Challenges, Opportunities, and Solutions for Converged Satellite & Terrestrial Networks

Tarik Taleb[1] Yassine Hadjadj-Aoul[2] Toufik Ahmed[3]

[1] Tohoku University. Sendai, Japan talebtarik@ieee.org

[2] University of Rennes-1, Rennes, France hadjadj@ieee.org

[3] University of Bordeaux-1, Talence, France tad@labri.fr

***Abstract* – The current trend in telecommunications services provisioning is shifting towards global ubiquitous networking and unified service architecture. Given the diversity of access technologies, this global ubiquitous networking cannot be possible without an efficient interworking between the different access players. This leads to the necessity of defining, implementing and deploying common services control architecture, able to support a wide variety of services for users in a variety of roles (consumer, producer or manager of communication and media).**

This paper defines some issues related to the interworking operation between the satellite and terrestrial domains. It suggests some solutions and discusses their potential.

I. Introduction

Wireless communications continues to pervade all aspects of our lives – wireless distribution of audio and video around the home; wireless solutions for logistics; wireless ticketing and access control; wireless sensors for agriculture, medical applications, etc. While many people appreciate the profound impact that wireless communications are having and will have on our lives, *it will be some time before the vision of wireless everywhere will be realized* – mainly because introducing large-scale changes to the way many systems work is complex and requires significant time, effort and energy.

While there have been many important advances in wireless technology in recent years, there are economic challenges in providing high-speed wireless access to less populated areas. This gap between those who benefit from digital technology and those who do not is known as *digital divide*.

A key technology which can help to *bridge the digital divide* is satellite communications as it can be used in areas where there is no terrestrial alternative. In the developed world, satellite networks can be "interworked" with existing terrestrial networks, being they wireless or fixed systems, core or access network, and function as a high-speed backbone network to support a wide variety of services for users in a variety of roles.

In this regard, there are many lessons to learn from the recent mobile satellite experience. Indeed, in urban/suburban areas, fixed and mobile technologies (e.g., ADSL, GSM) are well advanced. Satellites, performing in isolation, cannot compete with terrestrial systems in these urban areas. They can only provide niche services to areas inaccessible to terrestrial technologies. While these markets are politically important, they are small and of poor revenue for satellite operators. The future of next-generation satellite systems is clearly in an integrated architecture with terrestrial systems. Their success also hinges on their ability to provide, in full cooperation with terrestrial systems, broadband data rates applications; similar in spirit to today's Internet. This is also beneficial for terrestrial system operators as it will enable them to increase the capacity of their systems, support large-scale deployment of different emerging bandwidth-intensive services, and satisfy the ever-growing community of Internet users.

Two critical issues arise when considering satellite systems in this context: firstly, satellite systems are *very costly* in general; and secondly, there are *challenges in integrating* satellite and terrestrial networks, particularly when terminal mobility is necessary. This paper will give some insights towards solving both of these problems.

In this paper, we focus on interworking between the satellite part of the network and its terrestrial counterpart. Interworking related operations are performed at newly defined entities called Interworking Gateways (IGWs). The scope of this paper is to define the modules of the technological solutions that will be incorporated in IGWs and to evaluate their performances via computer simulations.

The remainder of this paper is organized as follows. Section II portrays the key components of the envisioned architecture. Section III describes our proposed “context-aware complete end-to-end QoS” approach devised for interworking between the satellite and the terrestrial domains. The paper concludes in Section IV.

II. Envisioned Network Topology

Although GEO systems are widely in use, and LEO/MEO will come onto the scene in the longer term, this paper does not target any particular satellite constellation type. The developed solutions will be designed in a way that they can be applicable to all constellation types (GEO, MEO, and LEO). The satellites are only assumed to be bidirectional interactive, acquiring onboard processing (OBP) capabilities and inter-satellite links (ISLs) [1]. Terminals are interactive. Terminals outside the reach of the terrestrial network have direct access to the satellites. Terminals within the reach of the ground Internet infrastructure have the ability to either connect directly to satellites or via the Interworking Gateways (IGWs). The overall objective of this paper is to define the necessary intelligence that should be added to IGWs to guarantee a “context-aware complete end-to-end QoS” for users. Thus, different levels of convergence will be considered. The first level of convergence concerns the efficient data transmission based on IP. The second level refers to the control and signaling for providing resource allocation and management. The third level of convergence deals with the provisioning of a generic service delivery platforms based in IP Multimedia Subsystem (IMS). Finally, mobility management and seamless connectivity are considered for both network-link handover and link-layer handover.

III. Context-Aware Complete End-to-End QoS Approach

A. *Efficient Data Transmission*

Firstly, in light of the rapid globalization of the Internet and the resultant universality of the Internet Protocol (IP), the data traffic load to be generated from the interworked satellite/terrestrial networks is expected to be all-IP as well. Investigating the interactions of IP protocols with the network is of vital importance.

The satellite systems are well known for their unique characteristics – long propagation delays, large delay-bandwidth product, errors due to propagation corruption and handovers, and variable Round Trip Time (RTT) and link handovers. These features put limitation on the working of most transmission With this regard, the authors have recently developed the Recursive, Explicit, and Fair Window Adjustment (REFWA) method to enhance the efficiency and fairness of TCP in satellite systems [2]. The use of the REFWA scheme has been extended further to the case of hybrid wired/wireless networks as well. While the REFWA scheme exhibits good performance, its performance remains limited in large bandwidth environments (such as satellite systems) due to its window-based nature. Development of a “*new rate-based congestion control protocol”* that is specifically tailored for satellites and can exploit well the “large delay bandwidth product” feature of the satellite systems is required. REFWA can be a good candidate for that by changing its window-based feature to a rate-based one. Indeed, this is possible by having IGWs send data at rates exactly equal to the feedback value computed by REFWA. This is similar in spirit to the concept of the eXplicit Control Protocol (XCP). Using such rate-based congestion control mechanism, and similar in spirit to the connection splitting in [3], the full end-to-end path will be decoupled into separate segments – end-terminal to IGW, segment traversing the satellite network, and the final segment between the IGW and the remote end-terminal (1). The use of the protocol pertains to the segment traversing the satellites. The other two segments will employ control mechanisms which are optimized for their characteristics (wireless or wired). Further, the necessary intelligence required for the coordination between the used data transmission mechanisms will be added to IGWs to ensure a reliable delivery of data while meeting the end-to-end QoS requirements.

To illustrate the idea with more clarity, we have conducted simulations using NS-2. The window-based nature of REFWA is replaced by a rate-based one as explained earlier. The performance of the modified REFWA is compared against that of XCP and TCP as shown in Fig. 2. We tested the system under homogeneous traffic conditions using 100 heavy FTP sources during 300 seconds, a duration long enough to capture and study the behavior of our proposed transport protocol. At the beginning of the simulation, the hosts behind each satellite terminal are activated randomly following a uniform distribution ranging from 10 ms to 100 ms.

Fig. 2 shows the average window size for the three protocols. From the figure, it is clear that the average window size, when REFWA is in use, converges immediately to the optimal window value. In contrast, in case of XCP, it takes 15 seconds before the system reaches its optimal window. In case of TCP/Reno, the system oscillates arround the optimal value without reaching a steady state. The simulation results also demonstrate the good performance of the modified REFWA as it achieves the highest goodput. Indeed, REFWA outperforms XCP and TCP/Reno by 6.93% and 20.90%, respectively, in terms of goodput.

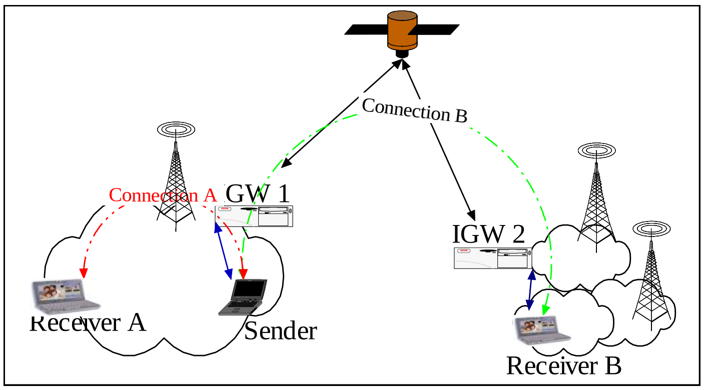

Fig. 1 RTT-based Connection Setup + Connection Decoupling

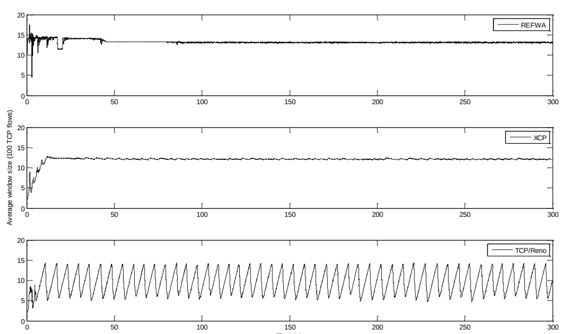

Fig. 2 The average window size for REFWA, TCP and XCP.

For the sake of further transmission efficiency and better QoS, short RTT connections should be established via the terrestrial wireless network. Indeed, terminals communicating with nearby users do not have to drain up their energy to connect directly to satellites. Long RTT connections can be set via satellites. Different techniques can be used for periodic monitoring of the network conditions. IGWs will be constantly updated with feedback on network dynamics. Based on this feedback and the RTT of connections, IGWs decide the path for communication – either via the satellites or via only the terrestrial network. For this purpose and in order to blur the separation between the satellite and terrestrial domains, there is need for exchange of state information (e.g., instant link loads) between the two domains. A hierarchical architecture of gateways can be considered. With this regard, the number of levels of this hierarchy, the size of each level, the amount of control traffic that should be exchanged and the length of the monitoring interval time should be decided in a way that enhances the accuracy in the assessment of network dynamics while minimizing the overhead in terms of signaling messages. Such *a context-aware routing scheme* will yield a better load balancing over the entire network and will enhance the E2E QoS [4].

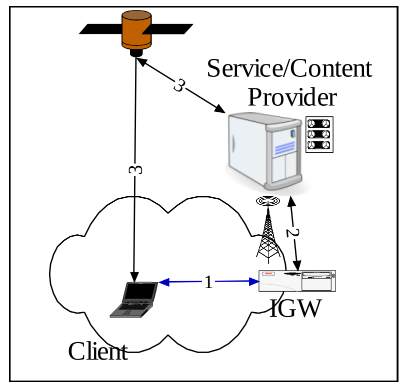

Fig. 3 Service Delivery Pattern.

In the considered interworked satellite/terrestrial network, IGW also provides the interface for any service/content provider who desires to provide service/content over both the terrestrial and the satellite networks. In case the provided data is bursty in nature (e.g., video data) and the targeted population of users is potential, it will be highly useful to send data from the provider to users using the satellite channels. In such communication scenario (**Error! Reference source not found.**), a service subscriber issues a request for a particular data/video title to IGW via the terrestrial network (wireless or wired). The IGW informs the service provider of the request and the latter allocates the necessary resources to satisfy the user's request. To ensure reliable transmission of data, (depending on the underlying transport protocol) the client keeps acknowledging IGW of successful receptions of data via acknowledgment packets sent over the terrestrial network. In case the ACK packets are delayed or lost, the overall network performance may be impacted. Adding intelligence in terms of "*an adequate delayed acknowledgment mechanism along with a robust error recovery mechanism*" to IGWs can help to cope with these issues.

B. Resource Allocation and Management

In light of the limited resources of any powerful network, QoS can be maintained only via efficient resource management/allocation mechanisms. In the case of converged satellite/terrestrial networks, devising an efficient resource allocation method is a highly challenging task due to the fluctuating nature of the wireless links. Indeed, in DVB-S2 networks for instance, a novel satellite-tailored Adaptive Coding and Modulation (ACM) technique is introduced to cope with the wireless channel fluctuations. ACM renders the resource reservation process even more difficult since the channel capacity changes frequently as the channel experiences noisy periods.

While there are many approaches to solving resource management issues in networks, one which is suitable when there are limited, costly resources is that of connection admission control (CAC).

A large library of CAC schemes has been proposed in the literature. These techniques can be classified as either: resource reservation-based or statistical multiplexing-based schemes. Resource reservation CAC systems have some known scalability issues and may often lead to self-induced congestion due to the heavy resource reservation process. Besides, static reservation falls short to satisfy flexibility requirements of typical network operators. Furthermore, statistical multiplexing CAC approaches cannot completely eliminate the congestion during some peak noisy periods. However, they enable resources sharing between users and yield a reduced waste of resources.

This feature renders statistical multiplexing schemes more suitable for converged satellite/terrestrial networks. Several research works have devised different CAC schemes to guarantee a reasonable QoS level in different network conditions [5]. A common shortcoming of these schemes resides in their inefficiency to deal with the varying nature of the physical layer capacity of the satellite network.

The authors' research work presented in [6] takes into account the satellite channel fluctuations and presents an interesting CAC mechanism that also ensures fairness among terminals competing for the capacity of the same satellite channel. A shortcoming of the proposed approach consists in its lack of a bandwidth allocation mechanism and multi-service support. Indeed, CAC should be exerted in conjunction with a bandwidth allocation mechanism especially in converged satellite/terrestrial networks where the link capacity may vary as a result of the ACM mechanism. In this case, a combined action among various layers, "*cross layer approach*", of the networks is likely to improve the performance of the overall system by protecting, for example, prioritized flows from packet drops during congestion events. In this area of research, there is a particular interest in the development of a "*cross-layer bandwidth allocation mechanism*" that can assist CAC and further enhance its functionality.

For this purpose, we suggest using a channel prediction mechanism based on the least mean square algorithm to tackle the excessive delay incurred by the feedback in satellite networks. Based on the proposed model, a self-configuring mechanism to cope with variable network conditions is derived. From this cross layer approach, both optimized bandwidth allocation and guaranteed per-class quality of service are expected.

To illustrate the idea, we have conducted simulations using Opnet. We test the system under homogeneous traffic conditions using FTP sources during 1500 seconds. In the first phase of the simulation, the system is maintained free from noise. At the beginning of the second phase of the simulation (i.e., t=245s), a source of noise is introduced, which impacts directly the SNR that decreases by approximately 2 dB. At the third phase of the simulation (i.e., t=790s), another source of noise is introduced. Finally, at the last phase which starts at t=1000s, the sources of noise are eliminated successively at t=1000s and t=1210s. This performance evaluation scenario allows us to see how quickly and accurately our scheme adapts to both degrading and improving conditions.

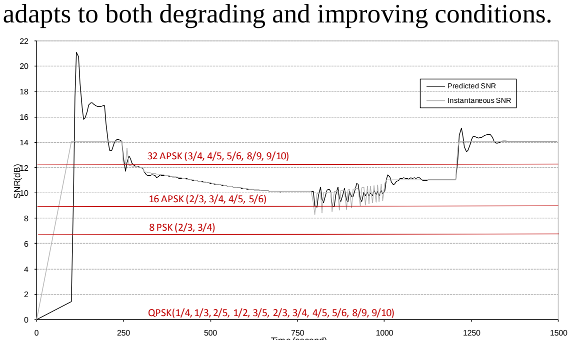

Fig. 4 The instantaneous SNR vs the predicted SNR.

Fig. 4 shows the instantaneous SNR experienced by a satellite terminal versus the predicted SNR. In the first stage of the simulation, we observe that the estimation error is relatively important. This is principally due to the random initialization used in Least Mean Square (LMS). After 105 seconds, we clearly see that the estimation becomes more precise which demonstrates the effectiveness of our proposed prediction mechanism. At this point, the accuracy of the proposed algorithm is approximately 98.5%.

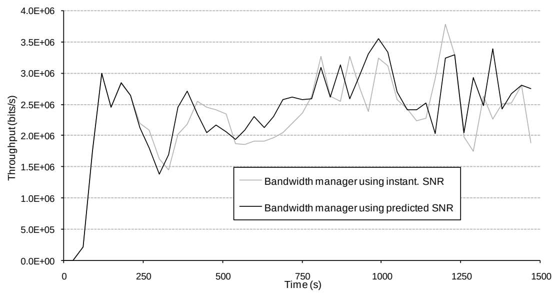

Fig. 5 Throughput using instantaneous vs predicted SNR.

Fig. 5 indicates the performance of the proposed mechanism in terms of throughput. The figures clearly shows that the bandwidth manager using predicted SNR values, which allow selecting the appropriate modulation and code rate, outperforms the conventional approach. As a consequence, the experienced throughput is increased by approximately 4.6%.

The proposed cross layer CAC mechanism, which relies on predicted SNR values, protects the network from congestion while maintaining a good trade-off between bandwidth utilization and end-to-end delay. These performances are particularly interesting for the provision of delay-sensitive and bandwidth-intensive applications over converged satellite/terrestrial networks.

C. Mobility Management

The success of any communication system hinges on its ability to provide acceptable QoS. In the context of mobile environments, QoS provisioning depends in turn on an efficient management strategy of mobility. In satellite networks, mobility management is a challenging task. Indeed, supporting continuous communication over satellite systems may require changing spot beams (and links in case of LEO/MEO systems) as well as the IP address of the communication endpoints. Thus, both link-layer and network-layer handovers are required for satellite networking. In case of non-geostationary (NGEO) systems, the mobility management becomes more complex as both the satellite network and the mobile users are on the move.

In satellites networks, handovers can be broadly classified into two categories: network-link handover and link-layer handover. The former occurs when one of the communication endpoints changes its IP address due to motion of satellites or mobility of the user terminal. The latter occurs when one or more links between the end terminals change. It consists of satellite handover, Inter-Satellite Link (ISL) handover, and spot beam handover. Spot beam handover are the most common type of handovers. They occur frequently due to the small area covered by spot beams and the mobility of users (or high speed of NGEO satellites). In this paper, we initially focus on spot beams handovers and then extend the study to the case of other handover types.

1) Spot beam Handover Management

For better frequency utilization, the footprint of an individual satellite is divided into smaller cells, called spot beams. To ensure uninterrupted ongoing communications, a current communication link should be handed off to the next spot beam when needed. A spot beam handover involves the release of the communication link between the user and the current spot beam and acquiring a new link from the next spot beam to continue the communication. Due to the small area covered by spot beams, users' mobility and high satellite speed in case of NGEO systems, spot beam handovers are the most common type of handovers experienced in satellite systems.

An efficient management of handovers is particularly linked to the resource allocation problem discussed above. Indeed, the selection of a suitable policy for channel allocation can ensure channel availability during handover. Thus, the channel allocation strategies and the handover guarantee are the prime issues in managing handover requests. It should be noted that it is more desirable to ensure and guarantee smooth ongoing calls rather than the blocking of a newly arriving call. To solve the spot beam handover problem, several handover schemes have been proposed in the recent literature. A thorough survey on these techniques is available in [7].

A sophisticated network planning is required to assign more capacity to spot beams when a high traffic rate is expected. Statistical methods, coupled with user behavior model and precise predictions of satellite tracks relative to the earth surface, allow general characterization of the traffic load for a particular satellite or spot beam. Via a cross layer design, this can help in anticipating imminent handover events and locating the new point of attachment to the satellite network [8]. While a cross layer optimization can be implemented either at end-devices or intermediate nodes in the network, such as IGWs, it is relatively easier and more feasible to implement changes at mobile nodes. Indeed, at the communication end-point, the physical layer of a mobile host instantly measures the radio strength or link quality. When the mobile node moves into the overlapping area of two or more spot beams, and different signals are consequently detected by the physical and data link layers, a warning message notifying an imminent handoff event, along with a list of the new possible spot beams, are sent to the application layer. In case of multiple spot beams, the application layer refers to a set of tools to sort out the spot beam to which the mobile node is most likely going to be connected to. Indeed the application layer may use history on the user's mobility pattern to predict the new spot beam. Referring to a spatial conceptual map, along with the user's personal information, its current position, and its velocity heading, the application layer can make an accurate prediction of the most probable future spot beam. Prior contextual knowledge on the coverage area of the satellite network and the type of the application can further increase the accuracy of the prediction. Once the next spot beam is determined, the mobile host informs the IGW of the next spot beam. Based on the current conditions (e.g., maximum number of free channels) of each spot beam, IGW decides whether the call should be accepted or denied. If the handover cannot be taken without degrading QoS of already existing users or causing network congestion, the IGW denies the handover request and sends an immediate negative acknowledgment to inform the mobile host that the request has been turned down. Simultaneously, a list of available spot beams can be sent along with the negative acknowledgment to induce the mobile host to hand over with another spot beam. The mechanisms by which IGWs admit or turn down handover requests (from a user) should be consistent with the underlying resource allocation strategy. An actual design and implementation of this *"context-aware cross-layer architecture"* at the mobile terminals and the intelligence required by IGWs to manage handovers define an interesting topic of research in this particular field of research.

*2) **Satellite and ISL Handover Management in NGEO systems***

The solution suggested above may be also highly interesting when it is put in the context of mobile satellite communication systems (e.g., LEO, MEO). In these systems, satellite handovers are more important as users need to first choose among different satellites and will then be served by the spot beam covering the user. In addition to the above solution, there is need to develop complementary solutions that select the most suitable satellite for communications that can reduce bandwidth wastage and the call blocking probability, and also fulfill the QoS requirements. In regard to QoS, the application type should be taken into account in the satellite handover management strategy.

Another important issue in the context of converged terrestrial/satellite networks is how to manage network layer handovers. While there are IP-based mobility solutions which have already reached the marketplace, they are unsuitable for converged satellite/terrestrial networks. This is mainly because they result in a very large amount of signaling traffic when employed in a satellite context, due to the constant and rapid motion of the satellites. Consequently, alternative network-layer handover mechanisms are required.

D. Seamless Connectivity

While in the above subsections, the focus was on defining issues pertaining to congestion control, resource allocation, and mobility management in the converged satellite/terrestrial network, and devising possible solutions, in the remainder of this paper, we discuss possible scenarios for the seamless use of the interworked satellite/terrestrial network and provide guidelines for their realization.

*1) **IMS-based Service Delivery Architecture***

A possible solution for service integration between satellite and terrestrial network can be based on the IP Multimedia Subsystem (IMS) that represents a key element in the satellite architecture, supporting seamless and universal access to personalised services. Indeed, the adoption of MS will favour a rapid emergence of new secure services and will enable seamless provisioning of multimedia services. IMS provides a Service Delivery Platform (SDP) on top of convergence network technologies. This will help in generating new revenues, reducing the complexity of the interworked gateway while significantly decreasing the cost of the satellite network management which impacts directly the satellite services cost. Additionally, IMS allows handling more efficiently the multicast and broadcast traffic initiated from the satellite terminals. The rest of this section describes how to handle multicast services using IMS-based Architecture.

Current mechanisms for delivering multicast services over satellite links use snooping (layer 2) or Proxying (layer 3) to allow the delivery of IGMP/MLD membership messages to satellite gateway over the air interface. A proxy and snooper are not changing anything in the IGMP/MLD messages but only forwarding the request further toward the gateway (IGW). In fact, over the satellite network, the broadcast property exists only on the forwarding link, i.e. the satellite return link provides only directional links. Host cannot listen to the signalling Report transmitted by other hosts on the return link. This leads them to individually send out their Report which generates an excessive traffic. This would result in flooding and high latency problems.

Flooding occurs when many hosts (i.e., IGMP/MLD clients) reply to a broadcast request from the IGMP/MLD Querier sent out by the router to sense the presence of clients in each multicast group. As highlighted earlier, unlike LAN, the satellite return link does not provide a broadcast property but only a unidirectional connection. Typically, hosts cannot listen directly to replies from other hosts. Thus, all the hosts have to response to the IGMP/MLD Querier after the expiration of their timer. Moreover, satellite multicast groups can be very large and very dynamic. This leads to a waste of bandwidth and CPU over the satellite link and the Gateway respectively.

Another important issue of IP multicast behavior over satellite networks is the latency in stopping transmission after the last host leaves a multicast group. The latency is the delay needed for the Querier to become aware that the multicast group is empty in order to stop multicast forwarding on it. The authors explain in [9] how to tackle the flooding and latency issues in providing multicast services. In this context, IMS service delivery platform is used which allows the IGMP messages to be aggregated and transmitted as SIP-based message for managing multicast groups. The argument for using SIP/IMS protocols is (1) to allow hosts to join and leave multicast group as IGMP does, (2) to verify the authentication and authorization of the user, (3) to signal any cryptographic context (example: using MIKEY), and (4) to support any future extension/augmentation easily that can be implemented in SIP.

IV. CONCLUDING REMARKS

This article highlights some of the opportunities behind integrated satellite and terrestrial networks. For the realization of such converged networks, it addresses a number of issues pertaining to transmission efficiency, resource allocation and management, mobility management, and seamless connectivity. Whilst the main objective of this article is to highlight the related issues and to define new directions for the community of satellite researchers, it also suggests a number of solutions, as seen from the networking perspectives, that, once they are put together, they form a complete context-aware end-to-end QoS approach that solves the aforementioned issues.

REFERENCES

[1]J. Farserotu and R. Prasad, "A Survey of Future Broadband Multimedia Satellite Systems, Issues and Trends," IEEE Communications Magazine, Vol. 38, No. 6, 2000.

[2] T. Taleb, N. Kato, and Y. Nemoto, "REFWA: An Efficient and Fair Congestion Control Scheme for LEO Satellite Networks", IEEE/ACM Trans. on Networking Journal, Vol. 14, No. 5, Oct. 2006. pp. 1031-1044.

[3]M. Marchese, M. Rossi, and G. Morabito, "PETRA: Performance Enhancing Transport Architecture for Satellite Communications," IEEE JSAC, Vol. 22, No. 2, Feb. 2004, pp. 320–22.

[4]T. Taleb et. al., "Explicit Load Balancing Technique for NGEO Satellite IP Networks with On-Board Processing Capability", to appear in IEEE Trans. on Networking journal.

[5] R. Abi Fadel and S. Tomhe, "Connection Admission Control and Comparison of two Differentiated Resources Allocations Schemes in a Low Earth Orbit Satellite Constellation" in ACM Wireless Networks J., Vol. 10, No. 10, May 2004.

[6] Y. H. Aoul and T. Taleb, "An Adaptive Fuzzy-based CAC Scheme for Uplink and Downlink Congestion Control in Converged IP and DVB-S2 Networks", to appear in IEEE Trans. Wireless Communications.

[7]P.K. Chowdhury, M. Atiquzzaman, and W. Ivancic, "Handover Schemes in Satellite Networks: State-of-the-Art and Future Research Directions," IEEE Communications Surveys and Tutorials, Vol. 8, No. 4, 4th Quarter 2006.

[8]T. Taleb et. al., "A Cross-Layer Approach for an Efficient Delivery of TCP/RTP-based Multimedia Applications in Heterogeneous Wireless Networks", in IEEE Trans. on Vehicular Technology Journal, Vol. 57, No. 6, Nov. 2008, pp. 3801-3814.

[9] T. Ahmed, et al. "IMS-based IP Multicast Service Delivery over Satellite Network" in 8th Annual IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC'07), September 2007.