A Framework for End-to-End Latency Measurements in a Satellite Network Environment

Anas A. Bisu, Alan Purvis, Katharine Brigham, and Hongjian Sun
Department of Engineering, Durham University, Durham, DH1 3LE, U.K.
Emails: {anas.a.bisu, alan.purvis, katharine.brigham, hongjian.sun}@durham.ac.uk

***Abstract*—In this work, a precise method for measuring end-to-end (E2E) latency in satellite Internet Protocol (IP) networks is proposed. Latency (i.e., time delay) is considered a key parameter that affects the quality of service (QoS) and the performance of communication systems. This is more pronounced in the IP over Satellite. Metrics such as throughput and bandwidth performance of communication systems are dependent on latency, which also has a direct impact on other QoS metrics, such as Internet packet transfer delay and delay variation (or jitter). The upper limits of QoS objective performance metrics are defined by E2E latency for different QoS traffic classes in this environment. Therefore, there is a need to develop efficient methods for the accurate measurement of E2E latency in a satellite IP environment. Two case study scenarios were developed for satellite heterogeneous networks to measure the latency in a satellite IP network. Two geostationary satellite network services were used to compare the performance of the different scenarios and networks. The results demonstrate that at least 50% of the E2E latency is due to the processing and transmitting of IP packets over the satellite in both scenarios. Inconsistent latency behaviour was also observed from daily results at different times of the day, which may degrade performance of jitter sensitive applications.**

***Index Terms*—Latency, QoS, Satellite Communications, Heterogeneous Network.**

I. Introduction

Satellite communication has interesting and unique attributes such as global coverage, resilience, scalability, multi/broadcast capability, bandwidth-on-demand flexibility, reliability, high data rate and high capacity [1]. These particular features may uniquely position satellite communications as a key technology to provide broadband Internet access to remote isolated communities areas for bridging the digital gap, and bring human and economic developments to these areas through smart city, smart grid, or smart home concepts. For example, about 62% of 1.2 billion population of Africa live in remote rural villages while 340 million of them travel about 50km to access the Internet [2, 3]. However, end-to-end (E2E) optimum performance and quality of service (QoS) at all network layers are required to meet current and future media-intense applications using satellite Internet protocol (satellite IP). Internet applications mainly use the transmission control protocol (TCP/IP) stack which was originally developed for terrestrial networks. Satellite communication links exhibit high latency and packets could be dropped, leading to requests for retransmissions, which congest the transmission path and thus cause high packet loss ratio [1, 4]. Latency (i.e., time delay) is the length of time taken by the transmitted signal (information) from the source or transmitter (Tx) to be received at the destination or receiver (Rx) end [5]. High latency has significant impact on performance and QoS of communications systems, especially when applications/services are TCP/IP-based [1, 6]. Optimum system performance can be achieved by improving latency, e.g., next generation 5G networks aim at achieving 1ms latency with 1000-fold capacity [7]–[9]. Performance metrics, such as throughput, IP packet transmission delay, and jitter (or packet delay variation), are directly affected by latency [1, 7, 10]. This is more pronounced in satellite IP networks, especially over geostationary earth orbit (GEO) satellites due to longer propagation paths with additional transmission and processing delays from end-to-end [11, 12]. Satellite networks exhibit long propagation latency and most assumptions of round-trip-time (RTT) of satellite are based on its propagation latency [11, 12]. Considering the E2E satellite network path, propagation latency contributes to about 50% of the total one-trip-time or RTT [13].

The increasing use of satellite IP in remote areas, and backhaul links, and the planned integration of satellite and terrestrial networks in future 5G networks [8, 9, 14] are a wakeup call for developing a more precise way to measure E2E latency in this environment. Today, applications that can convert smartphones and tablets into satellite phones and access terminals by different satellite network provider (SNP) are on the rise. An automated method was developed by Xiao *et al.* in [15] to measure and calibrate latency in real-time for global navigation satellite system (GNSS) transceivers. However, this work is limited to measuring GNSS transceiver latency as a point-to-point link and cannot be extended to the measurements for a complete multiple hop path. Furthermore, it is also limited to the use of a bidirectional cable link to couple and loop the signal from transmitter to receiver without leaving the transceiver system [15]. In contrast, the work proposed herein is a precise measurement technique for E2E latency in a heterogeneous network environment involving several links of different characteristics, including terrestrial, Wi-Fi, and satellite links for transmission and reception of voice over satellite IP.

The remaining sections are organized as follows. Section II details the related prior work. In Section III, the theoretical models used for the proposed framework are presented, and Section IV describes the experimental setup, scenarios, and the proposed E2E latency model. Section V subsequently discusses the experimental results, followed by concluding

978-1-5386-3180-5/18/$31.00 ©2018 IEEE

remarks given in Section VI.

II. Related Work on Latency Measurements

Latency is one of the key network performance parameters [16, 17] affecting many QoS metrics [1, 18]. Efforts have been made to measure and estimate latency and dependent parameters in different network environments [15]–[17, 19]–[23], especially under terrestrial network environments, with few conducted in pure satellite environments [13, 15, 24, 25]. The well-known methods used for network performance measurement are broadly classified as either *Active* or *Passive* approaches [16, 19]. The oldest and widely used IP network measurement tool is `ping` – an active approach in which Tx-user equipment (UE) generates *Internet Control Message Protocol* (ICMP) *echo request* to the Rx-UE. The Tx-UE will start a timer while Rx-UE retransmits the ICMP headers to Tx-UE as *echo reply* packet used to measure the RTT [16]. Another commonly used ICMP-based measurement technique is `traceroute` [16], which returns the RTT of *User Datagram Protocol* packet for each *hop* within the path from Tx-UE to Rx-UE with an increased value of *Time To Live* field in the IP header [16].

Different methods are used to measure either RTT, one-way delay (i.e., one-trip time), or latency. Well-known methods such as `ping` and `traceroute` are simple, active RTT measurement tools that are currently available but are not always accurate measures as some service providers/destination systems do not allow them for security reasons, and are therefore blocked by firewall at the Rx-end UE. Because these techniques measure the RTT by *elapsed time* of ICMP echo reply response from the target destination (Rx), a lack of an echo reply will not provide the needed E2E latency measurement or may give confusing results [16]. Another problem may be low scheduling priorities for ICMP-based traffic assign by routers within the measurement path, where the critical function is data packet routing, not ICMP packet test and measurement routing [16]. While `ping` and `traceroute` may provide useful data about network condition/status, these only measure the RTT *loop*, not simple E2E path conditions. Long RTT and high loss rates might be reported due to the target router processing load and scheduling algorithm priority [16]. Therefore, one-way delay methods are needed to measure accurately the E2E path. Useful research studies on one-way delay measurements can be found in [17, 20].

Our proposed framework uses an *active* method on real heterogeneous network (satellite and terrestrial) environment with two GEO satellite network providers to measure actual E2E latency as experienced by a real fixed satellite service end-user connected to a Satellite User Terminal (SUT) with UE. The results presented in this paper are real environment values (i.e., not ideal or simulated environment). Specialised end-UE in ideal simulated environment used in previous works minimised the impact(s) on user traffic, which may not reflect the true end-user experience in real communications environment operations. In addition, the proposed framework could be useful in Integrated Satellite Terrestrial Network of the future (5G and beyond) for measuring E2E network performance parameters such as throughput and available bandwidth.

III. Theoretical Framework

The proposed scenario models were developed from general latency theoretical framework models consisting of different component contributions along the E2E network path. The components of the network path contributing to the total latency are described below.

A. Propagation Latency

This is the time taken by the signal to travel from one node (Tx) to another (Rx) via the communication link [5, 20]. Expressed as a ratio of distance, d *(km)*, and speed, v *(km/s)*, of the medium (speed of light $c = 3 \times 10^8$m/s for wireless links), given by

$$T_{Prop} = \frac{d}{v}. \tag{1}$$

B. Transmission Latency

The time taken to transmit all of the bits in each packet via a communication link of capacity C (in bps) [5, 20] is the transmission latency. This latency is referred to as store-and-forward (serialization) delay, and is dependent on the link rate and packet (or message) size, M (in bits) [21], given by

$$T_{Tran} = \frac{M}{C}. \tag{2}$$

C. Processing Latency

The time taken to examine packets by hardware and software such as interfaces, applications and network protocols [5, 20] is the processing latency. Layer interactions with packets generate packet-switching (processing) delay. This may be negligible compared to other delays due to the current high speed of devices and software execution. The processing latency is mathematically expressed as the ratio of buffer size, B_S (in bits) and device rate of processing, S_P (in bps):

$$T_{Proc} = \frac{B_S}{S_P}. \tag{3}$$

D. Queuing Latency

The queuing latency is the time spent by the packet in the output buffer, waiting (queuing) before transmission onto the link [5, 20, 21]. This latency is dependent on the number of packets waiting for transmission. Therefore, it is proportional to the traffic (heavy or light) and average packet arrival rate on the link. Considering the M/M/1 queuing model [21], the latency can be expressed as

$$T_{Queue} = \sum_{i=1}^{N} \rho_i T_i\,; \quad \rho_i = \frac{\lambda_i}{C_i} \tag{4}$$

where ρ_i is i^{th} link loading, T_i is average i^{th} system (link/node) delay, λ_i is average (i^{th} packet arrival rate) external traffic on link in packet per second (pps), C_i is i^{th} link capacity (service rate), and N is the finite number of links (edges) or nodes (vertices) along the traffic path.

E. Correlation Function

Cross-correlation is used for computing the time delay between a pair of signals (Tx and Rx) through convolution to find the maximum similarity of the signals [23]. The cross-correlation for discrete time could be written as

$$R_{xy}[n] = \sum_{m=-\infty}^{\infty} x[m]y[m+n]. \tag{5}$$

where the peak value can be achieved when the two signals are exactly the same [22]. This peak can then be used to compute the time shift between two signals. Equations (1)–(5) form a general model of latency and the basis for the actual measurement of total E2E latency using the proposed scenarios in our framework [5, 21].

IV. Experimental Setup and Study Scenarios

Satellite networks on GEO were used as hypothetical reference paths to develop and test our method using two case study scenarios. This involved complete satellite and heterogeneous links to enable performance evaluation and analysis of different link types that are envisioned to be integral part of next generation 5G networks. These study scenarios are code-named satellite-satellite link (SSL), and satellite-terrestrial link (STL). The different experimental setups used are explained in the following subsections and are illustrated in Fig. 1.

A. Satellite-Satellite Link

The SSL established E2E connectivity between two semi-fixed SUTs at about 1m apart and pointed to the satellite through a Line-of-Sight link. Adjustments were made until satisfactory signal strength for communication was obtained within 70-90%. End-users and SUTs were connected using Wi-Fi link 1-20m apart. This scenario was developed to measure the actual E2E latency of satellite link testbed as shown in the top region of Fig. 1.

B. Satellite-Terrestrial Link

The STL E2E connection was achieved through a heterogeneous network consisting of satellite, Wi-Fi and Public land mobile network (PLMN) links. The user device on the satellite link end was connected to the satellite-pointed SUT via Wi-Fi, while the user device on the terrestrial link end was connected to PLMN via the nearest local base transceiver station (BTS). This scenario was developed to measure E2E latency of heterogeneous network testbed as seen in the lower region of Fig. 1.

Communications were routed and processed by a gateway station (GWS) on the ground after being reflected by a transparent satellite, forming a star topology [18].

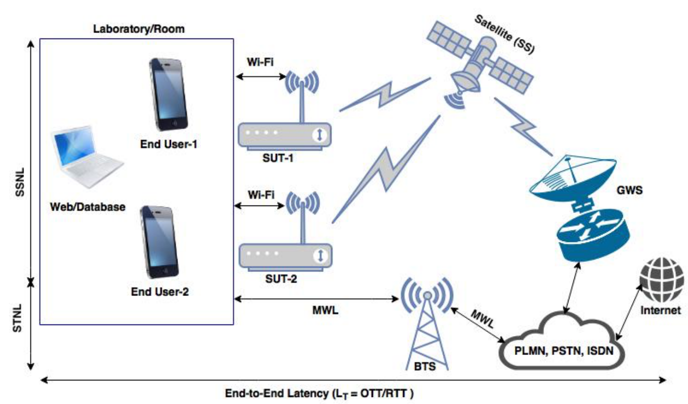

Fig. 1. Experimental Setup of Scenarios

C. Data Acquisition Process

Fig. 2 shows the experimental data acquisition process using the scenarios depicted in Fig. 1, by transmitting (Tx) audio signal from End-user 1 and receiving (Rx) by End-user 2. The correlation function in (5) was used to compute the time delay (latency) between the transmitted and received signals produced under different scenarios through SNPs.

D. End-to-End Latency Model For Scenarios

Three mathematical models were established from the two scenarios and network topology. The first is a general model described as linear summation of one-trip-time delay (propagation, processing, queuing and transmission) along the path as given by

$$T_L = KT_{Prop} + \sum_i T_i \ ; \ K = 1, 2, ..., N \tag{6}$$

where T_L is the total time delay (latency) in milliseconds (ms), K is the number of link(s) traversed by the traffic from source to the destination and sum of T_i represents non-propagation latency components such as queuing, processing and transmission latencies.

The other two models as described by (7) and (8) were developed from (4) and the two scenarios (SSL and STL) considering the E2E signal path topology.

$$T_{SSL} = K_{SL}T_{PropS} + \sum_i T_i \tag{7}$$

$$T_{STL} = K_{SL}T_{PropS} + \sum_i T_i + K_{TL}T_{PropT} + \sum_j T_j \tag{8}$$

Considering our two study scenarios, K_{SL} in (7) and (8) would assume a constant value of 4 and 2, respectively, while K_{TL} has respective values of 0 and 1 in (7) and (8).

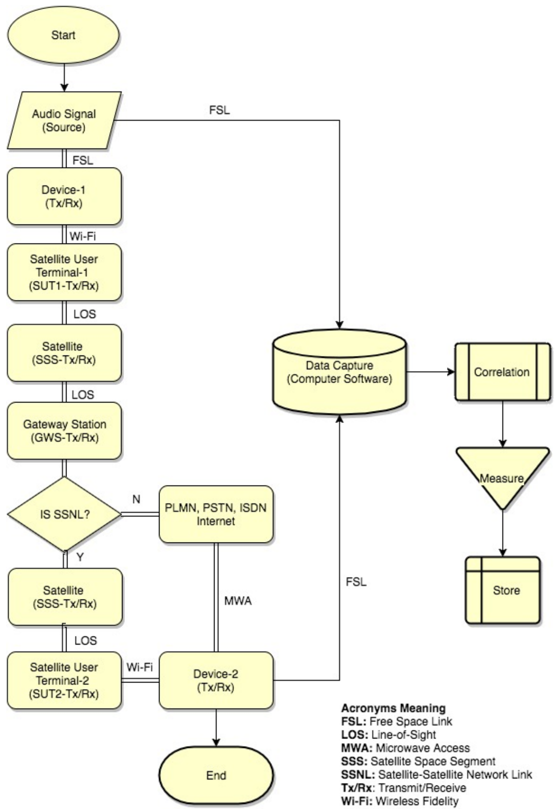

Fig. 2. Data Acquisition Flow Diagram

V. Results and Analysis

Results were obtained using two different SNPs and the two case study scenarios described in Section IV. The first set of experimental results were obtained daily using satellite network provider 1 (SNP1) terminals for sixteen (16) days as shown in Fig. 3, while the second set was obtained to increase the data resolution by collecting measurements in the morning (M), afternoon (A) and night (N) for fifteen days using SNP2 and is shown in Fig. 4. These data allowed us to study and analyze the latency performance of the two networks using the two developed scenarios as well as the time of day.

A. Daily Performance Analysis of Scenarios

The daily experimental measurements with SNP1 are shown in Fig. 3 and exhibit better performance of latency for STL with average of 956ms and standard deviation of 25ms. SSL has higher average of 1417ms and standard deviation of 31ms using the same network provider (SNP1). The latency was observed to be varying by the day having different values on different days as in Fig. 3 due to different traffic patterns each day and the stochastic nature of latency [21]. These variations may affect IP packet delay variation (jitter) thereby degrading performance of jitter sensitive applications. To improve the resolution of the data, measurements were conducted three times a day using SNP2 as shown in Fig. 4.

For SNP2, SSL was observed to have better performance with average latency of 957ms and standard deviation of 64ms, STL performed worst in this case with high average of 1239ms and 72ms standard deviation. The statistical results of the measurements are summarized in Table I.

The resulting analysis of the latency measurements shows that SNP1 performed better under the STL scenario, while SPN2 performed better under the SSL scenario. SNP1 provides broadband Internet service while SNP2 only provides limited Internet services, thus SNP1 may provide an additional advantage of fast Internet access in remote isolated areas and faster voice service to terrestrial destinations in the cities where they exist. SNP2 may serve better as a voice service to connect two remotely isolated users via satellite links in the absence of terrestrial infrastructure on both sites.

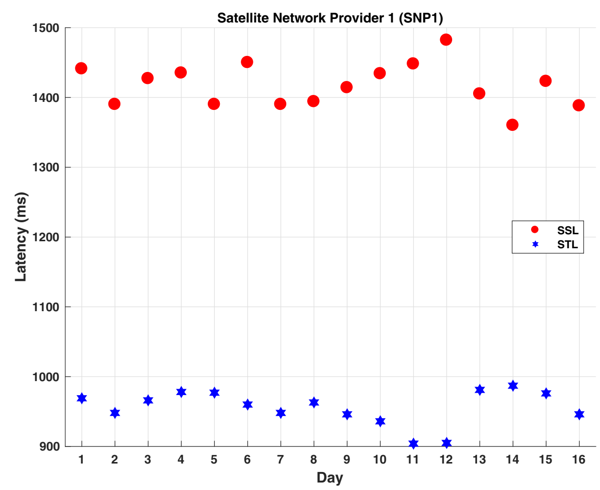

Fig. 3. Performance Comparison of Scenarios with Network Provider 1

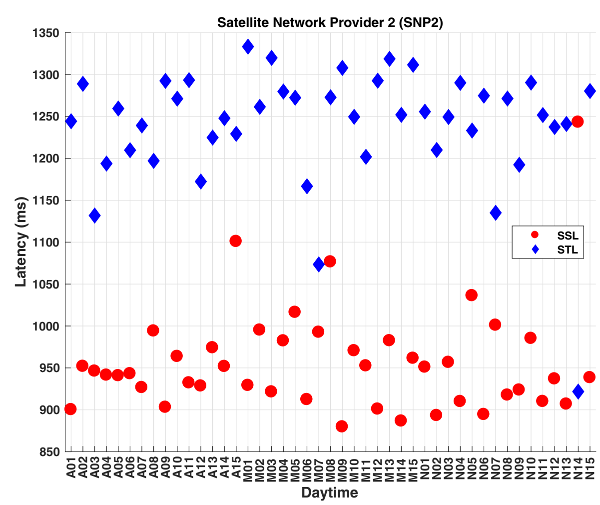

Fig. 4. Performance Comparison of Scenarios with Network Provider 2

TABLE I
PERFORMANCE SUMMARY AND COMPARATIVE ANALYSIS

Network		SNP1		SNP2	
Scenario		SSL	STL	SSL	STL
Latency (ms)	Max	1482	987	1243	1333
	Min	1360	904	880	992
	Avg	1417	956	957	1239
	Stddev	31	25	64	72

B. Daytime Performance Analysis of Scenarios

A performance analysis of different times of the day was considered by taking measurements in the morning (M), afternoon (A), and night (N) using the SSL and STL scenarios. Fig. 5 shows the daily performance using SSL, while Fig. 6 shows that of STL. The time of day with the lowest average latency (i.e., best performance) using SSL occurred in the afternoon with 953ms and 48ms standard deviation. The highest average latency (i.e., worst performing) occurred at night with 960ms and 88ms standard deviation. This time of the day also has the highest value of 1243ms with SSL scenario as seen in Fig. 5. The STL performance shown in Fig. 6 depicts better performance at night with average latency of 1222ms and 93ms standard deviation while the highest average of 1261ms occurred in the morning, although its standard deviation of 69ms is less compared to that of the night.

Overall, SSL has lowest latency of minimum 880ms in the morning and maximum of 1243ms at night, while STL has lowest standard deviation of 47ms in the afternoon. The statistical results of the measurements recorded throughout the day are summarized in Table II.

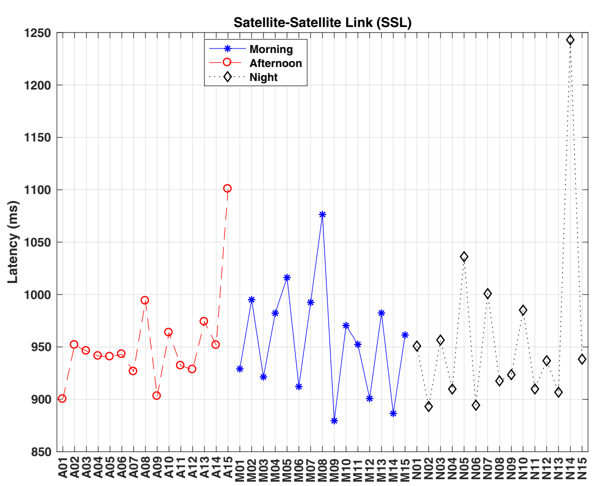

Fig. 5. Daytime Performance Comparison with Satellite-Satellite Link

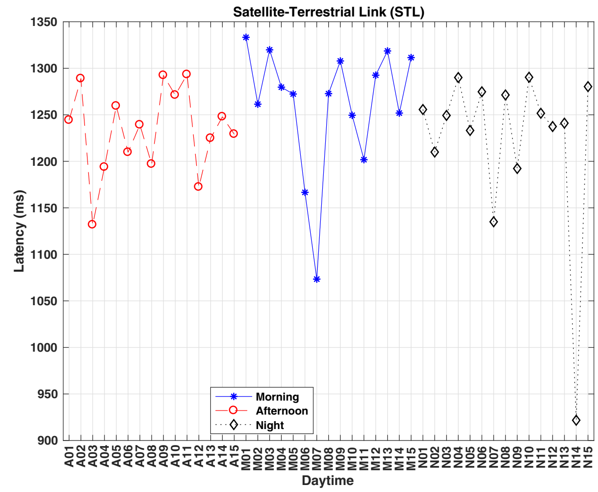

Fig. 6. Daytime Performance Comparison with Satellite-Terrestrial Link

TABLE II
DAYTIME PERFORMANCE COMPARISON OF SCENARIOS

Scenario	SSL			STL		
Daytime Latency (ms)	L_M	L_A	L_N	L_M	L_A	L_N
Max	1076	1101	1243	1333	1293	1290
Min	880	900	893	1073	1132	922
Avg	957	953	960	1261	1233	1222
Stddev	54	48	88	69	47	93

Statistical information such as the mean and standard deviation are important in latency characterization [21] in order to analyze the effects of E2E latency on data transfer across the network and to identify the achievable level of QoS and performance [1, 5, 6]. Uncertainties typically exist within a communications network; those due to latency can be modeled stochastically in (6), if (6) is viewed as a linear sum of random variables defined by (1)–(4) and (7) and (8), as is the case for our proposed scenarios. The main source of uncertainty likely lies in (4) [21] and in future work, this will be investigated further.

The performance differences observed across providers may have been due to the respective performance objectives of each network provider. For example, SNP1 may be more focused on (as determined by the customer offerings) the provision of Broadband Internet from remote areas using broadband global area network terminals that allow up to a 464kbps data rate. Therefore SNP1 may be less optimized for a SSL connection and more focused on heterogeneous links similar to those set up in the STL scenario. SNP2 may be more focused on voice communications than Internet provision over satellite and is only designed for light mobile Internet data usage such as email, instant messaging, and mobile web browsing on smartphones at rates of only 60-144kbps. Therefore, the performance for SNP2 may be more optimized for an SSL-type connection (compared to STL, in which it exhibited lower performance in terms of latency).

VI. Conclusions

This study proposes two scenarios for a precise E2E latency measurement in a satellite IP environment. The latency performance of the proposed scenarios were investigated with two different SNPs. Our results showed that latency performance in a satellite IP network depends on the type of scenario and network provider. Better performance of latency varies with scenario, network provider and time of day. The daily analysis showed, SSL performance was better using SNP2 while STL was better using SNP1. However, SSL exhibited the lowest latency (880ms) in the morning with maximum (1243ms) at night, while STL had lowest standard deviation (47ms) in the afternoon. Future work will investigate the performance of applications with the recorded maximum, minimum and average latency values. We also intend to study and apply techniques, such as performance enhancement proxies, delay-tolerant network architecture and User Datagram Protocol, to minimize latency and optimize performance of applications over satellite IP networks.

Acknowledgements

This work was supported by the UK EPSRC (grant no. EP/P005950/1), in part by the European Commissions Horizon 2020 framework programme (H2020/2014-2020) under grant agreement no. 734325 TESTBED project (http://testbed-rise.com/). The authors would like to thank Petroleum Technology Development Fund (PTDF), Federal Republic of Nigeria and Global Challenge Research Fund (GCRF), United Kingdom for funding the research activities.

References

[1] S. Kota and M. Marchese, "Quality of Service for Satellite IP Networks: A Survey", *International Journal of Satellite Communications and Networking*, vol. 21, pp. 303-349, 2003.

[2] W. B. Group. (2015) Rural Population (% of Total Population). [Online]. accessed: 31/03/2017. Available: http://data.worldbank.org/indicator

[3] G. Faria *et al.*, SatMagazine: Worldworld Satellite Magazine. Nov. 2016 edition, 8-79. [Online]. Available: http://www.satnews.com/magazines.php

[4] P. Woodhead, "Advances in Delay-Tolerant Networks (DTNs): Architecture and Enhanced Performance," in *Electronic and Optical Materials*, vol. 67, pp. 1-269, 2015.

[5] P. Hinton, E. Baker, and C. Hill, "Latency Time for lawyers to get up to Speed?", *Computer Law & Security Review*, vol. 28, no. 3, pp. 340-346, 2012.

[6] T. Braun, M. Diaz, E. J. Gabeiras, and T. Staub, *End-to-End Quality of Service Over Heterogeneous networks*. Berlin, Germany: Springer, 2010.

[7] A. E. Gharsellaoui, S. A. M. Ghanem, D. Tarchi, and A. Vanelli-Coralli, "Adaptive network coding schemes for satellite communications," in *Proc. 14th IEEE Advanced Satellite Multimedia Systems and Signal Processing for Space Communications Conference (ASMS/SPSC)*, Palma de Mallocra, Spain, 5-7, 2016.

[8] R. Ferrús *et al.*, "SDN/NFV-enabled satellite communications networks: Opportunities, scenarios and challenges," *IEEE Physical Communication*, vol. 18, no. 2, pp. 95-112, 2016.

[9] X. Artiga, J. Núñez-Martnez, A. Pérez- Neira, G. J. Lendrino Vela, J. M. Faré García, and G. Ziaragkas, "Terrestrial-Satellite Integration in Dynamic 5G Backhaul Networks," in *Proc. Advanced Satellite Multimedia Systems Conference and the 14th Signal Processing for Space Communications Workshop*, 2016.

[10] M. Marchese, *QoS Over Heterogeneous Networks*. West Sussex, UK: John Wiley, 2007.

[11] S. Utsumi and S. M. S. P. Zabir, Sumet, 11A new explicit congestion notification scheme for satellite IP networks," *Journal of Network and Computer Applications*, vol. 75, pp. 169-180, 2016.

[12] C. Caini, P. Cornice, R. Firrincieli, and D. Lacamera, "A DTN approach to satellite communications," *IEEE Journal on Selected Areas in Communications*, vol. 26, no. 5, pp. 820-827, 2008.

[13] D. Kirchner, "Two-Way Time Transfer Via Communication Satellites" *Proceedings of the IEEE*, Vol. 79, No. 7, pp. 983 - 990, 1991.

[14] C. Niephaus, M. Kretschmer, and G. Ghinea, "QoS Provisioning in Converged Satellite and Terrestrial Networks: A Survey of the State-of- the-Art," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 4, pp. 2415-2441, 2016.

[15] Y. Xiao, Y. Xu, H. Sun, X. Xu, and A. Zhou, "A Precise Real-Time Delay Calibration Method for Navigation Satellite Transceiver," *IEEE Transactions on Instrumentation and Measurement*, vol. 65, no. 11, pp. 2578-2586, 2016.

[16] G. Huston, "Measuring IP Network Performance," *The Internet Protocol Journal*, vol. 6, no. 1, pp. 2-19, Cisco Systems Inc., 2003.

[17] X. Zhang, H. Yin, D. Oliver Wu, G. Min, H. Huang, Y. Zhang, "SSL: A Surrogate-based Method for Large-scale Statistical Latency Measurement," *IEEE Transactions on Services Computing*, vol. PP no. 99, [IEEE Early Access Articles], 2017.

[18] T. Gayraud and P. Berthou, "A QoS Architecture for DVB-RCS Next Generation Satellite Networks," EURASIP Journal on Wireless Communications and Networking, Satellite Communications, pp. 79-87, 2007.

[19] G. Vankatesh and K. C. Wang, "Estimation of Maximum Achievable End-to-End Throughput in IEEE 802.11 Based Wireless Mesh Networks," in *Proc. 34th IEEE Conference on Local Computer Networks (LCN)*, pp. 1110-1117, Zurich, Switzerlanland, 20-23, 2009.

[20] T. Eylen and C. F. Bazlamacci, "One-Way Active Delay Measurement With Error Bounds," *IEEE Transactions on Instrumentation and Measurement*, vol. 64, no. 12, pp. 3476-3489, 2015.

[21] J. W. Stahlhut, T. J. Browne, G. T. Heydt, and V. Vittal, "Latency Viewed as a Stochastic Process and its Impact on Wide Area Power System Control Signals," *IEEE Transactions on Power Systems*, vol. 23, no. 1, pp. 84-91, 21, 2008.

[22] K. Dudacek Jr., K. Dudacek, and V. Vavficka, "Comparison of Short Delay Measurement Methods," in *Proc. International Conference on Applied Electronics (AE)*, Pilsen, Czech Republic, 8-9, 2015.

[23] X. Jin, Y. Hua, and Y. Cao, "Research on A New Method of Time Delay Measurement in Telephone Time Service," in *Proc. IEEE International Frequency Control Symposium (IFCS)*, 19-22, 2014.

[24] M. M. McMahon and R. Rathburn, "Measuring Latency in Iridium Satellite Constellation Data Services," *Directorate for Information Operations and Reports*, Report, US Naval Academy, 2005.

[25] D. S. Miller, N. Levanon, and A. Agrawal, "System and Method for Measuring Round Trip Delay in a Satellite Communication System" *United States Patent*, No. US006366762B, 2002.