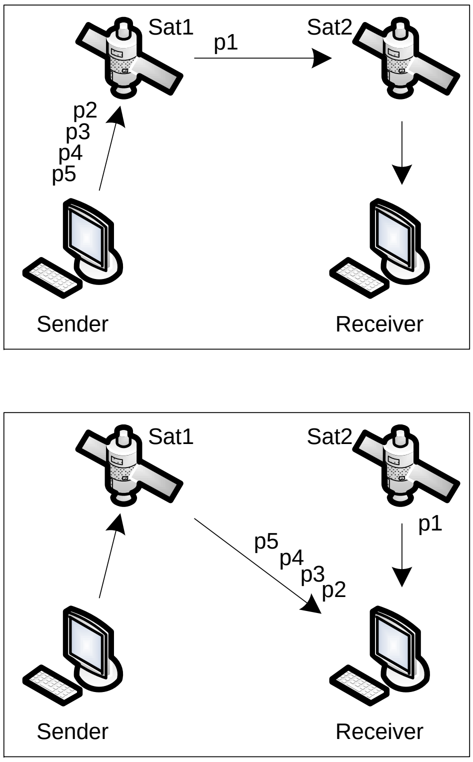

Figure 4: Communication link using two-hop satellite link (top) and packet disorder at the receiver when passing to single-hop link (bottom) (Adapted from [6]).

where α is based on the per satellite hop delay. The timer is initialized if the received packet comes with a sequence number that is greater than the expected one. Once that happens, the receiver waits for the in order sequence number to arrive until T_{rsv} expires. If the in order packet arrives in time, the receiver sends normal ACKs to the sender as if nothing happened whereas if the in order packet does not arrive during T_{rsv} it sends the duplicate ACKs [6].

The results obtained by testing the new approach are promising: fast retransmission has been avoided and therefore no bandwidth was wasted due to TCP congestion control in multi hop LEO satellite environment. It is also remarkable that the algorithm is compatible with IPv4 and IPv6 since it does not introduce any change in the IP and Transport layer protocol headers.

5 Multi-Layered approach for Real-Time applications

In Section 4 we have discussed the problem of real-time applications over LEO satellite networks and how we could achieve better performances for both delay and bandwidth utilization. The proposed solution has the prerequisite, for each of the involved satellites, to be able to process incoming data packets directly and take forwarding decisions. On board processing capabilities are thus essential and they define the difference between a "bent-pipe" satellite, like some GEOs that redirect all the traffic to one specific Earth gateway, and a packet switching and routing satellite like MEO and LEO nodes.

Inter-satellite links are at the core of packet switching but nevertheless they are not optimized to handle multiple connections with different types of layers. They are usually "single-layered" architectures, i.e. each satellite family (LEO or MEO) communicates and forwards traffic mainly to the same layer satellite group [13], as opposed to inter-orbit links (IOL) that interconnect satellites flying at distinct layer orbits.

Apart from the increased complexity of satellites supporting ISL and IOL, crucial is the traffic load experienced in different regions of the globe. We could assume that most of the developed countries generate more traffic than less developed countries and saturate their links more often. Uneven traffic load generates more packet drops and increases end-to-end delays.

5.1 Traffic Balancing and Packet Priority

The article written by Bahyan, Gür and Alagöz [2] explores a set of new ways of dealing with traffic balancing among LEO and MEO satellites by proposing a dual-layered model for delay-sensitive applications. They argue that by splitting the routing table calculations between MEO satellites and sending them back to the LEOs could greatly improve the traffic balancing, avoiding excessive queuing and disturbing delays for VoIP conversations. The algorithm designed by the authors, "Adaptive Routing Protocol for Quality of Service" (ARPQ), takes advantage of real-time network traffic information gathered by the LEO satellites and a dynamic routing table computed by the MEOs.

We could start by looking at Figure 5 where we have:

- One Source satellite that sends data.
- One Destination satellite that covers the final user, the target.
- Two Group Managers (GM_i and GM_{i+1}) that are covering a group of LEO satellites.
- Two Leo Groups (LG_i and LG_{i+1}) that are covered by the respective GMs, serving the users on Earth.

A LEO Group (LG) is composed of a set of LEOs that are under the footprint of the same MEO satellite. Every group has only one GM and each single LEO node belonging to a group knows about its GM. The role of the GM_i is to maintain link with all the LG_i, to collect information about their links status and to transmit relevant information to the Plane Manager MEO satellite (PM, not in the figure), who is in charge of the routing table computations.

Each LEO satellite holds a Neighbour status list (NSL_i) that stores one of the possible states in which every neighbouring satellite could be, i.e. free, fairly busy and busy depending on the load level, and a Link State (LS) that has the delay values of each of its output links. Once a LEO peer gets the information about the link status, it informs the GM by sending it the LS. Next, the GM informs all of his neighbouring MEOs, including the PM, about the link states. Then, the PM computes the routing table and distributes it to the GMs who, in return, distribute it to the respective LEOs.

Now that the basic routing table establishment has been clarified, we will focus on the actual algorithm that classifies the incoming packet as long-distance voice, short-distance voice or background packets. In case of a voice packet, the long or short distance characteristic is determined by a threshold time, based on the estimated delay time between the source and destination satellites. If the actual time to get to the destination satellite is higher than the threshold time, the packet is marked as a long-distance voice packet, otherwise it would be a short-distance voice packet.

What happens next depends on the previous categorisation. If the packet has been defined as long-distance, it goes directly from the source LEO to its GM, who will forward it to the neighbouring GMs via ISLs towards the final GM, who will send it to the destination LEO and finally to the end user. This reduces significantly the number of hops (and thus delay) from source to destination.

On the other hand, if the packet is a short-distance voice, it will be forwarded to the next calculated LEO satellite, no matter how busy it could be.

Different is the approach in case of a background packet: a non-voice packet will be forwarded to the preferred next hop LEO if and only if its state is "free", otherwise it will be forwarded to the LEO that has the lowest queue length and is closer to the destination than the current one.

The simulations [2] showed significant improvement in delay-sensitive applications by using the ARPQ algorithm. The most important issue, nonetheless, is the determination of the threshold time, D_{trs}, since packet categories are based upon it. By choosing a value of $D_{trs} = 500$ msec or more, all voice packets would be marked as short-distance and that would discard the possibility of using MEO satellites as alternatives to the possibly congested LEO networks. On the contrary, a too small value

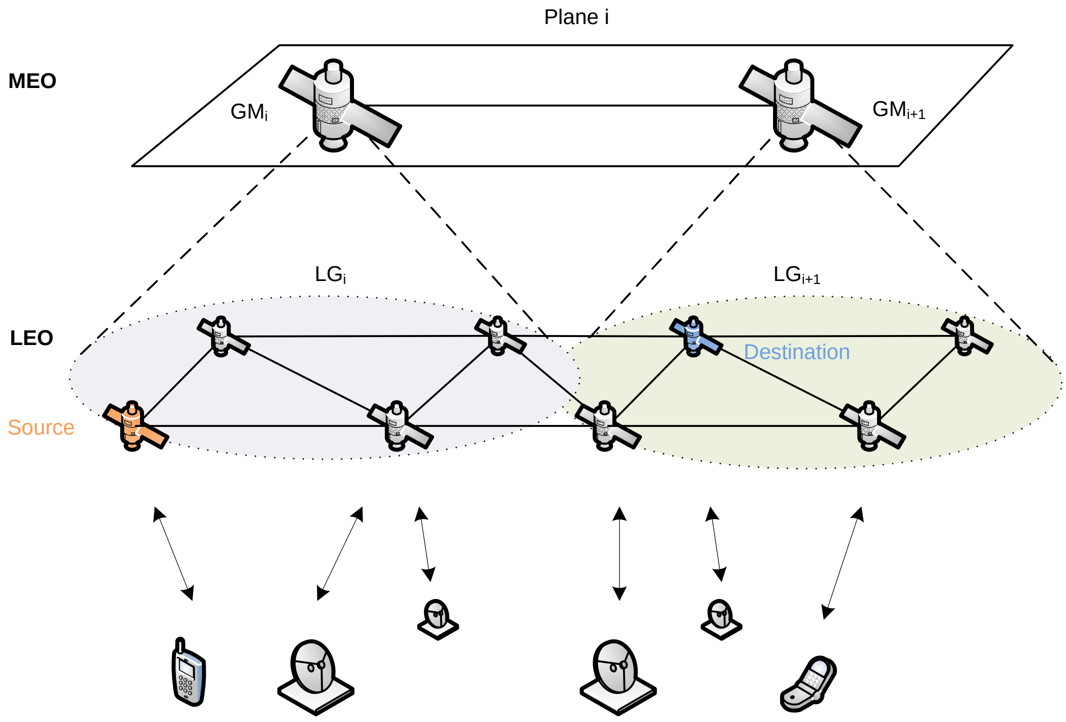

Figure 5: Multi-layered satellite network architecture for ARPQ protocol (Adapted from [2]).

of the threshold, like $D_{trs} = 80$ msec for instance, would have the opposite effect, by setting all voice packets to long-distance voice and using more inter-orbital link bandwidth. That, in return, would increase the delay in case the final user were just few LEO hops away from the source satellite.

5.2 Reserved Buffer and Dedicated Bandwidth

Eventhough Bahyan, Gür and Alagöz invented an algorithm that uses a dual-layered (LEO-MEO) satellite network architecture, there have been previous researches in the field. By prioritizing delay sensitive traffic already at the the source, great improvements could be achieved.

In the work of Dash, Durresi and Jain [16] it is argued that multi-layered satellite networks could be created as well by means of GEO, LEO and High Altitude Platforms (HAP), skipping de facto the intermediate MEOs. The routing strategy developed in their paper is based on the bandwidth availability, hop count and two path classes: one for real-time packets and one for all the remaining data. Also, it takes a rather different approach for the routing table calculations and synchronization when compared to ARPQ.

The information gathered by the LEOs concerning bandwidth usage and buffer is uploaded to the GEO satellites (which don't have much computing power) who send them back to the Earth gateway. Once on Earth, the data is analyzed and the routing table computed. When finished, the table is transmitted back to the GEO, forwarded to its peers and then back to LEOs under its footprint.

Moreover, the authors suggest a separate buffer and dedicated bandwidth for delay sensitive traffic. By separating the high priority packets from the rest, the drop ratio is clearly lower since the voice packets have their own buffer and by having a dedicated bandwidth, the QoS can be guaranteed. Indeed, new priority connections to a LEO satellite are accepted only if there is still some dedicated bandwidth available. In addition, the load balance among the uncongestioned LEO links assures that the delay insensitive applications do not cause excessive packet losses and delay for real-time requirements. HAP are then only used for delay sensitive traffic over a specific disaster or battlefield area, where ad-hoc communication infrastructure and high bandwidth demand have to be assured without affecting too much the higher level satellite network.

The drawback of such strategy is that the best-effort traffic, as opposed to real-time, is dropped whenever the priority link usage grows and there is no unused bandwidth left. By combining TCP and the packet drops, the proposed solution is suboptimal with respect to the bandwidth utilization for best-effort traffic if there are suddenly many voice calls over a specific region.

Nevertheless, it represents a remarkable improvement for load balancing and quality-of-service assurance and, together with the ARPQ algorithm, represents a possible way of offering good user experience with low delays and relatively low cost for a global voice and data service coverage.

6 Next-Generation Satellite Networks

The previous section covered a very interesting and promising field of studies. The use of different layers of satellites at the same time could overcome many of the disadvantages of using only one layer. It is therefore reasonable to assume that multi-layered satellite network architectures will become a subject of important debate and innovation in the near future.

6.1 Probabilistic Approach

What authors seem to agree on in general is the need to offer a differentiated quality of service for distinct service classes. Best-effort and real-time interactive applications have individual requirements in order to achieve a desired user satisfaction. If, for instance, the RTT value were 0.5 sec, the average user would not perceive the same level of disturb to download a web page as if he would for a VoIP call or during a video-conference. Additionally, it is also less annoying just not to initiate a phone call than to suddenly end an ongoing conversation.

The chance that a new conversation fails to initialize or a sudden drop of an existing conversation is defined as the blocking probability of a call and in this sense, Uzunalioglu et al. [17] analyzed the problem of route establishement in case of link connectivity changes.

The circular orbits travelled by the satellites have a crossing point exactly at the poles. When two nodes on different orbits arrive in proximity of a pole, they turn off their ISLs because they are no more in the respective antennae viewing spots. After passing this critical region, they turn the ISLs on again but on the opposite side, as represented on Figure 6. It is straight forward to realize that at each occurrence of the pole crossing, every ongoing call has to be rerouted through another set of satellites (link handover). This, in turn, could saturate the available bandwidth of other satellites which could result in call blocking or even drop of the existing sessions. If, however, there is an intersatellite handover (where the user switches the connection to another satellite that has a better coverage), there is no need for rerouting.

In order to mitigate the undesirable call drop caused by link handovers, the authors [17] have developed and implemented a routing protocol, the "Probabilistic Routing Protocol" (PRP), that decides the route of a new incoming voice connection,

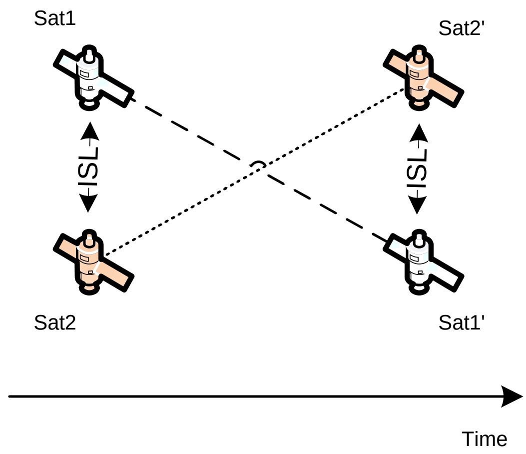

Figure 6: Two LEO satellites passing over the polar region (Adapted from [17]).

based on the probability that a link handover would occur before the conversation ends by itself.

In fact, the algorithm looks for the route from source to destination that would result in a spontaneuos ending or an intersatellite hand over with certain target probability p and discards the other routes. The call duration, the intervals between link handovers and connection handovers are expressed as random variables whose parameters are trained by statistics based on current and past calls.

If we define the T_c as the estimated call duration, T_{hr} as the time before the occurrence of an intersatellite handover and T_{lh} as the interval before a link handover, we have

$$Pr(min(T_c, T_{hr}) < T_{lh}) > p \quad (1)$$

All new routes violating (1) are excluded. Link handover probability can be well estimated thanks to the predictable topology of LEO satellites around the globe.

Note also that the authors suggest the use of PRP only for new incoming calls and not for ongoing ones. Using PRP for all calls would unnecessarily increase the blocking probability even more and since it is more annoying to experience a call drop than a call establishment denial, the ongoing calls should use all ISLs with enough capacity.

The simulations [17] showed that PRP could reduce the rerouting occurrences by 80% if the target probability p is 0.99 for calls with a 3 minutes average duration. This, however, resulted in a significant increase in the blocking probability since many routes were excluded. Thus, there is a trade-off between lower rerouting attempts and higher blocking probability. The optimal balance was achieved by setting p to 0.9, which provided 50% less reroutings and a tolerable blocking probability.

6.2 Systematic Approach

As we have seen, the discussion about the perceived quality of service for a specific type of traffic is a delicate matter that does not have a unique solution. It is a rather vast field and the mitigation strategies seem to be going on the right direction. What we have described are possible solutions that have some important drawbacks. Therefore, a proper QoS implementation for satellite networks needs to be assessed from a broader perspective.

Kota and Marchese [18] explored the problem by looking at the whole system and analyzing the service quality for all the communication layers, describing requirements, objectives and mechanisms that could enable efficient QoS over satellite IP networks.

According to the authors [18], there are different QoS parameters in satellite networks. Some of them are handover priority, blocking probability for ongoing and new connections, inactivity period during handovers, their speed (delay, jitter changes) and the packet losses that could occur during those handovers. Although link layer techniques such as modulation, coding, multiplexing, etc. are very interesting subjects, they are outside the scope of this paper and will therefore not be treated.

In existing terrestrial networks, there are already different QoS mechanisms and Table 2 shows some of the main aspects of each. As we can see, Integrated Services (IntServ) and Multiprotocol Label Switching (MPLS) are better suited for QoS guarantees but they require more resources than Differentiated Services (DiffServ), which in turn is more scalable but less efficient.

MPLS selects a path for each class of packets with the same label (delay sensitive and delay insensitive classes for example) and forwards packets to the respective routes according to the label. We could have then a route that has low delay but limited bandwidth and another route that has large bandwidth but with varying delay.

One important aspect of MPLS is that such packets can be forwarded by a router without having to inspect all the IP header. Only looking at the label of the packet suffices to take forwarding decisions. This allows for very fast processing times and therefore reduces the queue buffers. Considering the limited buffer size and processing power among satellites, Kota and Marchese have suggested MPLS for satellite IP QoS networks.

Results achieved in simulated scenarios show that when delay sensitive traffic has a distinct and independent route from the delay sensitive path, there is no delay or jitter increase for the former class by augmenting the latter traffic load on the network, since both traffic types have distinct routes and therefore they do not interfere with each other.

However, the improvements have to be done at all the different layers, from access techniques, coding, bandwidth allocation, fair utilization and especially in application level QoS.

Next-generation satellite networks for IP traffic depend largely on traffic engineering methods (like MPLS) and clever routing algorithms to cooperate with terrestrial fixed and mobile networks. Further studies have to take into consideration the complete system that interconnects networks with different physical characteristics and protocols. All this has to be done having in focus seamless interoperability and user satisfaction.

7 Conclusion

This paper analyzed several aspects regarding satellite Internet challenges and mitigation strategies. Particularly, the focus was pointed towards the applications that have stringent requirements for delay and jitter values such as real-time interactive traffic, where the communicating hosts need very fast information exchange in order to achieve a desirable level of user satisfaction.

The key issue for enabling seamless internet use among satellite and terrestrial networks is their cooperation and coordination. It is not trivial, however, to find a suitable solution for the integration because of the intrinsec difference between the terrestrial and space environment.

Internet services over geo-stationary satellite can be provided in a rather simple way since the satellite does not change during a connection time interval. However, delay sensitive application suffer greatly from the vast distance between satellite and Earth gateways, which makes the use of VoIP, teleconference and telemedicine hardly possible. The bandwidth utilisation, moreover, is not optimally utilized due to the limitations of current transport layer protocols like TCP. Those protocols were designed for use in low bit-error rate and constant propagation delay environment and as such, they are not adapted to the bursty nature of the error-prone satellite links.

Eventhough protocol enhancements and connection splitting solutions exist, they cannot solve the delay issues so crucial for real-time interactive applications. Since the speed of light is finite by nature, it is undesirable to use only GEO satellites for VoIP or interactive gaming applications.

The solutions for such services come from the lower orbit satellites, known as medium- and low-Earth orbit satellites. Indeed, LEO nodes offer a very interesting delay characteristic which makes them suitable for low-latency application use. Nevertheless, other challenges need to be addressed for those kind of orbits. First, they are rotating at different speed than the Earth around its axis. Therefore, in order to maintain orbit, they have to travel

Feature	IntServ	DiffServ	MPLS
QoS guarantee	Per-flow	Per-class	Per-label
Resource allocation	Defined by application	Defined by service level agreement	Aggregate source-destination &/or application
Challenges	Scalability limited by nr. of flows	No per-flow service, QoS not always guaranteed	Router Complexity

Table 2: QoS mechanisms for terrestrial networks (Adapted from [17]).

at greater speeds, which results in a dynamic constellation topology. This, in turn, means that their coverage area changes continuously with respect to the user terminal. The result is that connections need to be handed over to other satellites every 10-12 minutes, an interval corresponding to the average coverage area of a certain region for a single LEO satellite.

Efficient routing algorithms and traffic priority classes are used to increase the performances and the quality of service for LEO networks.

Interesting improvements come from the combined use of both LEO and MEO satellites. By having more distant MEOs in charge of the routing table computations and LEO topology analysis, significant performance gains could be achieved for both short distance and long distance connections thanks to inter-satellite links.

Moreover, special algorithms have been developed for traffic categorisation and load balancing. Since delay, jitter and hop count are the most sensible parameters for delay sensitive applications, routing algorithms have to consider their estimated values, in case of rerouting of an ongoing connection.

As it is, the future still looks appealing for satellite networks. Global coverage and bandwidth availability of satellite networks offer internet access and communication to remote areas, where terrestrial networks would hardly ever be deployed mainly because of economical reasons. It is therefore important to further study and analyse satellite network systems not only from network or transport layers, but from a complete system point of view.

References

[1] Bates, Regis J. Bud, "Satellite Communications Networking", Blacklick, OH, USA: McGraw-Hill Professional, 2002, pp. 390-410.

[2] Suzan Bayhan, Gürkan Gür, Fatih Alagöz, "Performance of delay-sensitive traffic in multi-layered satellite IP networks with on.board processing capability", International Journal of Communication Systems, Vol. 20, 2007, pp. 1367-1389.

[3] Fawzi Daoud, "Hybrid satellite/terrestrial networks integration", Computer Networks, Vol. 34, 2000, pp.781-797.

[4] Geoff Huston, "The future for TCP", The Internet Protocol Journal, Vol. 3, No. 3, Cisco Systems, 2000, http://www.cisco.com/web/about/ac123 /ac147/ac174/ac195/about_cisco _ipj_archive_article09186a00800c83f8.html, last visited on 03/02/2008.

[5] Associated Press, "Rural Advocates Call for More High-Speed Internet Access", 2007, http://www.freepress.net/news/25807, last visited on 03/02/2008.

[6] Hiroshi Tsunoda et al., "Avoiding TCP False Retransmission in LEO Satellite Network Using TTL Information", Electronics and Communications in Japan, Part 1. Vol. 88, No. 1, 2005, pp. 64-71.

[7] Information Sciences Institute, University of Southern California, "Transmission Control Protocol", RFC 0793, 1981, http://tools.ietf.org/html/rfc0793, last visited on 06/02/2008.

[8] G. Bharadwaj, John S. Barras, Norman P. Butts, "An architecture for Internet service via broadband satellite networks", International Journal of Satellite Communications, Vol. 19, 2001, pp.29-50.

[9] Jacobson V., Braden R., Borman D., "TCP Extensions for High Performance", RFC 1323, 1992, http://www.ietf.org/rfc/rfc1323.txt, last visited on 05/02/2008.

[10] Allman M., Floyd S., Partridge C., "Increasing TCP's initial window", RFC 2414, 1998, http://www.ietf.org/rfc/rfc2414.txt, last visited on 06/02/2008.

[11] Mathis M. et al., "TCP selective acknowledgement options", RFC 2018, 1996, http://www.ietf.org/rfc/rfc2018.txt, last visited on 06/02/2008.

[12] European Space Agency (ESA), "Inter-satellite Links overview",

http://telecom.esa.int/telecom/www/object /index.cfm?fobjectid=431, last visited on 10/02/2008.

[13] P. Loreti at al., “Mobile Internet access using satellite networks”, International Journal of Satellite Communications and Networking, Vol. 22, 2004, pp. 587-610.

[14] Roberto Arcomano, “Technical Info about VoIP”, VoIP How-to, 2002, http://tldp.org/HOWTO/VoIP-HOWTO-4.html, last visited on 11/02/2008.

[15] Sam Liand, David Cheriton, “TCP-RTM: Using TCP for Real Time Multimedia Applications”, Stanford University, 2002.

[16] Durga Shankar Dash, Arjan Durresi, Raj Jain, “Routing of VoIP traffic in multi-layered Satellite Networks”, Performance and Control of Next-Generation Communication Networks, Robert D. van der Mei, Frank Huebner, Editors, Proceedings of SPIE Vol. 5244, 2003.

[17] Hüseyin Uzunalioglu, Ian F. Akyildiz, Michael Bender, “A routing algorithm for connection-oriented Low Earth Orbit (LEO) satellite networks with dynamic connectivity”, Wireless Networks, Vol. 6, J.C.Baltzer AG, Science Publishers, 2008, pp. 181-190.

[18] Sastri Kota, Mario Marchese, “‘Quality of service for satellite IP networks: a survey”, International Journal of Satellite Communications and Networking, Vol. 21, 2003, pp. 303-349.