Radio Resource Management Across Multiple Protocol Layers in Satellite Networks: a Tutorial Overview (*)

Paolo Barsocchi, Nedo Celandroni, Erina Ferro, Alberto Gotta - ISTI C.N.R., Italy
Franco Davoli, Giovanni Giambene – CNIT, Italy
Francisco Javier González Castaño – University of Vigo, Spain
Jose Ignacio Moreno, University Carlos III, Madrid, Spain
Petia Todorova - Fraunhofer Institut FOKUS, Berlin, Germany

Abstract

Satellite transmissions have an important role in telephone communications, television broadcasting, computer communications, maritime navigation, and military command and control. Moreover, in many situations they may be the only possible communication set-up. Recent trends in telecommunications indicate that four major growth market/service areas are: messaging and navigation services (wireless and satellite), mobility services (wireless and satellite), video delivery services (cable and satellite), and interactive multimedia services (fibre/cable, satellite).

The major drawback when using geostationary satellites (GEO) is the long delay, which can have a great impact on the end-to-end delay user requirements. Moreover, atmospheric conditions may seriously affect the data transmitted via satellite. Since the satellite bandwidth is a relatively scarce resource compared to the terrestrial (e.g., the optical transport networks), and the environment is harsher, resource management of the radio segment assumes an important role in the system's efficiency and economy. This tutorial aims to give the basic elements of telecommunications via GEO satellite, with emphasis on the issue of radio resource management. The low earth orbit (LEO) satellite constellations are briefly discussed. Currently, after the IRIDIUM service disaster, LEO constellations are no longer a hot research field; however, in the future they may again play an important role. Aspects considered will include basic transmission and multiple access techniques, channel modelling and fade countermeasures, and their performance analysis done by means of theoretical, experimental, and simulation tools.

1. Introduction

It is not easy to treat the subject of resource allocation in satellite communications, since it encompasses practically all layers (application, transport, network, data link or MAC - Medium Access Control, - and physical) of a protocol' architecture. In this paper we present an overview of the way each of the afore mentioned levels acts, regarding to the resource allocation problem. The quality of service (QoS) requirements of the applications (application layer) is the basis upon which all the actions undertaken by the

(*)Work funded by the European Commission in the framework of the "SatNEx" NoE project (contract No. 507052).

lower layers depend. Immediately under the application layer, the transport layer must manage allocation problems to satisfy the upper levels' requirements. Nowadays, the TCP/IP stack is so widespread that TCP is the most popular transport protocol; whatever TCP version one uses, it may work inefficiently on a satellite link, due to the effect that the long satellite round-trip time (RTT) has on the congestion window. Our work does not focus on TCP over satellites, but we certainly cannot avoid mentioning it, since it has a great influence on the MAC layer, which runs the algorithms for the bandwidth resource allocation. We will not explicitly deal with the QoS solutions at the network layer, as they are not specific for the satellite environment; nevertheless, there is an interaction in mapping, for instance IP QoS classes onto DVB (Digital Video Broadcasting) classes. The MAC layer must manage the type of access to the satellite link, and the algorithms for utilizing the common resource (i.e., the channel capacity). Most of those algorithms implement some optimisation of the bandwidth allocation, which influences the TCP goodput. Call Admission Control (*CAC*) is also managed at the MAC level, being itself a way to block requests that cannot be satisfied, according to the QoS specified by the application.

At the physical link level, most of the games have already been done, in the sense that, optimised or not, a stream of bits must be transmitted over a satellite link, which is often affected by signal fade due to bad atmospheric conditions. Nevertheless, the original QoS of the data, which at this level is expressed in terms of Bit Error Rate (BER), must be maintained as much as possible. The physical link must deals with several problems, such as the signal fade, which must be counteracted by means of techniques that are transparent to the upper layers, but which interfere with the MAC layer, in the sense that there must be a strong correlation between the two layers in order to select the best countermeasure. In LEO satellite constellations the handoff problem has to be faced as well, as shown in Section 6.

Section 2 of this paper briefly presents the satellite networks and the broadband satellite scenario requirements. Section 3 deals with QoS requirements of applications that may use a satellite network. Sections 4, 5 (which are the core Sections), and 6 cover aspects pertaining to the transport, data link and physical layer, respectively. The survey in this paper is far from exhaustive: rather, the goal is to touch upon some of main aspects related to resource allocation in this specific environment for the purpose of QoS management.

For the reader's convenience, references are cited in the form [*x.y*], where *x* is the number of the Section referred to by *y*.

2. Digital satellite networks

The advantages of combining the high bandwidth, wide area coverage, reconfigurability, and multicast capabilities of satellites with terrestrial networks offer vast new market opportunities. In those areas where terrestrial high-bandwidth communications infrastructure is impractical or non-existent, satellite communications may be the only solution. It is expected that the satellite component will play an important role in the universal delivery of third-generation wireless multimedia services, due to the large coverage area offered. The satellite component can be used to complete the coverage of the terrestrial network in areas

where deployment of the latter would be uneconomical or technically infeasible. The UMTS (Universal Mobile Telecommunications System) integrates cellular, cordless, and paging technologies. Integration of the Satellite-UMTS (S-UMTS) component with the third-generation (3G) terrestrial mobile networks is regarded as a key factor for the success of the system, as it poses a valuable complement to the terrestrial UMTS network. S-UMTS is currently in an advanced standardization phase. It is expected to have the same set of features as terrestrial UMTS (T-UMTS) [2.1], in particular concerning the channel allocation process. However, the actual allocation techniques will have to be adapted to the final specifications about the frame length, the allocation frequency and, more generally, to the chosen transmission scheme and its constraints. Moreover, the actual channel allocation policies will have to cope with satellite delays, frequent handoffs, and channel impairments, in order to maintain the required QoS. Some satellite characteristics can be utilized for new services or result in satellite-specific QoS classes (e.g., wide area coverage). The major drawbacks of satellite communications are the high propagation delay, due to their altitude, and the SNR (signal to noise ratio), which can dramatically decrease with adverse atmospheric conditions, particularly in the Ka band, i.e., at frequencies above 14 GHz.

Taking into account that different constellations and orbits can be envisioned for the exploitation of S-UMTS services, several considerations regarding the choice of satellite constellation can be made in terms of the characteristics and geographical locations of the targeted users, as well as the desired services. In particular, geostationary satellites can be proven to be the right choice for providing complementary services over regions already served by T-UMTS networks. On the other hand, constellations of satellites in Low or Medium Earth Orbit (LEO or MEO) might prove to be better suited to collect the traffic of users evenly distributed over the entire globe, such as sea vessels and airplanes [2.2].

Geostationary satellites orbit the earth at an altitude of about 36,000 km. This embeds a typical figure of about a quarter-second round-trip (uplink plus downlink) propagation delay in a communication system. Many potential customers of a future global communications system would not tolerate these delays, for instance in voice applications. The performance of protocols with acknowledgements and a time-out-based congestion control mechanism, e.g., TCP, is inherently related to the delay-bandwidth product of the connection. If this product is high, as it is in GEO satellite links, a non-negligible packet loss due to data corruption may cause significant performance degradation. In fact, TCP is unable to distinguish between congestion and corruption losses; thus, any lost packet causes a reduction in TCP packet sending rate, even if the link is not congested [2.3]. In order to reduce the BER (and then the packet loss due to corruption), the employment of FEC (Forward Error Correction) coding is mandatory in satellite transmissions; this allows trading bandwidth for data reliability, as well as a certain gain in power.

LEO satellites travel at altitudes ranging from 700 to 2,000 km above the earth. They have the potential to be successful in future global telecommunications systems, due to their greatly reduced propagation delays as well as their ability to communicate with less powerful earth terminals. LEO systems also yield smaller satellite cells, allowing greater frequency reuse and hence higher capacity.

The major disadvantage of LEO satellites, compared to GEO satellites, is that they move rapidly around the earth. A LEO satellite at a height of 1,000 km introduces a round-trip propagation delay of 7-20 ms and travels over the earth at a speed of 7 km/s. Satellite movements can increase the typical number of handoffs during a call, thus increasing the amount of control and/or reservation traffic.

Table I summarizes the main advantages and disadvantages in using LEO and GEO satellites.

ADVANTAGES	DISADVANTAGES
LEO	
1. Portables can use low power and non-directive antennas 2. Direct portable-to-portable connections 3. Achieve large frequency reuse at L-band due to geographical separation 4. Short propagation delay 5. World-wide coverage 6. Allows high elevation-angle operations from portables 7. More fail-safe due to the number of satellites 8. Provides position location	1. Large number of satellites needed 2. Relatively complex payloads: - on board switching, routing - multi-hop operation for long distance connections - handoff (Sect. 6), cross-links - power management along with the orbit 3. Complex network operations: - satellite monitoring, control, replenishment 4. Individual satellites spend a large % of time over areas with little traffic
GEO	
1. Wide area coverage from one satellite 2. Coverage can be directed at traffic concentrations 3. Cost-effective for limited area coverage 4. Frequency reuse at Ka band due to antenna pattern separation 5. Relatively few satellites needed for world-wide coverage	1. Path delay 2. Antenna directivity needed at portable 3. Outages due to rain fades 4. Portable operation unproven at Ka band 5. Requires clear line-of-sigh path to satellite 6. Extended network needed for global coverage

Table I. Main advantages and disadvantages of LEO and GEO satellites.

2.1 Broadband satellite scenario requirements

Next-generation multimedia broadband satellite networks require the development of key technologies to increase their capacity and efficiency, as well as to decrease the total cost for the end user. These requirements call for very high throughput, flexibility, multi-beam processing, efficient modulation and coding techniques, and system adaptability.

Current bent pipe satellite technologies in Ku-band (12-14 GHz) do create difficulties in developing profitable multimedia satellite systems. In progress development of Ka-band, spot beams and frequency re-use will be effective for a near-term business model. However, the reduced coverage of each spot beam can be highly advantageous. In fact, it offers a higher G/T (receiving antenna gain over system noise temperature), thus allowing higher return channel burst rates or lower power level requirements, which decrease the price of the terminals significantly. Thus, the employment of Ka-band capacity will greatly affect multimedia satellite business and will probably lead to more successful network architectures and profitability.

Typical Ku-band broadcasting links are designed with a clear-sky margin of 4-6 dB and a service availability target of about 99% of the worst month (or 99.6% of the average year). Since the rain attenuation curves are

very steep in the region 99-99.9% of the time, many decibels of the transmitted satellite power are required for increasing system availability, in a given receiving location, by a few dozen minutes per year. While this waste of satellite power/capacity cannot be easily avoided for broadcasting services, where millions of users spread over vast geographical areas receive the same contents simultaneously, for unicast communications the spatial and temporal variability of end-user channel conditions allows increased average system throughput, by adapting the power required by each point-to-point link. This is achieved by Adapting Coding rate and Modulation format (ACM) to best match the user SNIR (Signal-to-Noise plus Interference Ratio), thus making the received information data rate location- and time-dependent. The inclusion of advanced coding and modulation schemes has been the first objective of the DVB-S2 (Digital Video Broadcasting via Satellite) working group [2.4]. In particular, ACM has been considered a powerful tool for further increasing system capacity, allowing for better utilization of transponder resources and hence providing additional gain with respect to current DVB-S systems [2.5]. Therefore, in DVB-S2, ACM is included as normative for the interactive application area and optional for DSNG (Digital Satellite News Gathering) and professional services. The standardization for the use of ACM by the DVB-S2 standard introduces an adaptive physical layer, which calls for the development of optimum adaptive resource management strategies for fully exploiting ACM potentialities.

The need to increase bi-directional data rates, so that multimedia broadband satellite solutions can be closer to the specs of terrestrial networks, is undoubtedly a core need for any DVB-based or DOCSIS (Data Over Cable Service Interface Specification)-based network, due to the increase in video and large file transfers in enterprises. Future broadband satellite networks should aim to create more symmetry between forward and return links, due to a predicted future demand for symmetric applications, such as videoconferencing or interactive e-learning. Moreover, satellite solutions that integrate into and coexist with existing enterprise infrastructure must include features and functions similar to a terrestrial solution: VPN (Virtual Private Network, i.e., a network that is constructed by using public networks to connect nodes) support, firewalls, multicast services, Service Level Agreements (SLAs), guaranteed QoS, and e-mail.

3. APPLICATION LAYER: QoS requirements for multimedia traffic

Supporting quality of service (QoS) in telecommunication systems nowadays is of great importance, as demands the determination of the requirements to be met when a service is provided. This task should take into consideration the fact that the user is not interested in how a particular service is provided, but in the degree of service quality he or she ultimately enjoys. There are different QoS requirements to be guaranteed according to the chosen multimedia traffic; they will be briefly analyzed in sub-sections 3.1-3.4.

3.1 Conversational services

The most common use of this scheme is real-time conversation, such as telephony speech. However, Voice over IP (VoIP) and video conferencing are two new applications that will require this scheme, since the Internet and multimedia services are developing rapidly. This scheme raises the most stringent QoS

requirement: the transfer time must be short while, at the same time, variation between information entities of the stream must be preserved in the same way as for real-time streams. The limit for acceptable transfer delay is very strict, since failure to provide low enough transfer delay results in unacceptable lack of quality. When referring to real-time conversation, the fundamental characteristics for QoS are: to preserve the time relation (variation) between information entities of the stream, and the conversational pattern (stringent and low delay). A real-time streaming application is one that delivers time-based information in real-time, where time-based information is user data that has an intrinsic time component. Video, audio and animation are examples of time-based information, as they consist of a continuous sequence of data blocks that are presented to the user in the right sequence at pre-determined instants. Examples of applications that use this stream are conversational voice, videophone, interactive games, two-way control telemetry and telnet. Table II summarizes these applications, providing explicit requirements for each.

Conversational voice

Audio transfer delay requirements depend on the level of interactivity of the end users. To preclude difficulties related to the dynamics of voice communications, ITU-T Recommendation G.114 [3.1] recommends the following general limits for one-way transmission time (echo control already included in):

0 to 150 ms	preferred range [< 30 ms, user does not notice any delay at all; < 100 ms, user does not notice delay if echo cancellation is provided and there are no distortions on the link]
150 to 400 ms	acceptable range (but with increasing degradation)
above 400 ms	unacceptable range

The human ear is highly intolerant of short-term delay variation (jitter), so the latter should be kept below a very low limit; 1 ms has been suggested. However, the human ear is tolerant of a certain amount of distortion of the speech signal. An acceptable performance is typically obtained with *Frame Erasure Rates* (FER) up to 3%. Finally, a connection for a conversation normally requires the allocation of symmetrical communication resources. As stated in Section 1, the round-trip delay between an earth station and a GEO satellite is around 250 ms; thus, the use of GEO satellites imposes severe constraints in order to achieve good QoS for conversational voice. Instead, LEO or MEO satellites can be used without any problem.

Videophone

Videophone implies a full-duplex system, which carries both video and audio and is intended for use in a conversational environment. Therefore, the delay requirements are the same as for conversational voice, i.e., no echo and minimal effect on conversational dynamics, with the added requirement that audio and video must be synchronized within certain limits to provide “lip-synch” (i.e., synchronization of the speaker’s lips with the words being heard by the end user). In fact, it is difficult to meet these requirements, due to the long delays incurred in video codecs. The human eye is tolerant of some information loss, so that some degree of packet loss is acceptable. It is expected that the video codecs will provide acceptable video quality with frame erasure rates up to about 1%. GEO satellites introduce the same constraints for QoS in videophone as

in conversational voice services. If we take into account the lip-synch problem, then GEO satellites are not suitable for providing this service.

Interactive games

Requirements for interactive games greatly depend on the specific game, but it is clear that demanding applications require very short delays, and a value of 250 ms is proposed, based on the studies done in [3.2], consistent with demanding interactive applications. The convenience or not to use GEO satellites (due to their round-trip propagation time -RTT - between the earth stations close to 250 ms) depends on the actual interactive game application.

Two-way control telemetry

Two-way control telemetry is included here as an example of a data service that does require a real-time streaming performance. Two-way control implies very tight limits on allowable delay and a value of 250 ms is proposed [3.3], but a key differentiator from the voice and video services in this category is the zero tolerance for information loss. Again, the convenience of using GEO satellites depends on the actual application. Here, besides the RTT delay introduced by GEO satellites, the "noisy" links may introduce information loss, thus making the use of geostationary satellites completely inadequate for this kind of applications. Error control techniques (see Section 6) may be adopted to solve this problem.

Telnet

Telnet (or, more likely, Secure Shell, *ssh*, http://www.ietf.org/ids.by.wg/secsh.html) is included here with a requirement for a short delay in order to provide essentially instantaneous character echo-back. With respect to GEO satellites, the same comments apply as for the two-way control telemetry case. Here, if the RTT introduced by GEO satellite makes the user perception of the telnet application bad, terminal auto-echo can be used.

Medium	**Application**	**Degree of symmetry**	**Data rate**	**Key performance parameters and target values**		
				End-to-end One-way Delay	*Delay Variation within a call*	*Information loss*
Audio	Conversational voice	Two-way	4-25 kbps	< 150 ms preferred < 400 ms limit	< 1 ms	< 3% FER
Video	Videophone	Two-way	32-384 kb/s	< 150 ms preferred < 400 ms limit Lip-synch < 100 ms		< 1% FER
Data	Telemetry two-way control	Two-way	<28.8 kbps	< 250 ms	Not Applicable (NA)	Zero
Data	Interactive games	Two-way	< 1 kB	< 250 ms	NA	Zero
Data	Telnet	Two-way (asymm.)	< 1 kB	< 250 ms	NA	Zero

Table II. Conversational services. End-user performance expectations.

3.2 Interactive services

The second service class is interactive service. This applies when a human or a machine is on-line requesting data from a remote server, and is characterized by the request/response pattern of the end user. An entity at the destination is usually waiting for a response message within a certain period of time. End-to-end round

trip delay time is therefore one of the key attributes. Another characteristic is that the content of the packets must be transparently transferred (with low bit error rate). The resulting overall requirement for this communication scheme is to support interactive non-real time services with low end-to-end round-trip delay.

The fundamental QoS requirements for of interactive traffic are the request-response pattern and the preservation of the payload content. Examples of this type of service are voice messaging and dictation, data, Web browsing, high-priority transaction services (E-commerce), and e-mail (server access). The corresponding requirements are summarized in Table III.

Voice messaging and dictation

Requirements for information loss are essentially the same as for conversational voice, but a key difference here is that there is more tolerance for delay, since there is no direct conversation involved. Therefore, the main issue becomes determining how much delay can be tolerated between the user's command to replay a voice message and the actual start of the audio. There is no precise data on this, but a few seconds' delay is considered reasonable for this application.

Web-browsing

In this category, we consider the retrieval and viewing of the HTML component of a Web page. The main performance factor is how fast a page appears after it has been requested. A value of 2-4 s per page is proposed [3.3]; however, improvement on these figures to a target figure of 0.5 s would be desirable.

Medium	**Application**	**Degree of symmetry**	**Data rate**	**Key performance parameters and target values**		
				One-way Delay	Delay Variation	Information loss
Audio	Voice messaging	Primarily one-way	4 -13 Kbps	< 1 s (playback) < 2 s (record)	< 1 ms	< 3% FER
Data	Web-browsing - HTML	Primarily one-way		< 4 s /page	NA	Zero
Data	Transaction services – high priority, e.g., e-commerce, ATM	Two-way		< 4 s	NA	Zero
Data	E-mail (server access)	Primarily One-way		< 4 s	NA	Zero

Table III. Interactive services. End-user performance expectation.

3.3 Streaming services

Streaming services are mainly unidirectional with high continuous utilization (few idle/silent periods) and short time variations between information entities within a flow. However, there is no strict limit for delay and delay variation, since the stream is normally aligned at the destination. In addition, there is no strict upper limit for packet loss rate. The fundamental QoS requirements of real-time streams are the unidirectional continuous streaming, and the preservation of the time relation (variation) between information entities of the stream. The resulting overall requirement for this communication scheme is to support streaming real-time services that have unidirectional data flows with continuous utilization. Application examples and corresponding limitations are presented in Table IV.

Audio streaming

Audio streaming is expected to provide better quality than conventional telephony, so information loss requirements (packet loss) is correspondingly tighter. However, no conversational element is involved and delay requirements can be relaxed.

One-way video

The main distinguishing feature of one-way video is that no conversational element is involved; therefore, delay requirements are not so stringent.

Still image

The human eye is somehow tolerant of information loss, as regards still images. However, single bit errors can cause large disturbances in some still image formats, so generally zero errors are expected when still image is transmitted, but delay requirements are not stringent.

Medium	**Application**	**Degree of symmetry**	**Data rate**	**Key performance parameters and target values**		
				Start-up Delay	Transport delay Variation	Packet loss at session layer
Audio	Speech, mixed speech and music, medium and high quality music	Primarily one-way	5 -128 kbps	< 10 s	< 2s	< 1% Packet loss ratio
Video	Movie clips, surveillance, real-time video	Primarily one-way	20 -384 kbps	< 10 s	<2 s	< 2% Packet loss ratio
Data	Bulk data transfer/retrieval, layout and synchronization information	Primarily one-way	< 384 kbps	< 10 s	NA	Zero
Data	Still image	Primarily one-way		< 10 s	NA	Zero

Table IV. Streaming services. End-user performance expectations.

3.4 Background services – applications

This scheme applies when the end-user, typically a computer, sends and receives data-files in the background. It is a classical data communication scheme characterized by the fact that the destination does not expect the data within a certain time. Also, the packets' content must be transparently transferred; however, the bit error rate must be kept at very low levels.

The fundamental QoS elements for background traffic are: the destination is not expecting the data within a certain time; and to preserve the payload content. The resulting overall requirement for this communication scheme is to support non-real time services without any special requirement on delay. A background application is one that does not carry delay-sensitive information. In principle, the only requirement for applications in this category is that information should be delivered to the user essentially error-free. However, there is still a delay constraint, since data is effectively useless if it is received too late for any practical purpose. Examples are: fax, low priority transaction services, e-mail (server to server), SMS, download of databases and measurement records. Here GEO satellites can be employed without the tight constraint on delay of conversational services. For applications that do not tolerate packet loss, forward error control techniques can be applied (see Section 6). Table V summarizes the delay requirements of the various categories; note that the delay values in the table represent one-way delay, i.e., from originating entity to terminating entity.

Service class	Conversational (delay << 1 s)	Interactive (delay ~ 1 s)	Streaming (delay < 10 s)	Background (delay > 10 s)
Error Tolerant	*Conversational voice and video*	*Voice messaging*	*Streaming audio and video*	*Fax*
Error Intolerant	*Telnet interactive games*	*e-commerce Web browsing*	*FTP, still image paging*	*e-mail arrival notification*

Table V. Application examples in terms of QoS.

Fax

Fax is not normally intended to accompany real-time communication. Nevertheless, it is expected that a fax be received within about 30 s.

Low priority transaction services

An example in this category is Short Message Service (SMS); 30 seconds is proposed as an acceptable delivery delay value.

3.5 Traffic patterns

As previously explained, although the important thing for the user is the QoS and not the actual mechanisms employed to provide this quality, operators must know the traffic requirements of the applications in order to plan their infrastructure precisely. Previous sub-sections are important for an operator to establish the delays and the frame-loss rate of the links. They also describe the bandwidth needed by the application, but the operator must have a bandwidth characterization closer to the physical level and must know the burstiness of the traffic. This section aims to describe this burstiness. Our results are based on the measures done in Madrid's Moby Dick trial [3.4] site; they are sure to reflect the applications that will be used in future 4G networks. Of these, Moby Dick is one of the first approaches [3.5]. In order to build the traffic pattern of some of the most promising applications, two kinds of measures were performed, which are presented in Tables VI and VII, respectively.

Application	Direction: Client (C)→Server (S) Peer A→ Peer B	Mean	Max	Min	Average deviation (Desv)	Desv/Mean
Audio stream[1]		380	380	380	0	0
Quake[2]	C→S	82.12	110	73	3.73	0.05
	S→C	161.79	1051	90	48.12	0.3
VoIP[3]	A→B	93	93	93	0	0
	B→A	93	93	93	0	0
Video Stream[4]		1363.93	1364	424	6.09	0
Tetris[5]	C→S	68.63	330	46	36.12	0.53
	S→C	87.72	329	60	60.45	0.68

Table VI. Packet size at IP level in bytes for the different applications.

Application	Client (C)→Server (S) Peer A→ Peer B	Mean	Max	Min	Average deviation (Desv)	Desv/Mean
Audio stream[1]		49.33	52	28	2.86	0.06
Quake[2]	C→S	124.35	144	94	6.53	0.05
	S→C	59.95	72	36	4.32	0.07
VoIP[3]	A→B	24.81	52	0	23.52	0.95
	B→A	32.42	52	0	20.67	0.64
Video Stream[4]		172.9	196	18	18.47	0.11
Tetris[5]	C→S	7.64	98	0	11.97	1.57
	S→C	7.63	88	0	11.82	1.55

Table VII. Packets per second generated by the different applications.

Notes: 1) By using Robust Audio Tool (RAT) with Linear 16 codec

2) 6 players match all against all

3) By using RAT with GSM codec

4) By using VideoLAN and a movie trailer of about 2 minutes

5) 6 players match

4. TRANSPORT LAYER: TCP/IP via satellite

The end-to-end performance of data transfer over paths with significant delay mainly suffers from interactions with TCP's window-based flow control. In fact, the high RTT of GEO satellite links greatly limits the TCP congestion window growth in time, while each packet loss, due to data corruption (which may be non-negligible), is interpreted as a congestion loss by TCP. The latter then reduces throughput by reducing the congestion window.

Much effort has been made to improve the TCP mechanism efficiency over satellite links. This subsection provides a survey of different solutions proposed to enhance the TCP end-to-end performance.

4.1 TCP standard mechanisms

Basically, TCP provides two functions: *flow control* and *congestion control* [4.1]. The flow control scheme allows an adequate exchange of data between two TCP nodes. The main element of this functionality is the mechanism called *sliding window*. In fact, the purpose of the transmission window is to allow the receiving TCP node to control the amount of data that is being sent to it at any given time. Besides, to avoid congestion problems, the congestion control scheme is based on two algorithms, called *Slow Start* and *Congestion Avoidance*, respectively. These algorithms are based on two variables: the *congestion window* (*cwnd*) and the *slow start threshold* (*ssthresh*). In particular, the end-systems probe the network state, by gradually increasing the window of segments that are outstanding in the network, until the network becomes congested and drops segments. Initially, the increase with time is exponential during the *Slow Start* phase. The Slow Start algorithm is quite simple and based on data sent for round trip. At the start, the source sends one TCP segment and waits for an acknowledgement. When it receives the acknowledgement, it sends two segments. Then, every time the sender receives the acknowledgements for the data sent, it sends the double

amount of packets in the next trip-time. When the window size reaches the *slow start threshold*, the increase becomes linear, thus allowing for a gentler probing of the available capacity (*Congestion Avoidance* phase).
Unfortunately, high latency, bandwidth asymmetry and transmission errors affect the TCP performance on satellite channels remarkably. These factors affect the standard TCP mechanisms that may worsen the performance. In particular, the TCP slow start mechanism may be too slow for broadband connections traversing long RTT links, resulting in low utilization. On the other hand, satellite channels increase the amount of time the congestion avoidance algorithm takes to increase the "*congestion window*", when compared to terrestrial links.
In order to improve TCP mechanisms' efficiency over the satellite links, many solutions can be adopted, some of them specifically proposed for satellites, and others for more general environments. A classification of such solutions is cited in the following. In recent years, enhancements of TCP standard mechanisms have significantly improved the performance of bulk transfers, for example by enlarging the size of the initial window [4.2], by reducing the number of acknowledgments, and by using the byte counting [4.3]. Several other solutions are based on implementations of modified flow control mechanisms and options. For example, TCP Westwood [4.4] dynamically estimates the available bandwidth to set the "congestion window" and the "slow start threshold" after a packet loss.
To improve the TCP performance on links characterized by high latency and high bandwidth-delay product, beyond protocol enhancements, several architectural enhancements or middleware techniques have been proposed. This approach has been generalized to the so-called "Performance Enhancing Proxy" or PEP [4.5]. The modification of the architecture (e.g., splitting the path and terminating connections at each step, acknowledging packet reception) can be based on the use of TCP on the non-satellite segment, and on using optimised protocols (e.g., XTP, SCPS-TP [4.6], etc.) on the satellite segment.

4.2 Cross-layer Approach

The transmission of TCP traffic over wireless data links poses a problem that is not usually apparent on wired links. As previously stated, TCP behaviour is very sensitive to packet loss, which is interpreted as a congestion signal, and consequently as a reason to throttle the data rate. The common behaviour of the wireless data link is to discard data packets in error, which are not even made available to the TCP/IP stack above. This means that the TCP/IP stack cannot distinguish packet loss due to data corruption, from packet loss due to other reasons that are interpreted as congestion signals, thus causing the reduction of the data rate.
In order to improve the efficiency of TCP over wireless links, where improving the error rate is generally expensive (if at all possible), a number of techniques are commonly used. These include various types of spoofers [4.7], which may or may not preserve the TCP end-to-end semantics, and Automatic Repeat reQuest (ARQ). These techniques, each operating at different levels of the protocol stack, exploit local knowledge of the wireless hop characteristics in order to add a shorter delay control loop underneath the end-to-end control loop in the TCP connection. This gives a prompter reaction to packet loss and consequently improves the end-to-end performance at the expense of local buffering of packets to retransmit in case of

loss and increased complexity. Methods have been postulated for a number of different techniques operating at different levels. For example, link-level FEC operates below TCP. Some methods, such as Explicit Loss Notification (ELN) [4.8], which make the TCP stack aware of packet losses due to link errors, need some sort of coupling between TCP and the link level, and require TCP modifications at the end points. Other methods proposed involve changing the TCP stack at the end points, for example TCP-Peach [4.9].

The cross-layer approach is orthogonal to all these techniques (apart from link-level FEC, a similar but less general concept). It operates at the physical or link level, by trading the bandwidth of the wireless hop for packet loss rate. It does not interfere in any way with the normal behaviour of the TCP stack; the wireless link parameters are simply appropriately tuned at the physical or link levels. Generally speaking, for any given wireless transmission method, a number of parameters are chosen in order to obtain a target performance in terms of BER and IBR (Information Bit Rate). For a given available radio spectrum, antenna size and maximum transmission power, the selection of a modulation scheme and various FEC types usually allow a wide range of choices. With a cross-layer approach, it is possible to obtain a better performance for TCP connections, by jointly choosing the BER and IBR of the wireless links that maximize a TCP connection goodput, i.e., the end-to-end transfer rate [4.10].

In some recent work regarding cross-layer optimization [4.10] [4.11], a general framework has been outlined for tuning some given transmission equipment, in order to maximize the throughput of a single TCP connection passing through the wireless link, by trading information bit rate for bit error rate. The optimal operating point is dependent on the channel conditions, and should be pre-computed once for some given equipment. During normal operation, the wireless equipment, which must be able to measure the channel conditions in real time, chooses its transmission parameters by using a simple lookup table, or multiple lookup tables in the case of dynamic bandwidth allocation. The performance analysis of these methods, conducted on specific cases with real data, has shown that relevant gains can be obtained with respect to fade countermeasures that only attempt to constrain the BER below a given threshold, and that a good range of flexibility can be attained in favouring the goals of goodput or fairness [4.10-4.13].

4.3 Interaction between TCP and RRM in satellite networks

A relevant issue that is worth approaching in the satellite network is the interaction between TCP and MAC protocols. MAC protocols play a fundamental role in guaranteeing good performance to higher-level protocols, by managing the arbitration of uplink access. In fact, decisions made within satellite Radio Resource Management (RRM) can significantly impact the end-to-end performance of TCP flows over a satellite network. Two cases are important - first, the case where the TCP protocol operates end-to-end (as is the norm in the general Internet, or is required when the IP Security Protocol (IPSEC) [4.14] is used), and second, when the RRM imposes significant delay and the operation of TCP is often “split” by a TCP Performance Enhancing Proxy (PEP) device [4.5]. Unless specifically engineered, satellite networks that employ RRM mechanisms can introduce significant and variable delay and capacity. Although PEPs are normally employed in GEO satellite networks with Bandwidth-on-Demand, there are cases where this is

either undesirable (because it breaks end-to-end semantics or the application requires end-to-end acknowledgements) or impossible (because IPSEC is used, or the PEP does not support the required protocol(s)). Where PEPs are not used, TCP currently expects the link to offer low levels of packet loss with controlled variation of the path delay; this has implications for RRM design. This requires specific research to survey the overall system performance and interoperability as a part of the wider Internet.

To mitigate the effect that RRM can have on TCP it would be desirable for RRM to be driven by TCP mechanisms. To analyze the effects of the access policy to the uplink channel on the end-to-end delay, a GEO satellite interactive network (DVB-RCS standard like) with fixed Return Channel Satellite Terminals (RCST) can be considered. Then, bi-directional Interaction Channels are established between the gateway and users for interaction purposes. A network control centre (NCC), installed on the ground, provides Control and Monitoring Functions (CMF). The satellite access scheme considered is a Multi-Frequency Time Division Multiple Access (MF-TDMA). With this discipline, a group of RCST communicates with a gateway by using a set of carrier frequencies, each of which is divided into time-slots. Therefore, the NCC will allocate a series of bursts to each active RCST as a result of terminal requests. Unfortunately, in this scenario, the allocation of new resources is very slow; in fact, it needs about 500 ms (time between the request transmission and the NCC response's arrival instant). At the same time, the TCP window may double if the "Slow Start" phase is running. This basically causes two effects: i) every RTT, the terminal will have to send new requests of resource allocations to the NCC; ii) the number of packets stored in the MAC queue will increase exponentially when the Slow Start phase occurs.

Then, there is an inefficient management of the channel resources that can drastically increase the average time that a packet spends in the queue. A possible solution, reducing the time that a packet must wait in the queue before being considered for transmission, may be based on a cross-layer interaction between MAC and transport layer. The idea is to use TCP parameters, such as the *cwnd* and *ssthresh*, to provide in advance an estimate of the resources needed to the transmission. In this way, a remarkable reduction of queuing delay and, consequently, an optimal utilization of channel bandwidth can be hoped for.

5. MAC LAYER: access schemes, allocation policies, CAC

5.1 Access Schemes

Multiple access is the "*ability of a large number of earth stations to simultaneously interconnect their respective voice, data, teletype, facsimile, and television links through a satellite*" [5.1]. It is a generic term to denote schemes to share the available capacity of a satellite transponder among several earth stations.

The following sharing techniques can be considered:

- by sharing the transponder's bandwidth in separate frequency slots (FDMA),
- by sharing the transponder's availability in discrete time slots (TDMA),
- by allowing coded signals to overlap in time and frequency (CDMA). Each earth station then separates the signals by recognizing which of the codes is destined for itself,
- by adopting a mix of the above techniques (e.g., combining TDMA and CDMA or FDMA and TDMA).

Another form of multiple access is allowed in the presence of a multi-spot beam antenna on the satellite. In this case multiple users can simultaneously access the satellite, if they are covered by distinct spot-beams. This technique is called *Spatial Division Multiple Access* (SDMA). In particular, a satellite has several directional antennas; some of these antennas may use the same frequency, provided that the cross-interference (due to antenna radiation pattern side-lobes) is negligible. Usually, beams separated by more than two or three half-power beam-widths can use the same frequencies; this frequency reuse technique permits increasing the utilization of the air interface resources.

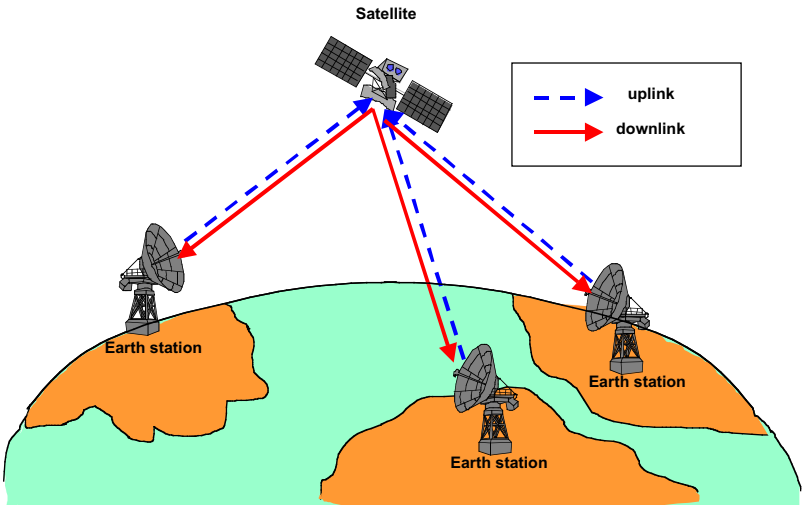

Fig. 1. *Uplink multiple access and downlink multiplexed transmissions.*

The management of TDMA, FDMA and CDMA resources can be made in a centralized (i.e., by the satellite with on-board processing or by an Earth master station) or decentralized way. The above multiple access schemes can be used to regulate both the access of distributed terminals to the satellite (*uplink*) and also to characterize the satellite transmissions to these terminals (*downlink*). The typical architecture of the sharing of satellite transmission resources among different earth stations is provided in Fig. 1.

FDMA

In FDMA, the total bandwidth is divided into equal-sized portions. An earth station is permanently assigned a portion (i.e., bandwidth) around a carrier or carriers. FDMA requires guard bands to keep the signals well separated. The traffic capacity of an earth station is limited by its allocated bandwidth and the Carrier power-to-Noise power ratio (C/N). The carrier frequencies and the bandwidths assigned to all the earth stations constitute the satellite's frequency plan. FDMA requires simultaneous transmission of a multiplicity of carriers through a common Traveling-Wave-Tube Amplifier (TWTA) on the satellite. The TWTA is highly non-linear (it produces maximum output power at the saturation point, where the TWTA is operating in the non-linear region of its characteristics) and the Inter-Modulation (IM) products produced by the presence of

multiple carriers generate interference, which degrades individual channel performance. The only way to reduce IM distortion in a given TWTA is to lower the input signal level, so that the tube can operate in a more linear region. For a given carrier, the dB difference between the single-carrier input power level at saturation and the input power level for that particular carrier in multi-carrier FDMA operations is called *input backoff*. The corresponding transmission power reduction in dB is called *output backoff.*

TDMA

In TDMA, the total bandwidth is divided into sub-channels by time. The channel is usually divided into time *slots*, organized according to a periodic structure, called *frame*. Each slot is used to convey one packet. Hence, TDMA is well suited for packet traffic. In TDMA uplink transmissions, earth stations take turns sending bursts through a common satellite transponder. As for TDMA downlink transmissions from a satellite, only one carrier is used. Hence, TDMA provides a significant advantage, since it permits a transponder's TWTA to operate at or near saturation, thus maximizing downlink C/N. In addition, there are no inter-modulation products and the resulting capacity reduction due to TWTA non-linearity is significantly reduced. However, non-linearity is not totally eliminated, since it is present in the form of non-linear inter-symbol interference, which must be minimized by transmission path filter function design. TDMA is easy to reconfigure for changing traffic demands, it is robust to noise and interference and allows mixing multimedia traffic flows. Digital data streams from many sources are transmitted sequentially in assigned time slots.

While in TDM (*Time Division Multiplexing*) all data come from the same transmitter and the clock and time frequencies do not change, in TDMA each frame contains a number of independent transmissions. Each station has to know when to transmit and it must be able to recover the carrier and the data clock for each received burst in time to sort out all desired base-band channels. This is not an easy task to perform at low C/N ratios. A long preamble is generally needed, which decreases system efficiency.

A group of earth stations, each at a different distance from the satellite, must transmit individual bursts of data in such a way that bursts arrive at the satellite in correspondence with the beginning of the assigned slots. Stations must adjust their transmissions to compensate for variations in satellite movements, and they must be able to enter and leave the network without disrupting its operation. These goals are accomplished by exploiting the TDMA organization in frames, which contain reference bursts that permit establishing absolute time for the network. Each frame contains a reference burst and a series of traffic bursts.

An earth station may transmit into or receive data from several transponders (transponder hopping). Reference bursts are generated from a control station on the ground (master station) in a centralised control satellite network. A reference burst contains at least the following items:

- carrier and Bit Timing Recovery Sequence (CBTRS) for synchronizing the transmitting and receiving carriers and modems;
- a Unique Word (UW);
- the station identification;
- the network housekeeping information.

Each burst starts with a preamble, which provides synchronization and signalling information and identifies the transmitting station. Reference burst and preambles constitute the frame overhead. The smaller the overhead, the more efficient the TDMA system, but the greater the difficulty in acquiring and maintaining synchronism. CBTRS consists of a short transmission of un-modulated carrier, followed by carrier phase transitions between 0 and π rad at the symbol clock frequency. The UW serves to mark the beginning of valid data. At the receiving end of a link, incoming bits are clocked into a shift register where they are compared with a stored version of the expected UW (UW correlator). When the bits in the register match the stored version, the correlator produces its maximum output voltage pulse. UW sequences and thresholds are chosen to minimize the false alarm and to maximize the UW detection. Depending on the framing error probability that can be tolerated, a TDMA system accepts as correct a received UW that differs from the expected one by not more than a specified number of bits. This means that when the output pulse from the correlator exceeds some minimum value, a UW is assumed to have been received and a new frame to be in progress. Thus, the UW correlator provides an accurate time reference for a receiving terminal.

Time access to the satellite link can be managed either in centralized or in distributed mode. Centralized control is generally more robust, because one station alone (the master) is responsible for the Burst Time Plan (BTP) consistency. In the case of distributed control, all stations have to listen to each other before deciding the access time; thus, the chance that some control information is lost or misinterpreted and the BTP corrupted is increased. On the other hand, the distributed control is more responsive to traffic variations, since it allows an update in one RTT. Hence, highly bursty traffic sources are more efficiently dealt with, thus increasing the global resource utilization. The lack of robustness may be compensated with complex recovery algorithms (some additional channel overhead is necessary).

CDMA

The signals are encoded, so that information from an individual transmitter can be detected and recovered only by a properly synchronized receiving station that knows the code being used (“signature code”). In a decentralized satellite network, only the pairs of stations that are communicating need to coordinate their transmissions (i.e, they need to exchange the transmission code). No frequency (as in FDMA) or time slot (as in TDMA) coordination with any central authority is required. On the average, a number of users occupy the whole transponder bandwidth for all the time. The concept at the basis of CDMA is spreading the transmitted signal over a much wider band (“Spread Spectrum”, SS). This technique was developed as a jamming countermeasure for military applications in the 1950s. Accordingly, the signal is spread over a band PG times greater than the original one, by means of a suitable ‘modulation’ based on a Pseudo Noise (PN) code. PG is the so-called Processing Gain. The higher the PG, the higher the spreading bandwidth and the greater the system capacity is. Suitable codes must be used to distinguish the different simultaneous transmissions in the same band. The receiver must use a code sequence synchronous with that of the received signal, in order

to correctly de-spread the desired signal. There are two different techniques for obtaining spread spectrum transmissions:

1. Direct Sequence (DS), where the user binary signal is multiplied by the PN code with bits (called *chips*) whose length is basically PG times smaller that that of the original bits. This spreading scheme is well-suited for Binary Phase Shift Keying (BPSK) and Quadrature Phase Shift Keying (QPSK) modulations.
2. Frequency Hopping (FH), where the PN code is used to change the frequency of the transmitted symbols. We have a fast hopping if frequency is changed at each new symbol, whereas a slow hopping pattern is obtained if frequency varies after a given number of symbols. Frequency Shift Keying (FSK) modulation is well-suited for the FH scheme.

Hybrid Multiple Access Techniques

The drawback of TDMA is the need to size earth stations for the entire system capacity (transponder bandwidth), even though the single terminal uses a small portion of that. An interesting solution is given by the hybrid combination of MF (Multi-Frequency) with TDMA systems, which takes some advantages of both FDMA and TDMA. In MF-TDMA the transponder spectrum is divided into several carriers, thus allowing the sizing of the station on a narrower bandwidth. Each carrier, on its turn, is 'divided' in TDMA mode. The transmission of the traffic stations occurs in time slots that may belong to different carriers. When slots do not overlap in time each other, the use of a unique modulator and an output power level relative to a single carrier are allowed.

MF-TDMA is a hybrid access technique also adopted in the uplink of the DVB-RCS standard [2.5]. The MF-TDMA technique allows for efficient traffic streaming, while maintaining flexibility in capacity allocation. Access to the satellite uplink employing this technique is characterized by a large number of connections that share limited system resources.

Hybrid combinations of FDMA/CDMA and TDMA/CDMA are also possible.

5.2 Packet Access Schemes and Scheduling Techniques

The basic problem is how to permit a variable group of earth stations to share satellite resources (i.e., FDMA, TDMA or CDMA resources) in a way that optimizes satellite capacity, QoS for multimedia traffic, spectrum utilization, cost, satellite power, user acceptability, interconnectivity, and flexibility. This is the task of Medium Access Control (MAC) protocols that are typically envisaged referring to uplink transmissions for what concerns the coordination of dispersed earth stations in accessing and sharing satellite resources. Several MAC schemes have been proposed for satellite systems [5.4], [5.5]. A taxonomy of MAC protocols can be envisaged as described below:

- *Fixed access protocols* that grant permission to transmit to only one terminal at a time, thus avoiding collisions of messages on the shared medium. Access rights are statically defined for the stations.

- *Contention-based protocols* that give transmission rights to several terminals at the same time. These policies may cause two or more terminals to transmit simultaneously and their messages to collide on the shared medium.
- *Demand-assignment protocols* that grant access to the network on the basis of requests made by the stations.

The reason for the presence of many different MAC protocols is that they are suitable for some applications (and related traffic types), but they do not often meet the QoS requirements for other applications. For instance, fixed access schemes are not efficient for bursty traffic, because they cannot adapt to varying traffic conditions. Below a short survey is provided for the main MAC schemes of the above three different classes.

Fixed access protocols (circuit level)

In fixed access protocols, transmission resources (i.e., frequency bands with FDMA, time slots with TDMA and codes with CDMA) are rigidly assigned to terminals. For instance, with TDMA, a given slot (or a group of slots) is periodically assigned on a frame-basis to the stations. Moreover, with FDMA a frequency band can be assigned in three different ways:

- Single Carrier Per Link (SCPL): each station is assigned a carrier for each link. In case of full two-way connectivity between N stations, the number of carriers is Nc = N (N-1).
- Single Carrier Per Station (SCPS): each station is assigned a single carrier for all its links. In this case, the number of carriers Nc is reduced to N.
- Single Carrier Per Channel (SCPC): each station is assigned a single carrier per channel. This scheme allows a flexible usage of frequency resources.

Contention-based protocols for packet data traffic

This class of MAC schemes encompasses pure Aloha and Slotted-Aloha, which are classical protocols for satellite systems [5.6]. In the Aloha scheme, earth stations are not coordinated in their transmission attempts. As soon as new data is available, packets are transmitted. Slotted-Aloha introduces synchronization, so that packets can only be transmitted at the beginning of time slots. These random access protocols are very simple to implement, but do not guarantee an adequate utilization of satellite resources (classical values are 18% with Aloha and 36% with Slotted-Aloha). This is the reason why new packet access schemes have been proposed to efficiently support multimedia traffic with differentiated QoS requirements. In particular, recent studies have been carried out [5.7], combining random access and spread spectrum (CDMA-like) schemes. Moreover, we consider below the adoption of the *Packet Reservation Multiple Access* (PRMA) scheme for TDMA-based mobile satellite systems.

PRMA

PRMA is a MAC protocol that combines Slotted-Aloha random access with slot reservation on a TDMA air interface [5.8]. The PRMA protocol was originally proposed for terrestrial micro-cellular systems [5.9], but it has been proved to be effective for multimedia transmissions in case of Low Earth Orbit (LEO) or Medium Earth Orbit (MEO) satellite communication systems [5.10]. PRMA can be viewed as a Dynamic TDMA (D-TDMA) protocol where time slots are allocated to the users for uplink transmissions after a contention procedure. PRMA is targeted mostly for mobile networks comprising a satellite and a number of mobile stations that exchange voice and data traffic. Such protocol is decentralized, with a form of control operated by the satellite that is supposed to be regenerating with on-board processing (OBP).

The efficiency of PRMA relies on managing voice sources with Speech Activity Detection (SAD): only during a talkspurt, a voice source needs to have reserved one slot per TDMA frame for transmitting its packets. Note that voice activity factors are typically lower than 50%, so that the rigid assignment of a TDMA slot per frame to a voice source would be rather inefficient. With PRMA, a TDMA frame consists of reserved and empty (i.e., available) slots. This information is sent to all stations via satellite broadcasts. As soon as a new talkspurt is revealed, the relevant station tries to transmit a packet in the first idle slot (contending state), depending on a permission probability scheme. When a station's transmission attempt is successful on a slot, the station obtains the reservation of this slot in subsequent frames. The reservation is released when the talkspurt ends. In conclusion, the assignment of time slots to users is not fixed, but is dynamically handled so as to multiplex many traffic flows on the available time-slot resources. From these considerations, it is evident that the satellite must be regenerative and with on-board processing (architecture with centralized control).

For an efficient use of the PRMA protocol, the TDMA frame length must be larger than the round trip propagation delay. Moreover, the packet lifetime (in the case of real-time traffic) must be greater than the round-trip propagation delay. These conditions are particularly important in satellite communication systems, since they pose a constraint for both the altitude of the satellite orbital configuration and the minimum elevation angle. Typically PRMA is well-suited for LEO systems and, with less efficiency, for MEO satellite systems as well. However, PRMA cannot be employed to support real-time traffic for communications with geostationary satellites.

Each data packet carries a header field that indicates the transmitting station. This information is important, because the satellite can immediately recognize the station to which the packet belongs. If just one attempt occurs on a slot (and if the related packet header is correctly received), the satellite is able to reserve that slot to the mobile terminal and broadcasts this information for the next frame. From these considerations, it is evident that the satellite must be regenerative and of the on-board processing type. If multiple transmissions occur on the same slot, the simultaneously-received packets collide; if we exclude the capture effect, no header can be correctly detected and the mobile stations recognize after a round-trip delay that their attempts have been unsuccessful, so that they need to attempt again. Hence, the presence of collisions increases the access delay, a critical aspect for managing real-time traffic with stringent time constraints. If the packet transmission deadline is exceeded, the packet is discarded. Since this phenomenon may occur at the

beginning of an access phase, it is called *front-end clipping*. A user may reserve more than one time slot, if needed and if available.

The PRMA access protocol must be suitably designed in order to avoid unacceptable values of the mean access delay and, hence, of the packet dropping probability. In particular, permission probabilities for real-time traffic sources need to be sufficiently high to guarantee prompt access. However, too high values cannot be used, otherwise repeated collisions may lead the access protocol to instability [5.11]. This is a crucial aspect of the correct design of PRMA-like protocols. Fig. 2 shows an instance of channel allocation for two simultaneous requests.

The limit of the PRMA scheme used in satellite systems is that, in case of a collision, the mobile station recognizes after a round-trip propagation delay that it has to attempt again. Even in LEO and MEO systems these delay values can be critical in comparison to packet transmission deadlines that may range from 30 ms up to 90 ms for real-time traffic. A modified PRMA protocol has also been proposed in [5.10], [5.12], where a station is also allowed to attempt transmissions (according to a permission probability scheme) while it is waiting for the outcome of a previous attempt. If the previous attempt has been unsuccessful, this modification permits a faster access. Otherwise, these further attempts are useless and may hinder the access to other terminals. Accordingly, this scheme has been called *PRMA with Hindering States* (PRMA-HS). It has been shown that the advantages of this fast retransmission scheme overcome the problems due to useless attempts [5.10], [5.12]. Moreover, PRMA-HS is well-suited to support multimedia traffic classes, such as voice, Web browsing and background data, by adopting differentiated permission probability values. Finally, in [5.13] the use of adaptive permission probability values has allowed increased utilization of air interface resources, while assuring the stability of the access phase.

Extensions of the classical PRMA access scheme have been proposed in [5.14] for a CDMA/TDMA air interface. In this case, terminals make transmission attempts by selecting a code in addition to a time slot in order to send their access bursts. The permission probability scheme is still adopted.

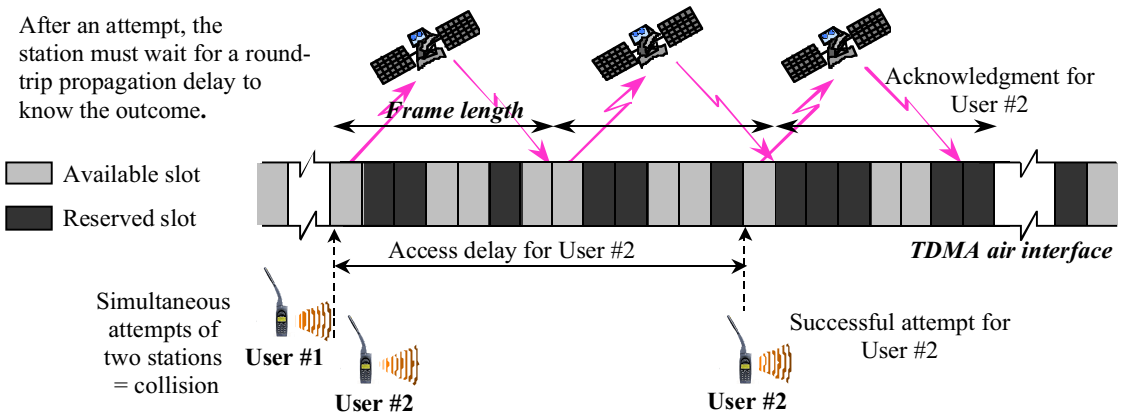

Fig. 2. Transmission attempt example for two voice traffic sources (user#1 and user#2).

Demand-Assignment Protocols

Circuit level

The DAMA technology allows for the dynamic allocation and re-allocation of satellite power and bandwidth based upon the communication needs of the network users. In the description of the DAMA scheme, we assume FDMA access and a bent pipe satellite so that the resource allocation is carried out by an earth master station. For applications where high-speed connectivity is needed, but is not required all the time, it can be efficient to employ DAMA technology rather than a full-time SCPS. Therefore, if a satellite network has multiple sites with voice and data service requirements, but does not have a 24 hour-a-day need for all sites to be in communication with the others, a smaller amount of satellite power and bandwidth can be shared by all stations. This solution allows an efficient use of satellite resources.

By exchanging signalling packets, the process of call set-up involves negotiation between the earth station, the satellite, and a master earth station that controls the satellite network. Once the connection is established, a certain amount of memory and bandwidth onboard is allocated to the new connection. Such a DAMA scheme is well-suited for the MF-TDMA air interface. It can also be used for SCPC systems. In both cases, suitable signalling channels must be used to configure the resource allocation in the network.

Packet level

Many new protocols have been proposed at packet level. Here we briefly mention the DRAMA and the FODA/IBEA (patented) protocols.

The *DRAMA (Dynamic Resource Assignment Multiple Access)* scheme is based on a TDMA frame and contains access slots and information slots. This protocol, originally proposed in [5.15] under the name of Reservation-Aloha (R-Aloha), has also been recently considered in [5.16], referring to a multimedia traffic scenario with packet access to allow an efficient use of satellite resources.

Access slots are mini-slotted: a station sends a short access burst on a mini-slot (at random) to request transmission resources. A persistency probability can also be used to control the access to contention mini-slots; accordingly, a station can decide whether it is allowed to attempt in a frame or it has to reschedule its attempt in the next frame. As soon as a station needs to transmit to the satellite, it waits for the next frame and transmits an access burst, by selecting a mini-slot at random among the mini-slots at the beginning of the frame (if a persistency scheme is employed, this transmission attempt is allowed only if this probabilistic check is fulfilled; otherwise the attempt is rescheduled in the next frame with the same modalities). Provided that the number of mini-slots per access phase is sufficiently high with respect to the number of stations, this access method is particularly efficient: according to the urn combinatorial theory, the collision probability on a mini-slot is quite low.

With the DRAMA protocol, requests coming from the stations are queued at the satellite that serves them according to a suitable scheduling policy. The organization of the transmission is dynamically updated on a frame basis; slot allocations are notified to the stations through a feedback channel (e.g., inserted at the

beginning of the frame). A priority scheme can be implemented to account for the QoS differentiation among the traffic classes. In general, the stations use a piggybacking scheme, in order to communicate new access requests with the information packets that they send on assigned resources to the satellite.
The information slots of the DRAMA TDMA frame can be divided into two parts: slots employed for uplink transmissions (as previously considered) and slots used for downlink transmission. Also for the downlink transmission case, the satellite decides the allocation of resources to the different traffic flows according to a suitable scheduling policy [5.17]. The sharing of TDMA resources between uplink and downlink can be dynamically updated to account for asymmetries in the two directions.
The DRAMA protocol performance in the presence of real-time traffic sources worsens as the frame duration increases. This is due to the fact that terminals have (at most) one access opportunity per frame. If the frame is too long, after a colliding attempt, a new one can be only made after a high delay. This is the reason why DRAMA-like schemes are not well suited for supporting real-time traffic flows in MEO and GEO systems.

The *FODA/IBEA (Fifo Ordered Demand Assignment/Information Bit Energy Adaptive)* scheme was designed to satisfy the requirements of: i) simultaneous transmission of real-time (stream) and non-real-time (datagram) data; ii) maintenance of the QoS close to user requirements; iii) channel optimization in any weather conditions; and iv) cost efficiency robustness [5.18], [5.19], [5.20]. The whole system (TDMA controller hardware plus the access scheme software) was a prototype used in Italian experiments of LAN interconnection via satellite. The system adopts a fade countermeasure technique that dynamically adapts the energy per information bit to each individual link status, which depends on atmospheric conditions. The total attenuation of each link (uplink plus downlink) is compensated for, by varying the transmitting power, the data coding and bit rates. In FODA/IBEA a master station is responsible for system synchronization and for capacity allocation on demand of the stations. These tasks are accomplished by sending a reference burst, which contains the transmission time windows time plan (BTP), at the beginning of each frame (20 ms in the implemented version). A maximum of one window per frame is allocated for each requesting station, in order to save the overhead due to the rather long preambles needed by the modem for the burst synchronization. Inside the transmission window each station sends its multimedia data by adapting the coding and bit rates to the current condition of the entire communication link (transmitting station uplink plus receiving station(s) downlink(s)), according to the BER required by each application. When broadcast and multicast transmissions occur, the worst link status is considered to adapt the transmission parameters. In faded conditions, the channel capacity reserved for datagram transmissions is reduced up to a minimum value in order to maintain the stream sessions already set up. In the implementation performed, the transmission bit rate varied from 1 to 8 Mb/s, and the coding rates used were 4/5, 2/3, 1/2 and uncoded.
In each time frame a small number of *control slots* (transmission windows of fixed size, devoted to transmit short data, such as updates of requests or other control messages) are assigned by the master on a round-robin scheme among all the active stations that did not receive any assignment in the current frame. The

number of the control slots is incremental (1 control slot when the system was sized for up to 8 earth stations, 2 for up to 16 earth stations, 3 for up to 24 stations and so on). The control slots are assigned with fairly high protection (1 Mbit/s, 2/3 coding in the implemented version). The position in the frame of the control slots is not fixed. If all stations have an allocation in a frame, the space devoted to the control slots is added as an additional space for the allocations. Every 32 frames (in the implemented version), a special fixed-size control slot, called FAS (First Access Slot), is present in the frame, in a fixed well-known position (just before the end of the frame); it is used in contention-mode among all those stations that want to become active, thus allowing one new station at a time to enter the network. In the frame where FAS is present, no other control slots are allocated. The FAS is sized as one control slot plus an uncertainty (fixed to ±150 μs) due to the current satellite position with respect to the nominal satellite position. As any other control slot, in the implemented version it is accessed at 1 Mbit/s, 2/3 coded.

Each station manages stream and datagram data by making different requests according to the type of traffic. Allocations for stream traffic, once accepted by the master, are guaranteed until they are released, while allocations for datagram traffic are not guaranteed. As the bandwidth assigned to each station is comprehensive of the portion for stream and for datagram traffic, the total allocation generally varies on a frame-by-frame basis, due to the time variability of the datagram part.

The request sent by a station for datagram is proportional to the datagram traffic coming into the station (*traffic*) plus the volume of data already waiting for transmission to the satellite (*backlog*). Thus,

$$request = backlog + h\ traffic$$

where h is a temporal constant of proportionality. Simulation results, obtained by loading the channel with Poisson generators of datagram traffic for 10 stations indicated a value of 0.4 s for the h parameter as a good trade-off between the average transmission delay at high and low traffic loads [5.20].

The stations issue their datagram requests as frequently as possible, in order to provide the master with the up-to-date information about the incoming traffic. The master organizes all the received requests into a datagram ring, which it scans cyclically to compute the assignments. The length of the assigned transmission window (a) is proportional to the request in a range of values between a minimum (T_{min}) and a maximum (T_{max}) threshold.

$$T_{min} \leq a = f\ r \leq T_{max}$$

where f is the coefficient of proportionality in the assignment. In the implemented version f was set equal to the number of active stations divided by 100, with 5% as minimum and 50% as maximum. T_{min} is introduced for efficiency purposes. This avoids small allocations when the transmission overheads are too big with respect to the information data. T_{max} prevents an overloaded station from removing too much capacity from the other stations. After each assignment, the relevant datagram request in the ring is decreased by the assignment itself and the next request is analyzed, if space is still available in the frame. The ring is not scanned more than once in a frame (assignment cycle). Thus, no more than one assignment cycle is made in a frame.

5.3 Dynamic Bandwidth Allocation

Some of the appealing advantages of satellite networks, such as the wide coverage and the configuration flexibility, make them an ideal candidate for providing multimedia services worldwide. However, satellite bandwidth is a commodity at a premium, and its inefficient utilization may negate some of the aforementioned advantages. To this end, an apportionment scheme, able to dynamically allocate the bandwidth among the satellite terminals while fulfilling the QoS requirements, is of paramount importance. Moreover, the satellite scenario adds a new dimension to the treatment of bandwidth, owing to the presence of variable physical channel operating conditions and of large bandwidth-delay products. Typically, control actions need to be exerted over a wide range of time scales, to cope with events that may occur with frequencies ranging from milliseconds to minutes or hours [5.19], [5.21], [5.22], [5.23], [5.24]. Satellite systems not only have to face variable load multimedia traffic, but also variable channel conditions and large propagation delays. The variability in operating conditions is due both to changes in the traffic loads and to the signal attenuation caused by bad atmospheric events, which particularly affect transmissions in the Ka band.

Nevertheless, efficient bandwidth utilization and QoS provisioning are, unfortunately, two competing goals; therefore, Dynamic Bandwidth Allocation (DBA) schemes seek for a trade-off between QoS provision and bandwidth saving. To address the vast majority of IP traffic, which is inherently bursty, a technique that implicitly evaluates the bandwidth requirement at each satellite terminal and manages the traffic flows on the allocated bandwidth is deemed essential. The purpose of this section is to present a number of solutions for the closely related problems of assigning the satellite bandwidth to different users (earth stations) and traffic types, and to exert CAC actions in the presence of guaranteed-bandwidth inelastic traffic.

The combined action between various layers of the network (from the physical up to the application layer) is likely to be a good way to combat channel variability. However, this procedure is complex and difficult to achieve to the widest possible extent, which would imply numerous cross-layer interactions for control purposes and the related exchange of signaling information. In order to obtain optimized policies for satellite bandwidth allocation, we could coordinate the actions taken in a satellite network at the physical layer (where the fade countermeasure technique is applied) with the functionalities performed at the data link layer (where the satellite bandwidth is allocated), thus obtaining a cross-layer optimization. The complexity of this procedure lies on the rapidly changeable measurements made at the physical layer, regarding the channel state (signal-to-noise ratio), which could produce an unstable allocation at the data link layer. The feedback information should be properly filtered with a proper hysteresis, capable of producing a stable allocation at the data link layer. Regarding resource allocation, another problem is the control network architecture, which could be centralized or distributed. A centralized allocation decision requires a station to play the role of master, also called *NCC*. The master collects all the information about the other stations (slaves) and performs the best choice in the sense of bandwidth allocation. This produces a heavy computational effort for the master. A distributed allocation technique solves the computational problem, but requires a robust

control channel and an efficient control protocol, which takes into account the large communication delay. In this sense, the available bandwidth would be significantly reduced by the signaling protocol.

Two categories of DBA schemes can be distinguished: *static* and a*daptive*. In *static* schemes, once a terminal is assigned a certain amount of capacity, this capacity remains constant for the connection's lifetime and can be handled dynamically, without involving the NCC. That is, the assigned capacity can be apportioned between High-Priority (HP) and Low-Priority (LP) traffic. In the case of *adaptive* schemes, each satellite terminal must send requests to the NCC in order to reserve or release channel capacity, based on its dynamic estimation of bandwidth requirements. To meet the QoS requirements of bursty and delay-sensitive traffic three approaches have been proposed:

- fixed allocation proportional to the maximum source rate;
- fixed allocation at a given rate using DBA for peak bursts;
- full DBA techniques.

As previously stated, the first approach is inefficient for satellite systems. Besides, the maximum source rate is usually unknown. As regards the full DBA techniques, these can exploit the channel capacity quite efficiently, since no capacity is reserved during inactive periods. Nevertheless, the common signaling channel may become overloaded during acquisition phases, leading to increased delays and congestion. Consequently, a mixed approach seems to be the most flexible choice, where each terminal is assigned some fixed channels of moderate capacity, while a number of DBA channels are used during peak traffic periods.

As far as adaptive schemes are concerned, one of the challenging problems that engineers must grapple with is the implementation of this technique in a GEO satellite system. The main problem stems from the long delay between the moment that a request is sent to the NCC and the moment at which the satellite terminal is informed about the bandwidth that has been allocated to it. This latency prevents the capacity from changing immediately. Since a low latency results in better performance, a GEO satellite system represents the worst case (about 500 ms when the NCC is on ground or 250 ms when the majority of processing is implemented on-board the satellite).

Adaptive DBA schemes are generally categorized as either *Reactive* or *Proactive* algorithms. *Reactive* schemes take into account the current queue length, packet loss, and average delay in order to react to traffic fluctuations, without trying to anticipate them (e.g., PRMA schemes in LEO satellite systems, where time slots are allocated to users on demand). Compared to Proactive algorithms, Reactive ones are easier to implement and can more efficiently utilize the channel capacity. However, QoS requirements are not easily met, since the requests (sent to the NCC) that represent the current needs for bandwidth are not always satisfied. In [5.25], the authors proposed a novel predictive bandwidth allocation and de-allocation scheme, which frees up bandwidth allocated to connections that are unlikely to be used. The look-ahead horizon of k cells is introduced, where $k = 2$. The scheme provides the lowest connection delay probability for real-time connections comparing with previous schemes. Even though Reactive schemes may perform well in LEO satellite networks, they are not well-suited to GEO systems on account of the high propagation delay.

Proactive schemes aim to analyze the traffic and predict the required bandwidth. Usually this is realized by providing a predictor with data (e.g., with the queue lengths, the input flows and output flows) up to time *t*, which in turn makes a prediction at time *t* of the aggregated traffic in the time interval [*t, t+k*) (e.g., the traffic within the next superframe), mainly based on statistical properties of IP traffic. Hence, the required bandwidth can be estimated. In order to make predictions adaptive, i.e., capable of following changes of the traffic characteristics in time, the parameters of the predictor can be regularly updated. The performance of these schemes relies heavily upon the accurate prediction of future traffic. On the basis of a fair policy of resource sharing among all the satellite terminals, the NCC receives the bandwidth requests of each terminal and decides whether or not to satisfy these requests. In order to meet the desired QoS, both the request algorithm and the NCC allocation strategy are of paramount importance. In [4.1] and [5.26] the authors compared some different allocation strategies based on traffic prediction.

In LEO and MEO constellations, the handoff problem can also affect the QoS of connections. In [4.2] bandwidth reservation for handoff is allocated adaptively by calculating the possible handoffs from neighboring beams, on the basis of users' location information. The reservation mechanism provides a low handoff blocking probability as compared with fixed guard channel strategy. However, employing user location information does not seem reasonable, since updating locations would cause high processing load to the on-board handoff controller and increase the complexity of terminals. The method seems only suitable for fixed users. In [4.4], the authors have introduced two different mobility models for satellite networks. In the first model, only the motion of satellites is taken into account, whereas in the second model other motion components, such as earth rotation and user movements, are considered. The key idea of the algorithm is that, to prevent dropped handoff during a call, bandwidth is reserved in a particular number *S* of spot-beams that the call would handoff into. In [5.27], a probabilistic resource reservation strategy for real-time services was proposed. The concept of sliding windows is proposed to predict the necessary amount of reserved bandwidth for a new call in its future handoff spot-beams. For real-time services, a new call request is accepted if the originated spot-beam has available bandwidth and resource reservation is successful in future handoff spot-beams. Non-real-time service new call requests are accepted if the originated spot-beam satisfies its maximum required bandwidth. In [5.28], the authors proposed a selective look-ahead strategy specifically tailored to meet the QoS needs of multimedia connections where real-time and non-real-time service classes are differently treated. The handoff admission policies introduced distinguish between both types of connections. Bandwidth allocation only pertains to real-time connections' handoffs. To each accepted connection, bandwidth is allocated in a look-ahead horizon of *k* cells along its trajectory, where *k* is referred to as the depth of the look-ahead horizon. The algorithm offers low connection delay probability, providing for reliable handoff of on-going calls and acceptable connection delay probability for new calls.
Moreover, a yet-undeveloped problem could arise from satellite-based meshed architectures. So far, the system model only considers the uplink part, relying on the assumption that the downlink is not a bottleneck. In a meshed architecture with multiple, limited bandwidth downlink spot-beams, in order to maintain the

overall QoS, the channel allocation has to take into account this point as well, which is particularly interesting in satellite-based switching systems.

5.4 Optimization techniques

Whenever the bandwidth allocation problem is posed in terms of the optimization of a certain performance index (revenue maximization or cost function minimization), some optimization technique must be employed in its solution. Satellite bandwidth allocation, possibly jointly with CAC in the presence of guaranteed bandwidth connections, can be formulated in many instances as such. In this respect, the satellite environment would not be much different from the cabled one, with the notable exception of the presence of possible adaptive fade countermeasures at the physical layer, which may effectively change (even over a short time scale) the bandwidth need of traffic that has been accepted and is ongoing in any given station. This creates the previously mentioned cross-layer interaction, and can be taken into consideration in the formulation of the overall optimisation problem. Indeed, among possible fade countermeasures, based on diversity, power control, FEC or transmission bit rate adaptation (among others), the latter produce variations in the redundancy applied to the data with respect to a clear-sky situation, thus varying the amount of bandwidth needed to achieve the same net Information Bit Rate (IBR).

In general, the optimization problems to be considered will also depend on the traffic models and on the performance indicators adopted. Moreover, they can be posed in terms of functional or parametric optimization, in either discrete or continuous variables. We briefly review these points in the following.

i) As regards traffic representation at the flow (also referred to as *connection* or *call*) level, usually real-time, with some bandwidth guarantees, it can be done exactly as in a circuit-switched environment, as far as the bandwidth allocation is concerned. This traffic may originate either as Continuous Bit Rate (CBR) guaranteed-bandwidth connections (voice or MPEG4 video), as EF (Expedited Forwarding) class in an IP DiffServ environment, as Guaranteed Service IP IntServ connections, or as MPLS (Multi Protocol Label Switching) requests for a LSP (Label Switched Path), along with its bandwidth requirement. It may even represent AF (Assured Forwarding) IP DiffServ classes or IntServ Controlled Load Service connections, if a suitable Effective Bandwidth can be assigned to them. Whichever the case, bits generated by these flows, whether unstructured or organized into packets, will eventually be carried within some specific DVB class, and identified by a PID (Program IDentifier). The dynamics of interest are at the connection level, and the relevant performance index is the call blocking probability (P_{block}); this is the steady-state probability that an arriving “call” (actually, a request for bandwidth) is refused because all the bandwidth devoted to the real-time traffic is busy. It is usually assumed that blocked calls are lost (not re-attempted); in practice, this means they are re-attempted after an amount of time such that they can be considered to be uncorrelated to the previous ones. For this type of traffic, the usual birth-death model with exponential distribution of call inter-arrival and duration times (Poissonian traffic) is adopted. If we assume that all connections sharing a fixed amount of bandwidth have the same peak rate, we face a particular single-class case, where the expression of

the blocking probability is given by the classical Erlang B loss formula [5.29]. If a certain amount of bandwidth is allocated to a station, independently of the others, at station i, given the Erlang traffic intensity $\rho^{(i)} = \lambda^{(i)} / \mu^{(i)}$ (where $\lambda^{(i)}$ [s^{-1}] is the arrival rate of the connection requests, and $1/\mu^{(i)}$ [s] is the average duration of each connection), the peak rate $B^{(i)}$ and a desired upper bound on the blocking probability $\eta^{(i)}$, the maximum number of acceptable calls $N_{\max}^{(i)}$ can then be derived.

More generally, either the whole system (if the available bandwidth is not partitioned among stations) or some stations might be in the multiclass case, where the connections that utilize a given bandwidth portion have different statistical parameters and peak rates (or they are assigned different data redundancies, as they are addressed to destination stations that experience diverse downlink attenuations). In this multiclass case, the blocking probability would result from a stochastic knapsack problem [5.29]. The situation can be maintained in the single class case if the bandwidth is assigned separately per fading or traffic class inside the earth station.

For what concerns best-effort traffic, there are two main categories that can be considered in modeling: i) TCP elastic traffic (with particular regard to long-lived connections) and ii) short-lived TCP connections and UDP traffic originating from applications that do not require bandwidth reservation. As regards the second category, it usually stems from the aggregation of packet bursts, generated by a high number of sources, whose data packets are fragmented into fixed-size cells (ATM or DVB) before transmission on the satellite channel. Assuming that at each station *i* cells are queued in a finite buffer of capacity $Q^{(i)}$, the quantity of interest is the cell loss probability (P_{loss}) in the queue of each station. In order to derive an approximate evaluation of this quantity, a discrete-time self-similar traffic model can be considered, as in [5.30]. This model represents the superposition of on-off sources, whose active periods (bursts) have Pareto-distributed 'on' time *t* ($\Pr\{\tau = l\} = c\, l^{-\alpha-1}$, $1 < \alpha < 2$, where α and *c* are the parameter of the discrete Pareto distribution and its normalization constant, respectively). The detailed description of the model, which yields an upper bound on P_{loss}, can be found in [5.34, 5.35]. Actually, various possible models can be adopted to approximate the cell loss probability, given the statistical characteristics of the burst generation and a fixed rate of extraction of cells from the buffer [5.35, 5.36, 5.37].

In a mixed environment of best-effort and guaranteed bandwidth flows, the performance index may be a cost associated with blocking and cell loss probabilities, or some revenue function that can be expressed in those terms or, more generally, in terms of state probabilities (either stationary or non-stationary) provided by the analytical models.

Still another interesting representation, which has been recently used [5.38, 5.39], is based on fluid models, where the dynamics at the level of buffers serving a link are modeled by a simple differential equation, and the loss volume is the performance index. The latter lends itself to interesting parametric optimization problems, where the functional form of the performance index may be unknown, and the optimization is based on on-line measurements only.

ii) The optimization problems related to finding the bandwidth allocation and CAC can be posed as functional ones, i.e., optimal mapping is sought between the available information and the control action. This approach often entails formidable difficulties, especially if the dimension of the state space is high. Therefore, especially as regards CAC (see also subsection 5.4 below), restrictions of the problem are often considered, where the functional form of the mapping is fixed *a priori* (e.g., a threshold policy), and only parametric optimization is performed. Moreover, in the case that information is transmitted in the form of packets (or DVB cells), the Service Separation principle [5.29] is often applied, in order to decouple low (e.g., on DVB cell loss) and high-level requirements (e.g., on blocking probability of connection requests). In this context, hierarchical strategies, where a master station allocates bandwidth to the traffic stations, and the latter independently apply CAC (and, possibly, bandwidth sharing among different traffic types - e.g., guaranteed bandwidth and best-effort), have been considered in [5.40, 5.41, 5.42, 5.33]. In these, as well as in other works, the bandwidth allocation was formulated as an optimization problem in a discrete setting (with the assignment's granularity determined by the minimum bandwidth unit – *mbu*); if the performance index is a separable function of the stations' parameters (e.g., a sum of terms, each depending only on the bandwidth to be assigned to a station), the problem can be solved numerically by applying Dynamic Programming [5.42], [5.33]. Moreover, within a cross-layer approach that takes into account fade countermeasures based on adaptive code and bit rate (ultimately, redundancy) assignment, the allocation can be made in either static or dynamic fashion. In the first case, the performance measure is averaged over the distribution of different forecasting scenarios, whose probability is derived from long-term statistics [5.39]. In the second case, on the basis of the measured fading attenuation values, a receding-horizon open-loop feedback control problem is stated, with the optimization performed on-line upon request from the stations [5.30]. Finally, it is worth noting that these model-based approaches can be by-passed by using a fluid approximation and by treating the bandwidth partitions as continuous variables. A gradient descent technique can be adopted, in conjunction with Infinitesimal Perturbation Analysis (IPA) [5.35] for gradient estimation [5.36].

5.5 Call Admission Control (CAC)

The public data network provides a resource that could profoundly impact high-priority activities for society, such as defence and disaster recovery operations [5.40]. Under stress, however, the public network has proved to be a virtually unusable resource [5.41]. Today's public network resource allocation mechanisms do not prioritise the way they allocate resources, working instead on a first-come-first-served basis. Loads on public networks reach up to five times the normal level during an emergency [5.42], and important traffic receives equally poor access to resources as low-priority traffic. In these, as well as in other multiservice cases, where a traffic differentiation is desired, call admission control (CAC) can improve the network performance considerably.

In many studies of resource allocation, a simple *complete sharing* (CS) policy is used, i.e. connections are admitted simply if sufficient resources are available at the time of the request, without considering the

importance of a connection when they are allocated. In the CS policy, the only constraint on the system is the overall capacity *C*; connections that request fewer resource units are more likely to be admitted (e.g., a voice connection will more likely be admitted compared to a video connection). As an almost opposite situation, in the set of policies of *complete partitioning* (CP) type, every class of traffic is allocated a set of resources that can be used only by that class.

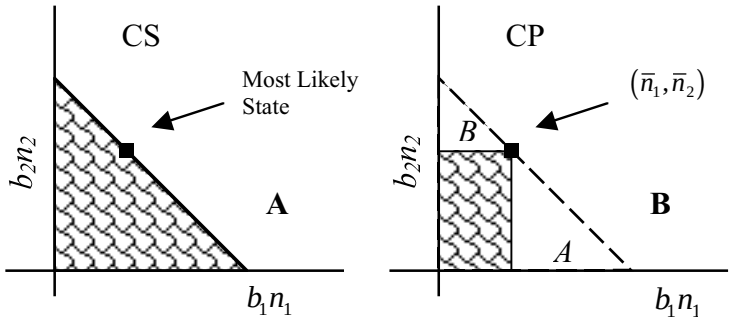

Fig. 3. Complete Sharing Vs Complete Partitioning feasibility regions.

Other policies have been derived to provide optimized access to resources, and Ross [5.29] provides an extensive discussion regarding a number of different solutions. As previously mentioned, optimal approaches should be based on Markov decision processes, given a certain cost functional to be minimized (or maximized) as a performance index; however, they must take into detailed account any allowable network state and state transition, which is impractical even for networks of modest complexity. The functional form of the optimal policies is usually unknown. Therefore, a set of generally suboptimal policies with fixed structure (which often can be described by a set of parameters), have been developed, which are simpler to implement and, in some special cases, do correspond to the optimal one: among others, the above mentioned CP, trunk reservation (TR) [5.43], guaranteed minimum (GM) [5.44], and upper limit (UL) policies [5.47, 5.48]. Comparisons have been made between these policies and the optimal one. The results indicate that the CP, TR, GM, and UL policies outperform the CS policy when significant differences between classes exist in requirements for bandwidth and offered load [5.46]. Figure 3 illustrates CS and CP policies.

5.6 DVB

The interest in digital TV has led to the *Digital Video Broadcast* (DVB) standards effort, which began in 1993. Key resulting DVB standards cover satellite (DVB-S) and terrestrial (DVB-T) delivery. In particular, recent DVB standards have defined satellite return channels (DVB-RCS) and terrestrial channels (DVB-RCT). While these return channels may support interactive TV (iTV), they also enable other telecommunications services over DVB infrastructure, including telephony and Internet access. Following successful demonstrations of broadcast quality compressed video in the late 1980s, a European digital television standardization effort began in the early 1990s [5.47]. This led to the DVB Project, formed in September 1993 and now comprising over 200 organizations. The DVB project, in conjunction with ETSI,

has produced a wide range of standards for cable, terrestrial and satellite Digital Television (DTV) services. The DVB-S standard, released in December 1994 [5.48], defines the framing, modulation and encoding schemes for satellite transmission of MPEG 2 streams. These comprise fixed-length (188 byte) packets, which may carry MPEG 2 video, audio, or data. The packets have RS FEC added, followed by interleaving, then convolutional coding. The bit rate of the resulting stream is a function of the transponder bandwidth (26 to 54 MHz are common ranges) and the coding rate. As outlined in [5.48], the bit rate for a 54 MHz transponder ranges from 38.9 Mbps (1/2 rate coding) to 68 Mbps (7/8 rate coding). ETSI has recently standardized DVB satellite return channels [5.49]. The standard outlines encoding, security and MAC protocols for these channels, which operate at rates of up to 2 Mbps [5.50]. Return channels comprise either ATM cells or 188 byte MPEG2-TS packets. Proprietary satellite return channel systems also exist, such as the Gilat Skyblaster [5.51]. The terrestrial DVB standard, DVB-T, appeared at the end of 1995. A key aim was to maximize commonality with DVB-S and DVB-C (cable) standards; hence, the coding and interleaving schemes have been reused [5.52]. OFDM transmission has been used, based on either 1705 carriers (known as 2K) or 6817 carriers (8K). Individual carrier modulation can be either QPSK, 16-QAM or 64-QAM, leading to potential tradeoffs between bit rate and robustness to noise. In order to avoid interference with existing analogue TV transmissions, DVB-T has been designed to operate at a power level that is 20 dB less than analogue TV systems [5.53]. The DVB-T return channel, known as DVB-RCT, was published by ETSI in March 2002 [5.54]. It is designed to support low power client transceivers (1W), with a potential aggregate bit rate of 30 Mbps within a given DVB-RCT channel [5.55]. While a 65 km maximum cell radius has been envisaged, DVB-RCT cell radii of 80 km have been successfully demonstrated [5.55]. In addition to defining forward and return channel characteristics, the DVB standards also specify operational features, such as service information (to indicate transmission parameters and context), teletext support, and the conditional access needed for Pay TV services [5.47].

DVB-S2

DVB-S was introduced as a standard in 1994 [5.48] and DVB-DSNG in 1997 [5.56]. The DVB-S standard specifies QPSK modulation and concatenated convolutional and Reed-Solomon channel coding, and is now used by most satellite operators worldwide for television and broadcasting services. DVB-DSNG (Digital Satellite News gathering) specifies, in addition to DBS format, the use of 8PSK and 16QAM modulation for satellite news gathering and contribution services. Since 1997 digital satellite transmission technology has evolved, and DVB-S2 is the latest advanced satellite transmission technique from DVB [5.57]. It makes use of the following improvements in the digital satellite transmission technology:

a) new coding schemes, which, combined with higher order modulation, promise more powerful alternatives to DVB-S/DVB-DSNG coding and modulation schemes. The result is an efficiency 30% greater than DVB-S;
b) variable coding and modulation (VCM), which may be applied to provide different levels of error protection to different service components. In the case of interactive and point-to-point applications, the VCM functionality may be combined with the use of return channels, to achieve adaptive coding

and modulation (ACM). This technique provides more exact channel protection and dynamic link adaptation to propagation conditions, targeting each individual receiving terminal.

While DVB-S and DVB-DSNG are strictly focused on a unique data format, the MPEG transport stream, DVB-S2 utilizes extended flexibility to cope with other input data formats (such as multiple transport streams or generic data formats without significant complexity increase. It improves on and expands the range of possible applications, by combining the functionality of DVB-S (for direct-to-home applications), and DVB-DSNG (for professional applications), and techniques such as adapting coding to maximize the usage of value satellite transponder resources. DVB-S2 is optimized for the following broadband satellite applications:

- broadcast services (BS) digital multi-programme television (TV)/high definition television (HDTV);
- interactive data services, including Internet access;
- digital TV contribution and satellite news gathering. Digital TV contribution applications by satellite consist of point-to-point or point-to-multipoint transmissions, connecting fixed or transportable uplink and receiving stations. They are not intended for reception by the general public.
- Data content distribution/trunking and other professional applications. These services are mainly point-to-point or point-to-multipoint, including interactive services to professional head-ends, which re-distribute services over other media. Services may be transported in (single or multiple) generic stream format. However, DVB-S2 is compatible with MPEG-2 and MPEG-4 coded TV services.

DVB-RCS

DVB-RCS (Return Channel via Satellite) [2.5] provides a number of bandwidth allocation methods, the most important being the rate-based and volume-based ones. The standard does not impose strict constraints on the algorithms to be used in the process; hence, it is possible to develop advanced techniques using the standard request types. The only standard weakness is related to the lack of information about the requests; hence, two requests of the same type will have to be considered as equal even if the requesting terminals have to deliver different kinds of traffic (e.g., volume based requests for high-priority and low-priority traffic).

Future improvements in DVB-RCS based allocation strategies will need to focused mainly on two topics, both related to a cross-layer approach. The first is to consider the effects of fading countermeasures, the second one consists of defining a simple interface for upper layers in order to develop a cross-layer QoS manager, able to tune the allocation process to the actual QoS requirements, possibly considering a pricing system, i.e., taking into account the user's willingness to pay.

6. PHYSICAL LAYER: FEC, ARQ, fade countermeasure, handoff

6.1 FEC

New generation satellite systems allow the performance of mechanisms to be adapted to the environment in which they operate. There is a wide range of possible adaptations. At the simplest level, modulation and

coding may be adapted, based on fixed properties (such as the terminal size, terminal location, statistical propagation data, etc.). Using several coding/modulation schemes within the same satellite network can result in significant saving in power/bandwidth (e.g., lowering the requirement for terminals located in the center of a satellite footprint). Although communication between link layer (L2) and physical layer (L1) is required, usually there is no significant impact on the network layer (L3) operation, since this sees a more or less fixed link characteristic, at least for the duration of most IP flows. Mobile terminals that travel over a wide geographic area will require hand-over procedures to accommodate the changes in propagation conditions.

The traditional way to design a link is to specify the physical, link and network properties separately. However, overall performance can be improved by considering the operation of several protocol layers together. A classic example of this *cross-layer* approach is the use of hybrid ARQ (*Automatic Repeat reQuest*) [6.1] where end-to-end data reliability is provided by a combination of physical layer (modulation and coding), link layer (framing and ARQ) and transport layer (TCP error control [6.2], [2.2], [6.3,], [6.4]). Satellite link designers can, and do, trade-off transmit power against protocol mechanisms to achieve the required overall performance. Further research on this topic is required to clearly identify the implications on satellite system design. In most current generation systems, the trade-off occurs during the design phase.

At the physical layer, the BER (bit error rate) can be improved by using data coding, which adds redundant bits to a data stream in such a way that an error in the data stream can be detected (*error detection codes*) and corrected (Forward Error Correction, FEC). Among many types of data codes (Linear block, Binary cyclic, BCH (Bose-Chaudhuri-Hocquenghem), Golay, Convolutional, Punctured, Erasures, Turbo, LDPC (Low Density Parity Check), etc), a great interest is recently dedicated to the turbo and LDPC codes, which seem to have great performances, as they almost reach the Shannon limit.

ARQ is an error-control system, where the receiver requests for a re-transmission when an error in transmission is detected. In a simple system, a positive acknowledgement (ACK) is returned when the data is received correctly in the right sequence and a negative acknowledgement (NACK) may be returned when an error is detected, i.e., an out-of-sequence data block is received. There are three main types of ARQ: *stop-and-wait*, *go-back-N*, and *selective-repeat.*

In *stop-and-wait ARQ*, the sender transmits a block of data and then waits for an ACK before sending the next block (Fig. 4).

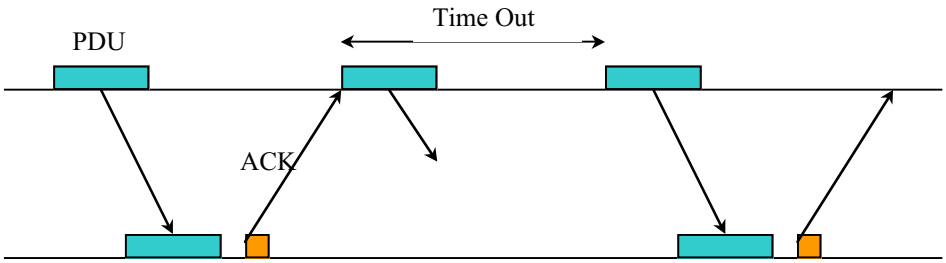

Fig. 4. Stop and wait ARQ.

When a block of data is lost, the receiver is unable to identify the loss (as it does not receive anything); it may or may not send a negative acknowledgement upon detection of an error. Unless the NAK is generated and received, the transmitter must then rely upon a timer to detect the lack of a response (Time Out). Stop-and-wait ARQ is simple, but it is inherently inefficient due to the idle time spent waiting for an ACK after each transmitted data block. This inefficiency becomes unacceptable when the delay-bandwidth product is large, as in satellite channels.

In a *go-back-N ARQ system* (Fig. 5), the transmitter continuously sends data blocks to the receiver, which sends ACKs or NACKs back to the transmitter. On receiving a NACK (or upon a time out expiring), the transmitter re-transmits the data block that was detected in error by the receiver (or timed out), and subsequent data blocks transmitted in the interval between the original transmission and the receipt of the NACK or timeout expiration. At the receiver, the data blocks following the erroneously-received data block are discarded even if they are correctly received.

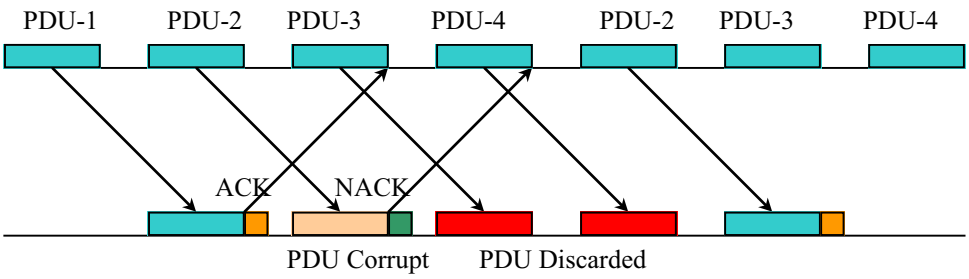

Fig. 5. Go-back-N ARQ

In a *selective repeat ARQ system* (Fig. 6), the sequence between transmitter and receiver is the same as in go-back-N systems. The difference is on receiving a NACK; here the transmitter re-transmits only the data block that was detected in error because the receiver stores any correct frames following the damaged one.

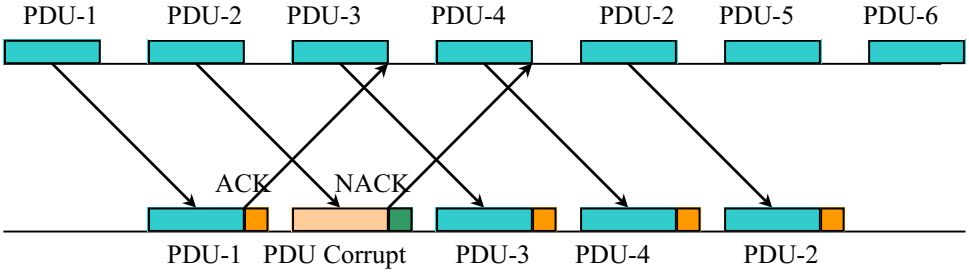

Fig. 6. Selective repeat ARQ.

6.2 The handoff

A handoff is the changeover of the physical radio channel without interrupting the connection. The handoff concept contributes to provide high-quality connections to mobile users. Thus, the handoff process is an essential part of QoS in mobile communication networks. It is typical in LEO satellite systems, as their

coverage area changes continuously. The coverage area of a satellite - a circular area of the surface of the Earth - is referred to as *footprint*. For special reasons of efficiency the satellite footprint is partitioned into slightly overlapping cells, called *spot-beams*. In order to maintain connectivity end-users must switch from spot-beam to spot-beam and from satellite to satellite, resulting in frequent intra- and inter- satellite handoffs. The satellite footprints/spot-beams are overlapped: thus, there is a time window for preparing and executing the handoff event.

To summarize, handoffs in mobile satellite communications are due to satellites or mobile stations' movements while being connected [6.5]. The connection parameters can change while using the service, e.g., the elevation angle can be lower than *min*. The quality of the signal declines under a certain value and the connection between mobile station and satellite cannot be maintained. In this case an *inter-satellite handoff* is necessary in order to select another satellite with a greater elevation. If too many mobile stations use the same radio channel, channel interferences can decrease the signal quality. Thus, the radio channel (carrier frequency) must be changed; this event is called *intra-satellite handoff.*

In the following an inter-satellite handoff strategy is briefly described.

Maximum elevation angle. At a handoff process, the satellite will be selected that has the maximum elevation angle to the mobile station. In this case several handoffs are necessary, but the best mobile link (signal power, signal quality) will be achieved.

To minimize the handoff events, the signal is transferred to another satellite only when the minimum elevation angle of the current satellite is reached. The selected satellite will be the longest reachable satellite with the constraint that the elevation angle must be greater than or equal to the minimum one. In this case the channel parameters undergo significant changes, possibly resulting in unacceptable quality.

Handoff management is the process of initiating and ensuring a seamless and loss-less handoff of a mobile user. The time required to effect the handoff should be appropriate for the rate of mobility of the mobile user. The management procedure should attempt to maintain the requested QoS after the handoff is completed, providing for reliable handoff of calls in progress, e.g., low Call Dropping Probability for handoff calls, acceptable Call Blocking Probability (CBP) for new call attempts and high resource utilization. A handoff management technique, queuing of handoff, which prioritizes handoff calls over new calls, has been widely studied [5.8], [6.6]-[6.10] to improve the service quality of calls that are in progress.

Further information can be found in references [6.11] and [6.12] for other handoff techniques or considerations regarding this problem.

6.3 Fade countermeasure techniques and signal quality estimation

Many methods to counteracting rain attenuation have been proposed in [5.19], and [6.13]-[6.20]. Some of these methods, such as space or frequency diversity ([6.17], [6.18], and [6.19]), permit a very high-level of link availability, but they are very complex and expensive. Other methods, based on the dynamic adjustment of the energy per information bit ([6.13], [6.14], and [5.19]), can be employed when a moderate level of link

availability is required. They are not very complex and are cost-effective. These methods require a modem that is able to change the transmitting power and the data bit rates, and a convolutional encoder/Viterbi decoder with puncturing, in order to realize variable coding rates.

In multiservice broadband networks, in order to support different classes of traffic, characterized by diverse statistical nature and QoS requirements, in the presence of limited resources (buffers, bandwidth, or processing capacity), several forms of control are exerted to maintain a desired level of performance for all users and traffic types. Moreover, to cope with possible variations in bandwidth demand and offered load, control actions should be devised to be dynamic, based on instantaneous and past information, or, at least, adaptive in nature [6.20]. This control scenario meets with still another one in networks involving satellite channels. Here, dynamically varying fade conditions, caused by adverse atmospheric events (e.g., rain, hail or snow) can heavily affect the transmission quality, especially when working in Ka band [20–30 GHz], unless adaptive fade countermeasure techniques are adopted. In essence, fade countermeasures address the physical layer requirement of keeping the BER below a given threshold, whatever the channel degradation may be, within a certain operating range, beyond which the station is declared to be in outage conditions.

All fade countermeasure systems require an attenuation meter that can make accurate estimates of signal degradation in real time, in order to trigger the countermeasure in a timely fashion. When the transmission power control is used to compensate for uplink attenuation, the latter must be estimated separately from the downlink attenuation. A traditional method for measuring the signal attenuation is to use beacon receivers at the earth stations. This requires beacon transmitters at frequencies that are very close to the signal frequency bands. Two beacon transmitters at the satellite and two receivers at each earth station are necessary if one needs to distinguish between the attenuations on the up- and downlinks. The big drawback of this method is the high cost of the hardware required. The other disadvantage is that the measurement of the attenuation is made out of signal band, and this leads to inaccuracies. In fact, even if the beacon and the signal frequencies are very close together, there is always some lack of correlation between the relative attenuation values.

Another method for estimating attenuation is by measuring the power level of the received signal, which, given a clear-sky reference level, depends on the up- and downlink attenuations. For a digital modem, this usually requires little additional hardware, because the modem already needs an instantaneous measure of the power level received, which is made with a fast AGC (Automatic Gain Control) in order to demodulate the data. This method is very cheap and does not lead to inaccuracies due to the measurement frequency offset, but it does not allow a separate evaluation of the up- and down-link attenuations.

An alternative way to estimate signal quality is to evaluate the Signal-to-Noise Ratio (SNR), from which an evaluation of the BER is straightforward [6.21]. This method is much more accurate than the signal attenuation measurement, since it takes into account the noise level variation due to the attenuation, i.e., the changes in sky temperature. Furthermore, SNR evaluation methods consider the effect of the total noise, including interference. Not even by using this method can the contributions of the up- and down-link attenuations be distinguished.

The SNR evaluation methods referred to in [6.22] deal with considering the power distribution of the received signal around its mean value. This measure can be made with very little additional hardware, provided that the receiver is equipped with soft decision levels of data demodulation. With respect to measuring the received power level, this method has the additional advantage of being independent of any reference level, thus requiring no tuning by an operator. Thus, it can be useful for end-user equipment, such as mobile, nomadic, or hand-held terminals, where ease of use is an essential requirement.

The dynamic adaptation of the fade countermeasure system must take into account the latency due to the transmission delay between the stations. This latency needs the use of a suitable power margin on the estimated channel conditions. The performance evaluation of a centralized control adaptive fade countermeasure system is given in [6.23].

7. Conclusions

This paper is far from being exhaustive on the subject of radio resource management. Nevertheless, the authors hope to have provided useful aid to the interested readers by touching the most relevant points. These have covered aspects related to almost all layers of the protocol architecture, and in some cases have touched issues of cross-layer interaction and optimisation. Besides the literature surveyed in some more detail (which is anyway a small fraction of what is available), some additional reading is furnished for integration of the topics discussed.

REFERENCES

[2.1] M. Karaliopoulos, P. Henrio, E. Angelou, B. G. Evans, “Packet scheduling for the delivery of multicast/broadcast services via S-UMTS”, in *Proc. 1st Internat. Conf. on Advanced Satellite Mobile Systems*, Frascati, Italy, July 2003.

[2.2] “Preparation of Next-Generation Universal Mobile Satellite Telecommunication Systems (S-UMTS) – Executive Summary” (Aug. 2000), ALCATEL Space Industries, France, ESA CR(P)-4309.

[2.3] M. Allman, V. Paxson, W. Stevens, “TCP congestion control ”, IETF, RFC 2581, April 1999.

[2.4] ETSI EN 302 307 (V 1.1.1), “Digital video broadcasting (DVB); second generation framing structure, channel coding and modulation systems for broadcasting, interactive services, news gathering and other broadband satellite applications”.

[2.5] ETSI EN 300 421 (V 1.1.2), “Digital video broadcasting (DVB); framing structure, channel coding and modulation for 11/12 GHz satellite services”.

****

[3.1] ITU-T Recommendation G.114 "International telephone connections and circuits – General Recommendations on the transmission quality for an entire international telephone connection”.

[3.2] P. Serrano, C. J. Bernardos, I. Soto, J. I. Moreno, “Medida y análisis del tráfico multimedia en redes móviles de cuarta generación”, Telecom I+D, Madrid, 2004 (in Spanish).

[3.3] 3GPP TS 22.105 V6.2.0 (2003-06) Technical Specification 3GPP; Technical Specification Group Services and System Aspects; WG1 Services; Services and Service Capabilities (Release 6).

[3.4] C. J. Bernardos, I. Soto, J. I. Moreno, T. Melia, M. Liebsch, R. Schmitz, "Experimental evaluation of a handover optimization solution for multimedia applications in a mobile-IPv6 network", *Europ. Trans. Telecommun.* (to appear).

[3.5] Moby Dick Project. Mobility and Differentiated Services in a Future IP Network. http://www.ist-mobydick.org/

****

[4.1] M. Allman, V. Paxson, W. S. Stevens, "TCP congestion control", RFC 2581, Apr. 1999.

[4.2] M. Allman, S. Floyd, C. Partridge, "Increasing TCP's Initial Window", RFC 3390, Oct. 2002.

[4.3] M. Allman, "TCP congestion control with Appropriate Byte Counting (ABC)", RFC 3465, Feb. 2003.

[4.4] M. Gerla, A. M. Y. Sanadidi, R. Wang, A. Zanella, C. Casetti, S. Mascolo, "TCP Westwood: congestion control using bandwidth estimation", in *Proc. IEEE Globecom 2001*, S. Antonio, TX, Nov. 2001.

[4.5]. J. Border, M. Kojo, J. Griner, G. Montenegro, Z. Shelby, "Performance enhancing proxies intended to mitigate link-related degradations", IETF, RFC 3135, June 2001.

[4.6] http://www.scps.org/scps/

[4.7] J. Ishac, M. Allman, "On the performance of TCP spoofing in satellite networks," in *Proc. IEEE Milcom*, Vienna, VA, Oct. 2001.

[4.8] H. Balakrishnan, R. H. Katz, "Explicit Loss Notification and wireless web performance", in *Proc. IEEE Globecom Internet Mini-Conf.*, Sydney, Australia, Nov. 1998.

[4.9] I. F. Akyildiz, G. Morabito, S. Palazzo, "TCP-Peach: a new congestion control scheme for satellite IP networks", *IEEE/ACM Trans. Networking*, vol. 9, no. 3, June 2001.

[4.10] N. Celandroni, F. Potortì, "Maximising single connection TCP goodput by trading bandwidth for BER", *Internat. J. Commun. Syst.*, vol. 16, pp. 63-79, 2003.

[4.11] C. Barakat, E. Altman, "Bandwidth tradeoff between TCP and link-level FEC", *Computer Networks*, vol. 39, no. 2, June 2002, pp. 133-150.

[4.12] N. Celandroni, F. Davoli, E. Ferro, A. Gotta, "A dynamic cross-layer control strategy for resource partitioning in a rain faded satellite channel with long-lived TCP connections", *3rd Internat. Conf. on Network Control and Engineering for QoS, Security and Mobility (Net-Con 2004)*, Palma de Mallorca, Spain, Nov. 2004; in D. Gaïti, S. Galmés, R. Puigjaner, *Network Control and Engineering for QoS, Security and Mobility, III*, Springer, New York, NY, 2004, pp. 83-96.

[4.13] N. Celandroni, F. Davoli, E. Ferro, A. Gotta, "Adaptive bandwidth partitioning among TCP elephant connections over multiple rain-faded satellite channels", to be presented at *3rd Internat. Workshop on QoS in Multiservice IP Networks*, Catania, Italy, Feb. 2005.

[4.14] http://www.ietf.org/html.charters/ipsec-charter.html

****

[5.1] J. G. Puente, W. G. Schmidt, A. M. Werth, "Multiple access techniques for commercial satellites", *Proc. IEEE*, vol .59, pp. 218-229, 1971.

[5.2] T. Pratt, C. W. Bostian, "Satellite Communications", John Wiley & Sons, 1986.

[5.3] J.D. Gibson, "The mobile communications handbook", IEEE Press, 1996.

[5.4] H. Peyravi, “Medium Access Control protocols performance in satellite communications”, *IEEE Commun. Mag.*, vol. 37, no. 3, pp. 62-71, March 1999.

[5.5] A. Andreadis, G. Giambene. Protocols for High-Efficiency Wireless Networks. Kluwer Academic Publishers, Nov. 2002.

[5.6] N. Abramson, “The ALOHA system - Another alternative for computer communications”, in *Proc. AFIPS Fall Joint Comput. Conf.*, vol. 37, 1970.

[5.7] S. Kota, M. Vazquez-Castro, D. Belay-Zeleke, A. Sanchez-Esguevillas, “Single code multiple access for the broadband satellite return channel”, in *Proc. IEEE Globecom 2002*, Taipei, Taiwan, Nov. 2002, pp. 2902–2907.

[5.8] J. Martin, Communications Satellite Systems, Prentice-Hall, Englewood Cliffs, NJ, 1978.

[5.9] E. Del Re, R. Fantacci, G. Giambene, S. Walter, "Performance analysis of an improved PRMA protocol for low earth orbit mobile satellite systems",. *IEEE Trans. Vehic. Technol.,* vol. 48, no. 3, pp. 985-1001, May 1999.

[5.10] S. Nanda, D. J. Goodman, U. Timor, "Performance of PRMA: A packet voice protocol for cellular systems", *IEEE Trans. Vehic. Technol.*, vol. 40, pp. 584-598, Aug. 1991.

[5.11] S. Nanda, "Stability evaluation and design of the PRMA joint voice data system”, *IEEE Trans. Commun.*, vol. 42, no. 3, pp. 2092-2104, May 1994.

[5.12] G. Benelli, R. Fantacci, G. Giambene, C. Ortolani, “Performance analysis of a PRMA protocol suitable for voice and data transmissions in low Earth orbit mobile satellite systems”, *IEEE Trans. Wireless Commun*., vol. 1, no. 1, pp. 156-168, Jan. 2002.

[5.13] G. Giambene, E. Zoli, “Stability analysis of an adaptive packet access scheme for mobile communication systems with high propagation delays”, *Internat. J. Satell. Commun. Network*., vol. 21, pp. 199-225, March 2003.

[5.14] N. Batsios, I. Tsetsinas, F. N. Pavlidou, “Performance evaluation of CDMA/PRMA techniques for LEO constellations”, in *Proc. IEEE Vehic. Technol. Conf. 2001*, Rhodes, Greece, May 2001, pp. 576-580.

[5.15] L. G. Roberts, “Dynamic allocation of satellite capacity through packet reservation”, in *Proc. Nat. Computer Conf., AFIPS NCC73 42*, pp. 711-716, 1973.

[5.16] A. Andreadis, R. Angioloni, R. Fantacci, G. Giambene, M. Michelini, G. Vannuccini, “Proposal of a packet access scheme to integrate isochronous and data bursty traffic in low Earth orbit-mobile satellite systems”, in *Proc. IEEE Globecom 2001*, S. Antonio, TX, Nov. 2001, pp. 2774 – 2778.

[5.17] R. Guerin, V. Peris, “Quality-of-Service in packet networks: basic mechanisms and directions”, *Computer Networks*, vol. 31, pp. 169-189, 1999.

[5.18] N. Celandroni, E. Ferro, "The FODA-TDMA satellite access scheme: presentation, study of the system and results", *IEEE Trans. Commun.*, vol. 39, no. 12, pp. 1823-1831, Dec. 1991.

[5.19] N. Celandroni, E. Ferro, N. James, F. Potortì, "FODA/IBEA: a flexible fade countermeasure system in user oriented networks", *Internat. J. Satell. Commun.*, vol. 10, no. 6, pp. 309-323, Nov.-Dec. 1992.

[5.20] N. Celandroni, E. Ferro., F. Potortì, "Experimental results of a demand-assignment thin route TDMA system", *Internat. J. Satell. Commun.*, vol. 14, no. 2, pp. 113-126, March-April 1996.

[5.21] F. Alagoz, D. Walters, A. AlRustamani, B. Vojcic, R. Pickholtz, “Adaptive rate control and QoS provisioning in direct broadcast satellite networks”, *Wireless Networks*, vol. 7, no. 3, pp. 269-261, 2001.

[5.22] R. Gibbens, F. Kelly, P. Key, “A decision theoretic approach to call admission control in ATM networks”, *IEEE J. Select. Areas Commun.*, vol. 13, no. 6, pp. 1101-1114, Aug. 1995.

[5.23] E. W. Knightly, N. B. Shroff, “Admission control for statistical QoS: theory and practice”, *IEEE Network Mag.*, vol. 13, no. 2, pp. 20-29, March/April 1999.

[5.24] M. Naghshineh, M. Schwartz, “Distributed call admission control in mobile/wireless networks”, *IEEE J. Select. Areas Commun.*, vol. 14 , no. 4 , pp. 711-717, May 1996.

[5.25] P. Todorova, S. Olariu, H. N. Nguyen, “A two-cell-lookahead call admission and handoff management scheme for multimedia LEO satellite networks”, in *Proc. 36th Hawaii Internat. Conf. on Syst. Sci. (HICSS-36)*, Big Island, Hawaii, Jan. 2003.

[5.26] L. Chisci, R. Fantacci, T. Pecorella, “Strategies for distributed bandwidth control in communication networks with high bandwidth delay product”, in *Proc. 43rd IEEE Conf. on Decision and Contr.*, Atlantis, Paradise Island, Bahamas, Dec. 2004.

[5.27] M. El-Kadi, S. Olariu, P. Todorova, “Predictive resource allocation in multimedia satellite networks”, in *Proc. IEEE Globecom 2001*, San Antonio, TX, Nov. 2001.

[5.28] P. Todorova, S. Olariu, H. N. Nguyen, “A selective look-ahead bandwidth allocation scheme for reliable handoff in multimedia LEO satellite networks”, in *Proc. ECUMN'2002*, Colmar, France, April 2002.

[5.29] K. W. Ross, Multiservice Loss Models for Broadband Telecommunication Networks, Springer-Verlag, London, 1995.

[5.30] N. Celandroni, F. Davoli, E. Ferro, “Static and dynamic resource allocation in a multiservice satellite network with fading”, *Internat. J. Satell. Commun. Network.*, vol. 21, no. 4-5, pp. 469-487, July-Oct. 2003.

[5.31] B. Tsybakov, N.D. Georganas, “On self-similar traffic in ATM queues: definition, overflow probability bound, and cell delay distribution”, *IEEE/ACM Trans. Networking*, vol. 5, no. 3, pp. 397-409, 1997.

[5.32] B. Tsybakov, N.D. Georganas, “Self-similar traffic and upper bounds to buffer-overflow probability in an ATM queue”, *Performance Evaluation*, vol. 32, pp. 57-80, 1998.

[5.33] I. Norros, "On the use of fractional Brownian motion in the theory of connectionless networks”, *IEEE J. Select. Areas Commun.*, vol. 13, no. 6, 1995.

[5.34] H.S. Kim, N.B. Shroff, “Loss probability calculations and asymptotic analysis for finite buffer multiplexers”, *IEEE Trans. Networking*, vol. 9, no. 6, pp. 755-768, 2001.

[5.35] C. G. Cassandras, G. Sun, C. G. Panayiotou, Y. Wardi, “Perturbation analysis and control of two-class stochastic fluid models for communication networks,” *IEEE Trans. Automat. Contr.*, vol. 48, no. 5, pp. 770-782, May 2003.

[5.36] F. Davoli, M. Marchese, M. Mongelli, "Optimal resource allocation in satellite networks: certainty equivalent approach versus sensitivity estimation algorithms", *Internat. J. Commun. Syst.* (to appear).

[5.37] R. Bolla, F. Davoli, M. Marchese, "A bandwidth allocation strategy for multimedia traffic in a satellite network", in *Proc. IEEE Globecom 2000*, San Francisco, CA, Nov. 2000, pp. 1130-1134.

[5.38] R. Bolla, F. Davoli, M. Marchese, "Adaptive bandwidth allocation methods in the satellite environment", in *Proc. IEEE Internat. Conf. Commun. (ICC 2001)*, Helsinki, Finland, June 2001, pp. 3183-3190.

[5.39] R. Bolla, N. Celandroni, F. Davoli, E. Ferro, M. Marchese, "Bandwidth allocation in a multiservice satellite network based on long-term weather forecast scenarios", *Computer Commun.*, vol. 25, pp. 1037-1046, July 2002.

[5.40] Federal Emergency Management Agency, "Technology applications by the federal emergency management agency in response, recovery, and mitigation operations," presented at the *27th Joint Meeting of the U.S./Japanese Panel on Wind and Seismic Effects*, Tokyo/Osaka, Japan, May 16–27, 1995.

[5.41] G. Philip and R. Hodge, "Disaster area architecture," in *Proc. IEEE MILCOM*, San Diego, CA, Nov. 1995, pp. 833–837.

[5.42] S. Adamson and S. Gordon, "Analysis of two trunk congestion relief schemes," in *Proc. IEEE MILCOM*, Boston, MA, Oct. 1993, pp. 902–906.

[5.43] P. B. Key, "Optimal control and trunk reservation in loss networks," *Probabil. Eng. Inform. Sci*., vol. 4, pp. 203–242, 1990.

[5.44] G. Choudhury, K. Leung, W. Whitt, "An Algorithm to compute blocking probabilities in multi-rate multi-class multi-resource loss models," *Adv. Appl. Probabil.*, vol. 27, pp. 1104–1143, 1995.

[5.45] C. C. Beard, V. S. Frost, "Prioritized resource allocation for stressed networks", *IEEE/ACM Trans. Networking*, vol. 9, no. 5, pp. 618-633, 2001.

[5.46] S. B. Biswas, B. Sengupta, "Call admissibility for multirate traffic in wireless ATM networks", in *Proc. IEEE INFOCOM*, vol. 2, Kobe, Japan, pp. 649-657, 1997.

[5.47] D. Wood, "The DVB Project: philosophy and core system", *Electronics and Commun. Eng. J.*, pp. 5-10, Feb. 1997.

[5.48] "Digital broadcasting systems for television, sound and data services; Framing structure, channel coding and modulation for 11/12 GHz satellite services", *ETS 300 421*, Dec. 1994.

[5.49] "Digital Video Broadcasting (DVB); Interaction channel for satellite distribution systems", *ETSI EN 301 790*, 2003

[5.50] J. Neale, R. Green, J. Landovskis, "Interactive Channel for Multimedia Satellite Networks", *IEEE Commun. Mag.*, pp. 192-198, March, 2001,.

[5.51] www.gilat.com

[5.52] U. Reimers, "DVB-T: the COFDM based system for terrestrial television", *Electronics and Commun. Eng. J.*, pp. 28-32, Feb. 1997.

[5.53] A. Mason, "Digital Video Broadcasting Standards for Satellite Terrestrial and Cable Television", in *Proc. IEEE MTT Internat. Microwave Symp.*, Baltimore, June 1998.

[5.54] "Digital Video Broadcasting (DVB); Interaction channel for Digital Terrestrial Television (RCT) incorporating Multiple Access OFDM", *ETSI EN 301 958, 2002.*

[5.55] A. Untersee, G. Connan, "DVB-Return Channel-Terrestrial: An Update", Harris Broadcast Europe, www.broadcastpapaers.com.

[5.56] "Digital Video Broadcasting (DVB); Framing structure, channel coding and modulation systems for Digital Satellite News Gathering (DSNG) and other contribution application by satellite ", *ETSI EN* 301 210.

[5.57] "Digital Video Broadcasting (DVB); Second generation framing structure, channel coding and modulation systems for Broadcasting, Interactive services, New Gathering and other broadband satellite applications ", draft *ETSI EN 302 307*, V1.1.1 (2004-06).

****

[6.1] G. Fairhurst, L. Wood, "Advice to link designers on link Automatic Repeat reQuest (ARQ)", BCP 62, IETF RFC 3366.

[6.2] J. Postel, “Internet Protocol”, STD 5, IEFT RFC 791, Sept. 1981.

[6.3] M. Mathis, J. Mahdavi, S. Floyd, A. Romanow, “TCP selective acknowledgement options”, IETF RFC 2018, Oct. 1996.

[6.4] S. Dawkins, G. Montenegro, M. Kojo, V. Magret, N. Vaidya, “End-to-end performance implications of links with errors”, BCP 50, IETF RFC 3155, Aug. 2001.

[6.5] E. Papapetrou, S. Karapantazis, G. Dimitriadis, F.-N. Pavlidou, “Satellite handover techniques for LEO networks”, *Internat. J. Satell. Commun. Network.*, vol. 22, no. 2, pp. 231-245, 2004.

[6.6] S. Olariu, S. Rashid, A. Rizvi, R. Shirhatti, P. Todorova, “Q-Win - A new admission and handoff management scheme for multimedia LEO satellite networks”, *Telecommun. Syst.*, vol. 22, no. 1-4, pp.151-168, 2003.

[6.7] P. Todorova, S. Olariu, H. N. Nguyen, “A leightweight call admission and handoff management scheme for LEO satellite networks”, in *Proc. 5th Europ. Workshop on Mobile/Personal Satcoms (EMPS 2002)*, Baveno, Italy, Sept. 2002.

[6.8] E. Del Re, R. Fantacci, G. Giambene, “Handover queuing strategies with dynamic and fixed channel allocation techniques in low orbit model satellite systems”, *IEEE Trans. Commun.*, vol. 47, no. 1, pp. 89-101, Jan. 1999.

[6.9] E. Del Re, R. Fantacci, G. Giambene, “Effective dynamic channel allocation techniques with handover queueing for mobile satellite networks”, *IEEE J. Select. Areas Commun.*, vol. 13, no. 2, pp. 397-405, Feb.1995.

[6.10] G. Maral, J. Restrepo, E. Del Re, R. Fantacci, G. Giambene, “Performance Analysis for a Guaranteed Handover Service in a LEO Constellation with a Satellite-Fixed Cell System”, *IEEE Trans. Vehic. Technol.*, vol. 47, no. 4, pp. 1200-1214, Nov. 1998.

[6.11] E. Papapetrou, F.-N. Pavlidou, “QoS handover management in LEO/MEO satellite systems”, *Wireless Personal Commun.*, vol. 24, no. 2, pp. 189-204, Jan. 2003.

[6.12] E. Del Re, R. Fantacci, G. Giambene, “Different queuing policies for handover requests in Low Earth Orbit mobile satellite systems”, *IEEE Trans. Vehic. Technol.*, vol. 48, no. 2, pp. 448-458, March 1999.

[6.13] N. Celandroni, S. T. Rizzo, “Detection of errors recovered by decoders for signal quality estimation on rain-faded AWGN satellite channels”, *IEEE Trans. Commun.*, vol. 46, no. 4, pp. 446-449, April 1998.

[6.14] N. Celandroni, E. Ferro, F. Potortì, “The performance of the FODA/IBEA satellite access scheme measured on the Italsat satellite”, in *Proc. ICDSC-10 Conf.,* Brighton, UK, May 1995, vol. 1, pp. 332-338.

[6.15] M. J. Willis, B. G. Evans, ”Fade countermeasure at Ka band for Olympus“, *Internat. J. Satell. Commun.*, vol. 6, pp.301-311, 1988.

[6.16] H. Kazama, T. Atsugi, M. Umehira, S. Kato, “A feedback-loop type transmission power control for low speed TDMA satellite communication systems”, in *Proc. IEEE Internat. Conf. Commun. (ICC ’89)*, Boston, MA, June 1989.

[6.17] E. Russo, “Implementation of a space diversity system for Ka-band satellite communications”, in *Proc. IEEE Internat. Conf. Commun. (ICC 93),* Geneva, Switzerland, May1993.

[6.18] L. Dossi, G. Tartara, E. Matricciani, “Frequency diversity in millimeter wave satellite communications”, *IEEE Trans. Aerospace and Electronic Syst.*, vol. 28, no. 2, pp. 567-73, April 1992.

[6.19] F. Carassa, E. Matricciani, G. Tartara, “Frequency diversity and its applications”, *Internat. J. Satell. Commun.*, vol. 6, pp. 313-322, 1988.

[6.20] F. Carassa, “Adaptive methods to counteract rain attenuation effects in the 20/30 GHz band”, *Space Commun. and Broadcasting*, vol. 2, pp. 253-269, 1984.

[6.21] B. Li, R. A. Di Fazio, A. Zeira, P. J. Pietraski "New results on SNR estimation of MPSK modulated signals", in *Proc. in 14th IEEE Internat. Symp. on Personal, Indoor, and Mobile Radio Commun. (PIMRC 2003)*, Beijing, China, Sept. 2003, pp. 2373-2377.

[6.22] N. Celandroni, E. Ferro, F. Potortì "Quality estimation of PSK modulated signals", *IEEE Commun. Mag.*, vol. 35, no. 7, pp. 50-55, July 1997.

[6.23] N. Celandroni, F. Potortì, "Fade countermeasure using signal degradation estimation for demand-assignment satellite systems". *J. Commun. and Networks*, vol. 2, no. 3, pp. 230-238, Sept. 2000.

****

ADDITIONAL READINGS

UMTS papers on the International Journal of Satellite Communications and Networking at site:
http://www3.interscience.wiley.com/search/allsearch?mode=quicksearch&WISindexid1=WISall&WISsearch1=S-UMTS

N. Celandroni, F. Davoli, E. Ferro, S. Vignola, S. Zappatore, A. Zinicola, “An experimental study on the Quality of Service of video encoded sequences over an emulated rain-faded satellite channel”, *IEEE J. Select. Areas Commun.,* vol. 22, no. 2, pp. 229-237, Feb. 2004.

N. Celandroni, E. Ferro, “A multi-frequency TDMA/TDM system for a VSAT terminal network operating in Ka band”, *J. Commun. and Networks,* vol. 3, no. 2, pp. 132-141, June 2001.

N. Abramson “The throughput of packet broadcasting channels”, *IEEE Trans. Commun.,* vol. COM-25, pp. 117-128, 1977.

A. Guntsch, “Analysis of the ATDMA/PRMA++ protocol in a mobile satellite environment”, in *Proc. 46th IEEE Vehic. Technol. Conf.,* Atlanta, GA, April 1996, pp. 1225-1229.

V. Kawadia, P. R. Kumar, “A cautionary perspective on cross layer design", *IEEE Wireless Commun. Mag.*, 2004 (to appear). http://black.csl.uiuc.edu/~prkumar/ps_files/cross-layer-design.pdf.

W. Kumwilaisak, Y. T. Hou, Q. Zhang, W. Zhu, C.-C. Jay Kuo, Y.-Q. Zhang, “A cross-layer quality-of-service mapping architecture for video delivery in wireless networks”, *IEEE J. Select. Areas Commun.*, vol. 21, no. 10, 2003.

V. T. Raisinghani, A. K. Singh, S. Iyer, “Improving TCP performance over mobile wireless environments using cross-layer”, in *Proc. IEEE Internat. Conf. on Personal Wireless Commun.*, New Delhi, India, Dec. 2002.

ITU-T Recommendation F.700: “Framework recommendation for audio-visual/multimedia services”.

ETSI TS 101 851-1 v1.1.1, “Satellite Component of UMTS/IMT 2000; A-family; Part 1: Physical channels and mapping of transport channels into physical channels”, Dec. 2000.

ETSI TS 101 851-2 v1.1.1, “Satellite Component of UMTS/IMT 2000; A-family; Part 2: Multiplexing and channel coding”, Dec. 2000.

ETSI TS 101 851-3 v1.1.1, “Satellite Component of UMTS/IMT 2000; A-family; Part 3: Spreading and modulation”, Dec. 2000.

ETSI TS 101 851-4 v1.1.1, “Satellite Component of UMTS/IMT 2000; A-family; Part 4: Physical layer procedures”, Dec. 2000.

K. Feher, Digital Communications. Satellite/Earth Station Engineering. *Prentice Hall*, 1981.

3GPP RAN WG#1, Technical Specification Group (TSG), 25.211-25.214 V4.1.0.

R. Fantacci, T. Pecorella, I. Habib, “Proposal and performance evaluation of an efficient multiple-access protocol for LEO satellite packet networks”, *IEEE J. Select. Areas Commun.,* vol. 22, no. 3, pp. 538-545, April 2004.

G. Giambene, F. Miano, E. Zoli, “Energy-efficient packet access scheme for MF-TDMA in non-GEO satellite systems”, in *Proc. VTC 2004-S*, Milan, May 2004.

3GPP TS 23.246, Multimedia broadcast/multicast service; architecture and functional description, release 6.

K. Narenthiran et al., “S-UMTS access network for MBMS service delivery: the SATIN approach”, *Internat. J. Satell. Commun. Network.*, Jan.-Feb. 2004.

T. Severijns et al., “The intermediate module concept within the SATIN proposal for the S-UMTS air interface”, *Proc. IST Mobile Summit 2002*, Thessaloniki, Greece, June 2002.

Y. Bernet et al, “Integrated Services operation over Diffserv networks”, *IETF draft,* June 1999.

A. Iera, A. Molinaro, “Designing the interworking of terrestrial and satellite IP networks”, *IEEE Commun. Mag.,* vol. 40, no. 2, pp. 136-144, Feb. 2002.

A. Iera, A. Molinaro, S. Marano, “IP with QoS guarantees via GEO satellite channels: performance issues”, *IEEE Personal Commun. Mag.*, vol. 8, no. 3, pp. 14-19, June 2001.

L. S. Ronga, T. Pecorella, E. Del Re, R. Fantacci, “A gateway architecture for IP satellite networks with dynamic resource management and DiffServ QoS provision”, *Internat. J. Satell. Commun. Network.*, vol. 21, no. 4-5, pp. 351-366, July-Oct. 2003.

T. Inzerilli, S. Montozzi, “Design of an efficient CAC for a broadband DVB-S/DVB-RCS satellite access network”, in *Proc. 1st Internat. Conf. on Adv. Satell. Mobile Syst., ASMS 2003*, Frascati, Italy, July 2003.

S. Cho, I. F. Akyildiz, M. D. Bender, H. Uzunalioglu, “A new connection admission control for spotbeam handover in LEO satellite networks*, Wireless Networks*, vol. 8, no. 4, July 2002.

E. Del Re, R. Fantacci, G. Giambene, “Efficient dynamic channel allocation techniques with handover queueing for mobile satellite networks”, *IEEE J. Select. Areas Commun.,* pp. 397-405, Feb. 1995.

S. Olariu, P. Todorova, “Resource management in LEO satellite networks”, *IEEE Potentials*, pp. 6-13, April/May 2003.

Ruiz, T. L. Doumi, J. G. Gardiner, “Teletraffic analysis and simulation of mobile satellite systems”, in *Proc. IEEE Internat. Conf. Commun. (ICC’99)*, , Vancouver, Canada, June 1999, vol. 2, pp. 1074-1078.

M. Allman, S. Dawkins, D. Glover, J. Griner, D. Tran, T. Henderson, J. Heidemann, J. Touch, H. Kruse, S. Ostermann, K. Scott, J. Semke, “Ongoing TCP research related to satellites”, *IETF RFC 2760*, Feb. 2000.

L. Chisci, R. Fantacci, T. Pecorella, “Predictive bandwidth control for GEO satellite networks”, in *Proc. IEEE Internat. Conf. Commun. (ICC 2004*), Paris, France, June 2004, pp 3958-3962.

S. Cho, “Adaptive dynamic channel allocation scheme for spot-beam handover in LEO satellite networks”, in *Proc. IEEE Vehic. Technol. Conf. 2000 (Fall VTC 2000)*, Boston, MA, Sept. 2000, pp. 1925-1929.

T. V. Lakshman, U. Madhow, “The performance of TCP/IP for networks with high bandwidth-delay products and random loss”, *IEEE/ACM Trans. Networking*, vol. 5, no. 3, pp. 336-350, June 1997.

M. Mathis, J. Semke, J. Mahdavi, T. Ott, “The macroscopic behavior of the TCP congestion avoidance algorithm”, *Comput. Commun. Rev*., vol. 27, no. 3, July 1997.

V. Jacobson, R. Braden and D. Borman, “TCP extensions for high performance”, IETF, RFC 1323, May 1992.

J. Padhye, V. Firoiu, D. F. Towsley, J. F. Kurose, "Modeling TCP Reno performance: a simple model and its empirical validation", *IEEE/ACM Trans. Networking*, vol. 8, no. 2, April 2000.

Mertzanis, R. Tafazolli, B. G. Evans, “Connection admission control strategy and routing considerations in multimedia (Non-GEO) satellite networks”, in *Proc. IEEE IEEE Vehic. Technol. Conf. 1997 (VTC 1997)*, Phoenix, AZ, May 1997, pp. 431-435.

C. Partridge, T. J. Shepard, “TCP/IP performance over satellite links”, *IEEE Network*, vol. 11, no. 5, pp. 44-49, Sep./Oct. 1997.

M. Marchese, “Performance analysis of the TCP behavior in a GEO satellite environment”, *Computer Commun.*, vol. 24, Issue 9, pp. 877-888, May 2001.

M. Allman, D. Glover and L. Sanchez, “Enhancing TCP over satellite channels using standard mechanisms”, RFC 2488, Jan. 1999.