[ITID.11.2.Zheleval(2).ID1228.1-14-15.ms.author.FINAL]

Research Article

Internet Bandwidth Upgrade: Implications on Performance and Usage in Rural Zambia

Mariya Zheleva
University at Albany SUNY[1]

Paul Schmitt
Morgan Vigil
Elizabeth Belding
University of California, Santa Barbara

Abstract

Broadband Internet access has become a critical part of socioeconomic prosperity; however, only six in 100 inhabitants have access to broadband in developing countries. This limited access is driven predominately by subscriptions in urban areas. In rural developing communities, access is often provided through slow satellite or other low-bandwidth long-distance wireless link, if available at all. As a result, the quality of Internet access is often poor and, at times, unusable. In this article we study the performance and usage implications of an Internet access upgrade, from a 256 Kbps satellite link to a 2 Mbps terrestrial wireless link in rural Zambia. While usage did not immediately change, performance improved soon after the upgrade. By three months post-upgrade, subscribers began to use the faster connection for more bandwidth-hungry applications such as video streaming and content upload. This change in use resulted in a dramatic deterioration of network performance, whereby the average round-trip time doubled, the number of bytes associated with failed uploads increased by 222% and failed downloads by 91%. Due to this deteriorated performance, the use of more bandwidth-hungry applications as observed three months post-upgrade did not persist over the long term. As uploads became largely unsuccessful, users stopped initiating uploads and, instead, switched to using the increased capacity for heavier downloads. Thus, while an Internet access upgrade should translate to improved performance

To cite this article: Zheleva, M., Schmitt, P., Vigil, M., & Belding, E. (2015). Internet bandwidth upgrade: Implications on performance and usage in rural Zambia. *Information Technologies & International Development*, *11*(2), XXX–XXX.

[1] This work was completed as a part of Mariya Zheleva's PhD dissertation at UC Santa Barbara.

and user experience, in rural environments with limited access speed and growing demand, it can bring unexpected consequences.

Keywords: ICT4D, rural area networks, digital divide, bandwidth increase, satellite, wireless terrestrial

1. Introduction

Access to the Internet is critical for improving the wealth of nations and promoting freedom. Bright examples of advancements facilitated by Internet access span democratic change (Allagui & Kuebler, 2011), government (Ndou, 2004), e-learning (Sife, Lwoga, & Sanga, 2007), and health care (Fraser & McGrath, 2000). Broadband Internet access, however, is still largely unavailable in developing countries, with only 6% of the population having broadband connectivity (ITU, 2013), the majority of whom reside in urban areas.

Recent efforts to bring connectivity to rural areas of the developing world employ asymmetric satellite or other low-bandwidth wireless links (Matthee, Mweemba, Pais, van Stam, & Rijken, 2007; Surana et al., 2008). At the same time, the bandwidth demand of online applications is increasing; for example, the average web-page size has grown 110 times since 1995 (King, 2012). As a result, residents of developing rural regions access the web with inadequate connectivity for the bandwidth requirements of modern content. These opposing trends in content growth and limited capacity render Internet access frustrating or even unusable (Du, Demmer, & Brewer, 2006; Johnson, Pejovic, Belding, & van Stam, 2011) in many developing areas.

Previous work on traffic analysis shows a “strong feedback loop between network performance and user behavior” (Johnson et al., 2011, p. 1), whereby residents in bandwidth-constrained environments tend to focus more on bandwidth-light applications such as web browsing, as opposed to bandwidth-intensive applications such as multimedia streaming, content upload, and real-time user interaction. In the face of limited bandwidth, the failure rate of uploads is high (Johnson, Pejovic, Belding, & van Stam, 2012), discouraging rural residents from contributing Internet content and resulting in their consumption of largely Western content (Vannini & le Crosnier, 2012). While recognizing the benefits of the Internet, residents of developing regions express concerns that the flood of Western culture, coupled with decreased ability to document and transfer their own traditions, threatens the existence of local cultures (van Hoorik & Mweetwa, 2008).

Our work's focus is in Africa, where the increased fiber-optic capacity (Song, 2014), coupled with higher-bandwidth, lower-latency technologies such as terrestrial microwave wireless, offers hope for improved Internet access in remote areas. In this article we study the implications of an Internet access upgrade from a satellite to a microwave terrestrial link on the performance and Internet use in the rural community of Macha, Zambia. Our work builds on our prior study of rural networks' performance (Johnson et al., 2011) and presents the first real-world comparative study of pre- and post-upgrade Internet use and performance. We collect a longitudinal network traffic trace that captures an upgrade of the gateway capacity and offers a unique opportunity to study the change in user behavior and network performance following an eight-fold increase in access bandwidth. We analyze 3½ months of performance both before and after the upgrade in order to evaluate the immediate and long-term effects of the upgrade. Our results show that while use did not change immediately, application performance improved. The Internet access upgrade broadened users' abilities to access content, use online applications, and express themselves on the Internet. As time passed, however, subscribers began to change their Internet usage behavior, which ultimately resulted in network performance degradation and a subsequent deterioration of the user experience. In particular, in this article we show that as bandwidth increased, users first aggressively tried to access more bandwidth-hungry applications such as P2P file downloads. As their attempts failed, users reverted to using predominantly low-bandwidth HTTP web browsing. The results of our fresh analysis make a strong case that one should not assume that advanced technologies and higher access speed lead to a better user experience and increased adoption of the Internet in rural communities; rather, one should carefully consider the evolution of use and performance so as to assess the actual impact and adoption of Internet technologies.

2. Related Work

Several previous research efforts focus on rural network traffic characterization. Web traffic from Internet cafés and kiosks in Cambodia and Ghana is analyzed in Du et al. (2006). Here, the focus is on the characterization of HTTP traffic to guide caching techniques for web users in developing regions. Ihm, Park, and Pai (2010) focus on understanding the network traffic in developing regions as compared to their OECD counterparts. This article characterizes national traffic patterns based on network use, with the goal of improving caching techniques for developing regions. Anokwa, Dixon, Borriello, and Parikh (2008) identify the impact of latency on network performance in developing regions and propose a flow-based prioritization scheme as a solution. In contrast, our work focuses on a smaller scale and characterizes web traffic to ascertain the impact of a network upgrade on use and performance.

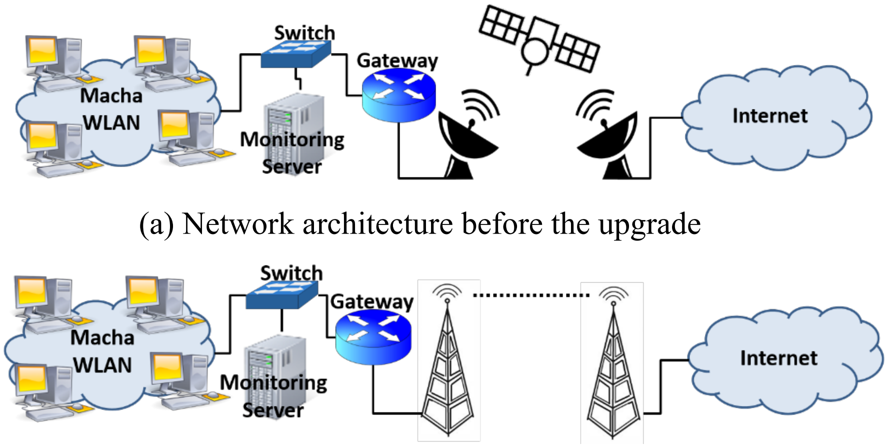
Our analysis of TCP performance builds on the measurements employed by Johnson et al. (2011) by engaging a more in-depth analysis of TCP performance, including measurements of TCP bytes in flight, retransmissions, interpacket arrival times, RTTs (round-trip times), and packet sizes. A *TCP flow* defines a packet stream between a single source-destination pair and can be evaluated using metrics such as control overhead (i.e., fraction of acknowledgments), RTT, and retransmissions. Performance of different types of TCP such as CUBIC TCP, Compound TCP, and TCP Reno interactions is measured via simulation of high-delay wireless networks in Abdeljaouad, Rachidi, Fernandes, and Karmouch (2010). This work has an explicit interest in measuring goodput and TCP fairness. Our analysis is based on TCP performance in a low-bandwidth, high-latency real network. We measure TCP performance in aggregate and as separated by operating system network stacks. We then draw conclusions about TCP fairness of different variants found on the network.

3. Macha

Economics and Demographics. Macha is a typical poor rural village located in the southern province of Zambia. Approximately 135,000 people live in the area, spread in a radius of 35 km around the village. The primary occupation in the village is maize farming. The average estimated income is US$1/person/day—one-fifth the cost of a round trip to the closest town and one thirtieth the cost of a monthly Internet subscription limited to 1 Gb.

Macha has been a local leader in health care and technological innovation. Active village organizations include a hospital and health care research facility as well as MachaWorks, an NGO that maintains a local wireless network, LinkNet. LinkNet distributes Internet access from an Internet gateway over an area of six square kilometers, including the schools, hospital, research institute, and residential areas.

Internet use and provisioning. While Macha is connected to the national power grid, electricity is rarely available in individual households. The lack of electricity, coupled with the high prices for user equipment and Internet provisioning, makes it virtually impossible for Machans to use Internet at home. They typically access the Internet at work, from an Internet café, or at school.



(a) Network architecture before the upgrade

(b) Network architecture after the upgrade

Figure 1. Network architecture and traffic monitoring.

Internet access is distributed from the village Internet gateway to central facilities via a local wireless mesh network (Figure 1) maintained by LinkNet (Matthee et al., 2007). Between 2008 and April 2011, the village was connected to the Internet through a satellite connection that cost US$1,200/month and provided 256 Kbps downlink bursting to 1 Mbps, and 64 Kbps uplink bursting to 256 Kbps. In April 2011, the village Internet access was upgraded to a higher quality microwave terrestrial link [Figure 1(b)] with speeds up to 2 Mbps that cost US$3,600/month. At the time of the upgrade, approximately 300 residents connected to the Internet.

4. Network Analysis

We evaluate network performance and use for three months. We select one month immediately before (we term this *Pre-upgrade* throughout the article) and one month immediately after the upgrade (termed *Post-upgrade*) to measure the short-term impact on network use and performance. We also evaluate one month of traffic approximately three months after the upgrade to determine whether performance changed over time (termed *Long-term*).

We start by describing our traffic collection methodology and our approach to calculating evaluation metrics. We continue with detailed results from our network analysis. We first focus on our overall network performance analysis, which indicates that 93% of traffic traversing the network is TCP. Thus, we focus our analysis of TCP performance following the increased bandwidth. We describe trends in uplink and downlink performance of TCP flows and assess the success and failure rates of these flows. We conclude our TCP analysis by outlining the most popular services. We evaluate network use, focusing on popular URIs (uniform resource identifiers). We conclude by analyzing the geography of network flows initiated in Macha to determine whether Machans used more global services once they had better Internet access.

Pre-upgrade, the network was typically saturated, resulting in high RTTs, congestion, and aborted sessions. Post-upgrade, we observed a decrease in the number of retransmissions and RTTs due to improved network performance and movement away from the saturation point. By three months after the upgrade, the traffic had increased once again to saturation. Our analysis shows a difference in network performance and use Post-upgrade and Long-term: While Post-upgrade user behavior did not change, automatic programs such as software updates were suddenly able to complete, resulting in an increase in traffic demand. In Long-term, subscribers used the faster Internet access for more bandwidth-hungry applications such as video streaming. Once the saturation point was reached in Long-term, network performance deteriorated, but was still better able to support bandwidth-intensive applications than Pre-upgrade. We describe these network use and performance patterns in detail in the following sections.

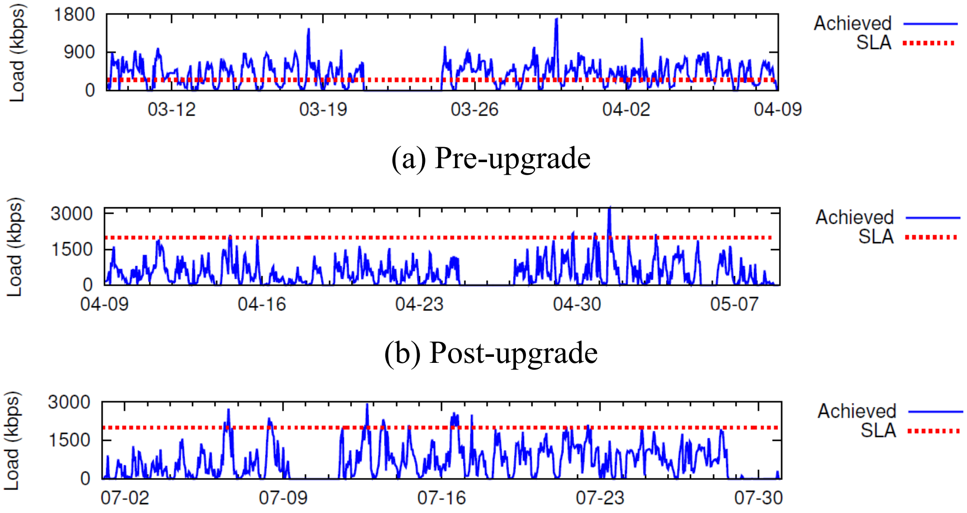
4.1 Methodology

We capture traffic at the Internet gateway in Macha. As shown in Figure 1, we connect a monitoring server to the switch that bridges the Internet gateway and Macha's WLAN. We configure a mirror port at that switch, allowing us to capture all the traffic crossing the WLAN. With user consent, we capture packets and store traces on the monitoring server. During a 2012 field trip to Macha, we offloaded the collected traces to an external hard drive and brought them to our research facility for offline analysis.

Most (93%) of the traffic crossing the Macha network is TCP. Thus, a large portion of our analysis focuses on TCP flow. We extract these metrics by running the network analysis tool *tshark* in an offline mode on the collected traces. To evaluate TCP flow completion and failure, we developed a tool that reassembles unidirectional flows from a list of packets based on packet signature (source IP, source PORT, destination IP, destination PORT, timestamp). In the process of flow reassembly, we count the number of packets and bytes associated with this flow and calculate its duration. We calculate the *packet inter-arrival time* (*IAT*) as the difference in time of consecutive packets. To obtain bidirectional flows, we combine the unidirectional flows based on flow signature and timestamp.

4.2 Overall Network Performance

Traffic load. We start with an evaluation of the traffic load. We calculate the *traffic load* as the aggregate number of bits that traverse the gateway each hour divided by the number of seconds in an hour. Our results capture the average combined uplink and downlink loads. We find that the average traffic load Pre-upgrade is 367.3 Kbps, Post-upgrade is 495.3 Kbps, and Long-term is 648.1 Kbps. Figure 2 plots over time the traffic load averaged per hour in blue and the service level agreement (SLA) with the Internet provider in red.[2] In the period before the upgrade, demand frequently exceeded the SLA of 256 Kbps. This is less often the case during the period immediately after the upgrade, as users have not yet adapted to the increased bandwidth. However, three months after the upgrade, the demand often approaches the SLA. As detailed later in our analysis, this is likely due to changed patterns of use whereby users began to access more bandwidth-hungry applications once more bandwidth was available. Gaps in the plots correspond to periods in which traffic captures were unavailable due to power or network outages.



(a) Pre-upgrade

(b) Post-upgrade

(c) Long-term

Figure 2. Traffic load over time.

[2] Note that while the guaranteed speed was 256 Kbps, bursts of up to 1 Mbps were possible depending on link utilization. This is why the actual traffic load Pre-upgrade consistently exceeds the SLA of 256 Kbps.

Table 1. Average TCP Statistics.

	Total Gb	Total packets (10^6)	Total control packets (%)	Average RTT(s)	Total retransmissions (%)
Pre-upgrade	123	373	56.59	0.1436	1.12
Post-upgrade	163	338	47.69	0.1085	1.09
Long-term	210	432	49.72	0.3190	1.16

General trends. We continue our evaluation by discussing general trends over the three observed periods. Table 1 presents a detailed look into performance. As can be seen, the total bytes traversing the gateway nearly doubled over the course of three months. The number of packets dipped Post-upgrade as the same traffic demand was first accommodated with fewer retransmissions. Usage changed over time, resulting in a drastic increase in the number of bytes traversing the gateway and a corresponding increase in the number of packets.

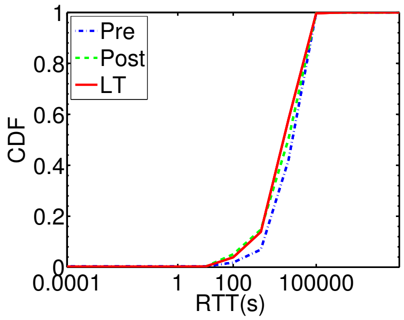


Figure 3. RTT.

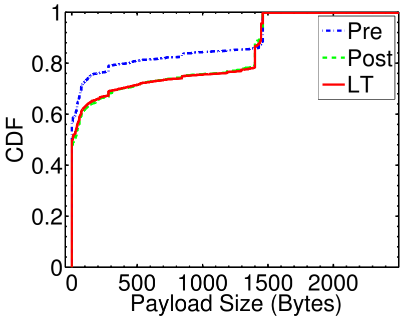


Figure 4. Payload size.

A similar trend is observed in RTTs. While immediately after the upgrade the average RTT decreased by about 35 milliseconds (ms), it nearly tripled as time progressed. We explore RTT dynamics in detail in Figure 3, which plots a cumulative distribution function (CDF) of RTT for the three periods. We observe a long-tail distribution of RTT in Post-upgrade and Long-term performance; however, the median RTT values for those two periods are lower than those observed Pre-upgrade. As can be seen later in Section 4.5, the long-tail distribution of RTT after the upgrade is due to changed browsing habits and a tendency to use services that are physically farther away (such as streaming video from servers abroad). We discuss usage patterns in depth in Section 4.4 to validate our hypothesis.

Next, we analyze the control overhead in the network. Ideally, the control overhead should be minimal to ensure reliable packet delivery while maximizing the available bandwidth for actual data transmission. We calculate the control overhead as the fraction of control packets (e.g., retransmissions, ACK, SYN, FIN) from all packets in a TCP flow. As Table 1 and Figure 4 indicate, the fraction of control packets decreased after the link upgrade, from 56.59% to 47.69%, then slightly increased in Long-term to 49.72%. The number of retransmissions follows a similar trend. This overall decrease in control overhead is attributable to improved network performance, resulting in less protocol overhead from retransmissions and repeated acknowledgments as well as fewer attempts to re-establish failed TCP sessions. The uptick in retransmissions and control packets over the Long-term is attributable to decreased performance due to the increase in offered load to the new saturation point.

4.3 TCP Performance Analysis

TCP performance improved significantly after the link upgrade. One factor that indicates this improvement is the *bytes in flight* (or *congestion window*), which is the fraction of sent data that has not yet been acknowledged. Intuitively, the better the link performance, the larger the congestion window, which allows more data to be sent on the link before an acknowledgment is received. Figure 5 presents a CDF of bytes in flight for the three periods. Immediately after the upgrade, the bytes in flight drastically increased and continued to grow over the Long-term.

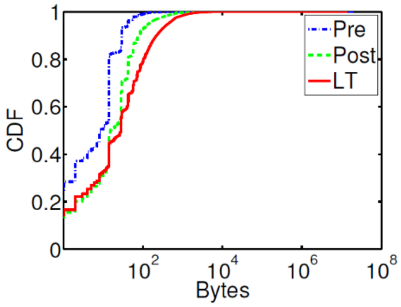


Figure 5. Bytes in flight.

We continue our analysis by exploring TCP flow trends following this improved TCP performance. We start by presenting general trends of TCP flows in Table 2. The bytes associated with TCP flows increased after the upgrade and continued growing in Long-term. This increase in bytes is due to increased demand in browsing and streaming applications as well as increased rate of completion of larger TCP flows. We evaluate flow success and failure rates later in this section.

We next examine the average flow size across the three periods. As can be seen in Table 2, the flow size doubled Post-upgrade and continued to increase in Long-term. The increase of flow size is attributable to different applications using the link immediately after the upgrade and in Long-term. Indeed, we see many software updates Post-upgrade, which are then replaced by other applications. These we explore in Section 4.4. The average packet IAT decreased Post-upgrade, then increased Long-term.

Table 2. TCP Flow Analysis.

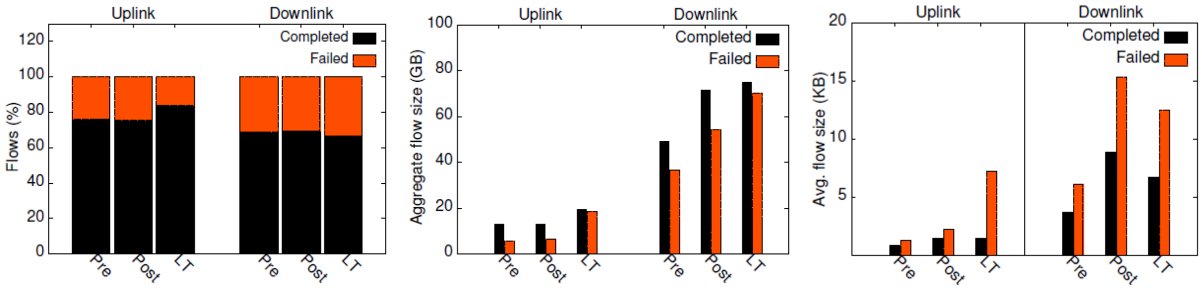
Period	Total Gb	Flow size (B)	IAT(s)
Pre-upgrade	105	3,445	1.92
Post-upgrade	145	7,708	1.49
Long-term	183	8,103	1.91

Uplink and downlink flows. Next, we differentiate flows into uplink and downlink to analyze direction-specific trends. In Table 3 we present aggregate bytes in each direction. Both uplink and downlink bytes increased after the link upgrade. While downlink increased rapidly, uplink remained almost unchanged Post-upgrade, but increased drastically over the Long-term. Average uplink packet and flow sizes increased Post-upgrade and in Long-term. Concurrently, downlink packet and flow sizes increased Post-upgrade, then slightly decreased in Long-term. These trends can be explained with differences in applications accessing the Internet as well as with changes in network performance due to link saturation in Long-term. The rapid increase in downlink activity Post-upgrade is due to an increase in automated activities such as software updates. The increase in uplink happens more gradually and is attributed to a slower change in user behavior and, in particular, a gradual increase in content upload attempts.

Table 3. TCP Flow Uplink (UL) and Downlink (DL) Characteristics.

	Total Gbs		# of flows (x 10^5)		Packet size (B)		Flow size (B)	
	UL	DL	UL	DL	UL	DL	UL	DL
Pre-upgrade	18.65	85.9	189	194	132.7	616.0	988.9	4,427
Post-upgrade	19.26	125.7	114	116	158.6	877.0	1,691	10,856
Long-term	38.14	145	157	168	227.7	787.6	2,422	8,613

Finally, we concentrate on the number of flows. As can be seen in Table 3, the number of flows in both up- and downlink directions decreased dramatically Post-upgrade, then increased. The initial decrease is attributable to a higher rate of successful flow completions, which directly results in fewer flow re-initializations. The subsequent increase in the Long-term is due to a combination of increased user activity as well as an increase in the flow failure rate as user demand again reaches link capacity.



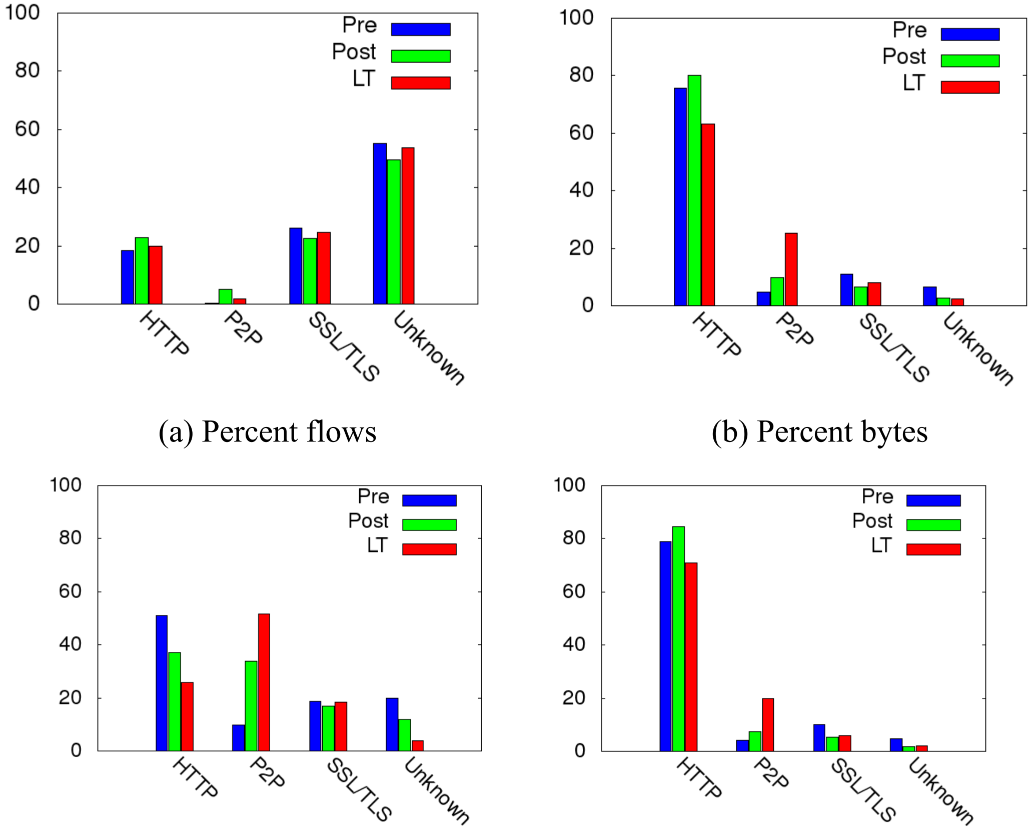
(a) Percent of flows (b) Aggregate bytes (c) Average flow size

Figure 6. TCP flow success and failure in uplink and downlink direction.

TCP flow completions and failures. We now focus on flow completions and failures. In compliance with RFC 793 (which mandates the operation of the TCP protocol), we accept that a FIN packet indicates a completed flow, while lack of a FIN packet or exchange of an RST (reset) packet indicates a failed flow. Figure 6(a) presents the fraction of completed and failed flows in uplink and downlink in each period. The completion rate of uplink flows remains unchanged Post-upgrade and slightly increases in Long-term. On the other hand, the downlink flow completion rate remains unchanged. In Figure 6 we also analyze success and failure trends with respect to byte volume and flow size. Figure 6(b) plots the aggregate flow size in bytes for each direction. The aggregate size of both completed and failed uplink flows remains unchanged Post-upgrade but increases in Long-term. Unfortunately, the number bytes in failed flows approaches the number of bytes in completed flows, which indicates that while users were likely more successful in uploading content, over the Long-term, half of all content that users generate fails to upload. Similarly, in terms of the size of downlink flows, we see a gradual increase in successful downloads; however, over the Long-term the aggregate size of download flows that fail also increases, nearly reaching the aggregate size of successful downloads.

We evaluate average flow size of completed and failed flows in Figure 6(c). We find the size of an individual flow by summing the packet sizes of all packets associated with the flow. In the uplink direction, the average size of failed flows over the Long-term is four times larger than the size of completed flows. This implies that smaller content uploads such as Facebook posts and small images are more likely to succeed, while larger uploads of videos or high-quality pictures had a higher probability of failure. Similarly, the average size of failed downlink flows is persistently higher than that of completed flows. This points to the success of smaller flows such as email and web access, while the increase in downlink average flow size for failed flows is likely due to increased attempts to download larger files such as video content.

Most popular services. We analyze the most popular services accessed by Machans during the three periods. For this analysis we use a tool called Tstat,[3] which performs layer-7 packet inspection to determine service type.



(a) Percent flows

(b) Percent bytes

(c) Percent UL bytes

(d) Percent DL bytes

Figure 7. Tstat analysis of service types.

Our initial results show that the most popular services across periods were HTTP, P2P, and SSL/TLS; thus, the remainder of this section focuses on those services. Figure 7 presents our results for (a) the percent of bidirectional flows to each service, (b) the percent of total bytes, (c) uplink bytes and (d) downlink bytes. We examine trends across the three periods and look for correlations among services to capture changes in user behavior. A substantial number of flows could not be classified by the layer-7 inspection and were labeled Unknown. As can be seen in Figures 7(a) and (b), Unknown traffic constitutes the majority of the flows observed across all periods; however, the fraction of bytes due to Unknown flows is minimal. We postulate that the Unknown traffic is due to malware such as port scans, which generates a large volume of small

[3] http://tstat.tlc.polito.it/

flows. The fraction of bytes due to Unknown traffic [Figures 7(b), (c), and (d)] decreased Post-upgrade and over the Long-term. This is likely due to the successful completion of software updates, which allow computers to better defend against malware. SSL/TLS is the next most accessed service, followed by HTTP. On the other hand, in terms of generated bytes, HTTP is far more prevalent than SSL/TLS. P2P is the least popular in terms of percent of flows; however, the bytes due to P2P flows are substantial.

Exploring trends across the three analyzed periods, we observe a reverse correlation between the bytes due to HTTP and P2P flows over the Long-term [Figures 7(b), (c), and (d)]. As Figure 7(b) shows, there is an increase in both services Post-upgrade. Over the Long-term, the bytes due to HTTP activity dropped and those due to P2P flows increased significantly. This indicates a shift in user interest from web browsing to P2P file downloads. Further analysis of the upload and download bytes [Figures 7(c) and (d)] confirms this trend. We see a gradual increase Post-upgrade in both uplink and downlink P2P bytes, while a more substantial increase is observable over the Long-term. Facebook, Google, and software updates are among the top applications accessed through HTTP. Of the P2P traffic, 40% is through BitTorrent applications and the remainder is other unclassified P2P traffic. The nature of BitTorrent applications provides a partial explanation of the increase in both uplink and downlink P2P bytes. When a user downloads a torrent, the user is termed a *seeder* (simultaneously a source of the file and an uploader of that file to other torrent clients). As a result, 28% of all uploads are BitTorrent uploads, in other words, (potentially unintentional) seeding activity. The remaining 72% of uploads are user initiated and consist of HTTP (23%), SSL/TLS (19%), unclassified P2P (23%), and other (7%).

4.4 Network Usage

The most prevalent application protocol used in Macha is web browsing. Most (87%) of the Pre-upgrade traffic in up- and downlink directions is a combination of HTTP and HTTPS. This number remains almost unchanged Post-upgrade and drops to 71% in the Long-term. At the same time, P2P and Unknown traffic, which includes services to unspecified ports (e.g., Skype and BitTorrent) increased in the Long-term. This is a strong indication of a shift of usage habits to more real-time services, which is typical for well-connected Internet users.

In this section we investigate web traffic to understand user behavior. We correlate our findings regarding popular applications with network performance and make inferences about the user experience based on this correlation.

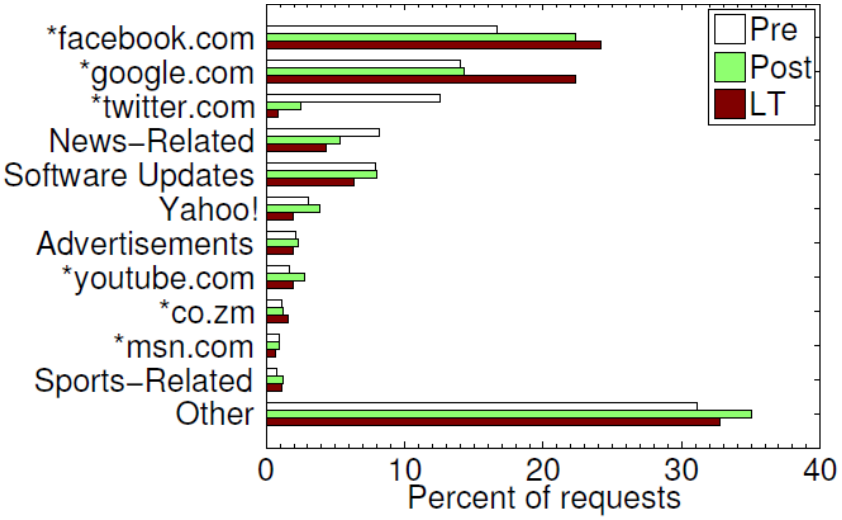
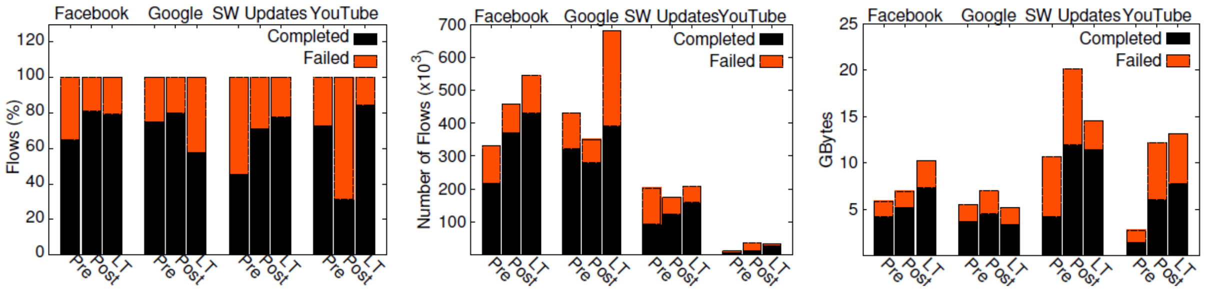


Figure 8. Popular URI requests.

Popular URIs. We begin our analysis by evaluating popular web services. Figure 8 shows web URI requests classified by destination domain. For clarity of presentation we combine related sites (e.g., Facebook with the associated content delivery networks). Facebook and Google are clearly the most popular sites. Both sites see a significant increase in the percentage of requests after the link upgrade, further extending their dominance. At the same time, access to Twitter, the third most popular domain Pre-upgrade, dropped significantly. The “News” classification includes *postzambia.com, *lusakatimes.com, and BBC news sites. The popularity of these websites is important as it shows user interest in local content, a pattern also seen in Johnson, Belding, Almeroth, and van Stam (2010).

Software update sites such as those associated with Windows, Adobe, and Ubuntu remain relatively unchanged throughout the measurements; however, as shown later in this section, their completion rate significantly increases Post-upgrade. In general, while requests for multimedia-rich sites or large binary downloads remain the same across periods, the actual traffic associated with such requests increases as more requests are successfully completed. We explore TCP sessions count and size later in this section.

Advertising-related sites are the seventh most popular request type, representing roughly 2% of all requests. Traffic generated by such requests is equivalent to wasted bandwidth as most advertisements are targeted at more affluent urban consumers and are likely of no interest to users in rural Zambia. As bandwidth is clearly a scarce resource in this network, such wasteful access to advertisements can lead to further deterioration of the user experience.



(a) Percent of flows (b) Number of flows (c) Aggregate bytes

Figure 9. TCP flow success and failure for URIs of interest.

Following our URI findings we evaluate TCP flow patterns associated with four of the most accessed online services: Facebook, Google, YouTube, and software updates. For this analysis we combine the previously extracted unidirectional flows into bidirectional sessions based on flow signature and timestamp. We then extract flows of interest based on the URIs accessed in the corresponding session. Figure 9 plots (a) the percentage and (b) the number of flows as well as (c) the aggregate bytes over each period for the four services. The results are divided in terms of flow completion and failure. Both the number of flows and bytes for Facebook access increase over the three periods. This trend differs from that followed by Google, which in terms of number of flows remains almost unchanged Post-upgrade, but increases over the Long-term. Similar to Facebook, YouTube also increases immediately after the upgrade both in terms of flows and aggregate bytes. Interestingly, the failure rate of YouTube flows is high Post-upgrade, then it decreases. This may be due to software updates that use a large fraction of the bandwidth Post-upgrade, causing YouTube to fail more often. Of note, while only 16% of YouTube flows fail in the Long-term, they account for 40% of the YouTube flow bytes. This implies that large flows are the ones that fail most often. Due to the increased interest in access to real-time streaming services such as YouTube, the network quickly achieves its maximum capacity, inhibiting these services with substantial flow failures.

We note the rapid increase in the number of failed flows to Google in the Long-term. A closer look at these requests indicates that a substantial number of them are destined for the mobile version of Google Maps. Depending on which Google server those requests hit, a large fraction of the requests receive a 501 response (Not Implemented), which means the server does not recognize or cannot fulfill the request. Such errors often indicate unavailability of a certain feature at the server side. In Macha's case, a single user with an Android phone attempts to use Google Maps using outdated Google Maps servers, resulting in multiple retries and failures and causing the rapid increase in failed flows to Google.

Lastly, we look at TCP flows from software updates. The number of such flows decreases slightly Post-upgrade but increases in the Long-term. Our analysis indicates that the short-term decrease is due to improved network performance, resulting in fewer TCP session re-initializations. Furthermore, the quantity of bytes associated with software updates doubles immediately after the link upgrade. This is likely due to long-postponed software updates that can finally be completed. We see a decrease in software update bytes in the Long-term due to successful completion of updates in the period Post-upgrade.

Next, we measure HTTP response codes to uncover discernible differences between observation periods. We find noticeable changes in three response types: 200, 400, and 408. OK responses (type 200) indicate a valid request for which an HTTP server can correctly respond. As shown in Table 4, the percentage of HTTP 200 responses increases more than 10% after the link upgrade. Bad Request (type 400) errors indicate a request that the web server does not understand. These errors typically are caused by bad syntax or potentially a host infected with malware that sends poorly defined HTTP requests. Table 4 shows that type 400 errors decrease significantly after the link upgrade. We believe that this could be due to two changes. First, immediately after the upgrade, hosts could have implemented overdue software updates, which could rectify browser version issues associated with the request format. Second, in a fashion similar to operating system software, antivirus software is updated to newer versions, which potentially allows for malware detection of and removal from hosts. The final response code we investigate is 408 (Request Timeout), which indicates that the server expects a request from the client in some amount of time and the client fails to produce the request. Such errors occur in networks with very limited bandwidth or where multiple packets are dropped along the path. The number of 408 errors decreases dramatically after the link upgrade. This is an encouraging result, as it shows that even a small bandwidth increase can make a large difference in the user experience.

Table 4. HTTP Response Codes.

Response	Pre-upgrade	Post-upgrade	Long-term
200	4,289,578	3,333,240	4,667,380
400	5,933,008	2,627,842	3,514,872
408	17,146	68	162
Total	12,638,744	7,507,975	10,186,110

4.5 Flow Geography

To evaluate the geographic characteristics of website accesses, we investigate the network traffic using geographical information. For each traffic flow we identify the external node IP address. Using these IP addresses, we query the MaxMind GeoIP database[4] to correlate each flow with geographic coordinate information. Our preliminary investigation involves calculating the straight-line distance between Macha and the given coordinates for the other side of each connection using the Haversine formula (Robusto, 1975). Figure 10 shows the CDF of the flow distances from Macha in each of the three observation periods. We find that flows generally occur over longer distances in the periods after the network upgrade. Of note is the large increase between the Pre period and the Post period in the roughly 8,000–12,000 km range. While Long-Term flows show even longer distances compared to Post-upgrade, the increase is not as pronounced. We posit that a possible reason for the increase in distances from Macha is the result of a better user experience after the network upgrade, which encourages users to access such services that are physically farther away.

We also use the GeoIP database to find the country code for each external node. We calculate the number of bytes associated with each country code and rank them. Interestingly, traffic to and from nodes in Zambia itself increase dramatically after the network upgrade. In the Pre period, Zambia ranks as the 13th most popular country in terms of bytes, representing 0.9% of all traffic. In the Post period, Zambia jumps to second most popular country, representing 23.4% of all bytes; in Long-Term it ranks third, with 12.1%.

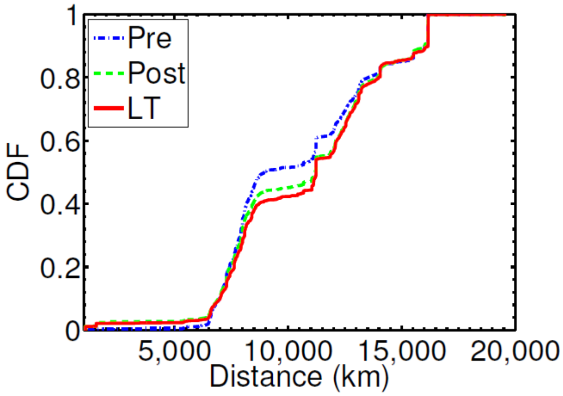


Figure 10. Flow distance.

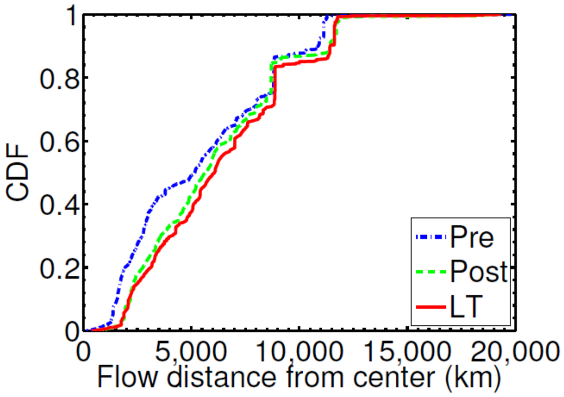


Figure 11. Flow distance from center-mass.

Our initial distance findings lead us to investigate not only the distance from Macha that flows represent, but also the overall "worldliness" of the network flows. That is to say, to further characterize network usage, we investigate the distribution of the geographic coordinates. We employ the radius of gyration metric to provide a value for the spread of the data. *Radius of*

[4] http://dev.maxmind.com/geoip/

gyration is used extensively to characterize user mobility in wireless networks (Gonzalez, 2008) and provides a technique to measure dispersion. Radius of gyration can be understood as the range of observed points up to time *t* and can be calculated by this formula:

$$r_g^a(t) = \sqrt{\frac{1}{n_c^a(t)} \sum_{i=1}^{n_c^a} (r_i^a - r_{cm}^a)^2}$$

where r_i^arepresents the *i*th coordinate for period *a*, and r_{cm}^a is the calculated center-mass for the period. $n_c^a(t)$ represents the number of points measured up to time *t*. Table 5 indicates that each successive period shows an increase in the radius compared to the prior periods. This means that not only are flows connecting to locations farther away from Macha as seen in Figure 10, they are also spreading out. Assuming users are behind most of the network traffic, we can argue that network users are connecting to content from more geographically diverse parts of the world.

Table 5. Measured Radius of Gyration.

Period	Distance (km)
Pre-upgrade	6,363.26
Post-upgrade	6,851.41
Long-term	7,096.86

We verify these results in two ways. First, we find the distances between each flow and the center-mass and plot the CDF as shown in Figure 11. As expected, we see longer-distance flows in the periods after the network upgrade. Further, the increases seem to be incremental and uniform, rather than drastic changes. We also investigate the center-mass values to determine whether the upstream provider (which changed with the network upgrade) drastically alters the distribution of external nodes. While we expect the center-mass values to be different for each period, we also expect them to be somewhat clustered. Should the upstream providers apply unexpected policy-based traffic routing [e.g., resolving all CDN (content delivery network) queries to a particular location], we expect to see the center-mass values vary dramatically between the Pre period and those after the upgrade. Figure 12 shows the center-mass points for each period as well as each calculated radius of gyration. We find the three center-mass values are within a reasonable range of each other, given the global scale. As such, we do not credit the upstream provider with the radius of gyration increase. Given these results we are confident that the increase in spread can be credited to an increased geographic diversity of external nodes.

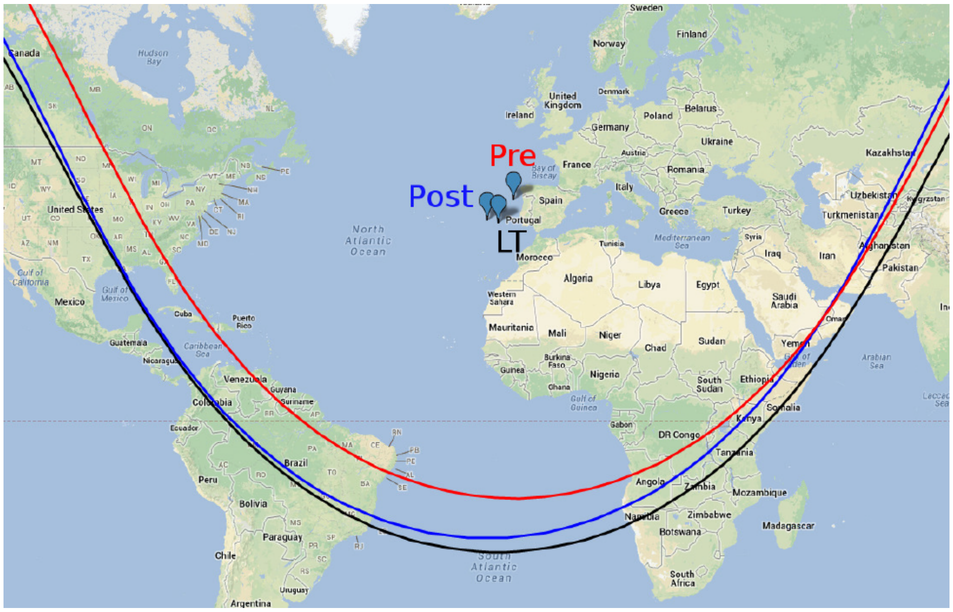


Figure 12. Center-mass points with radii of gyration.

4.6 Persistence of Long-Term Trends

To understand if the observed long-term trends persist over time, we examine network performance and usage in one additional time interval seven months post-upgrade. To assure that a change in user population would not skew our results, we focus on November, since this month has roughly the same number of users as the previous months of interest. We measure the same set of characteristics as for the previous three months and compare results from November and July (our Long-term month). Our results show that November's overall network usage resembles that of July's; however, user behavior changes. Users no longer attempt to upload large files to the Internet, but instead, increase their download activity. As a result, November's uploaded traffic decreases and is comparable to that before the upgrade, while the total downloaded bytes increases significantly, by 60%.

We analyze 2½ weeks of November. The reason we do not analyze the full month is twofold. First, the initial four days of the month are not captured due to a failure of the network management switch. Second, the Android phone phenomenon that we observe in July is aggravated throughout November, whereby approximately 4 million of 5 million requests sent to Google are produced by this phone. To conduct an objective comparison of network performance and usage, we discard the minute intervals throughout November where this phone is present on the network, as the phone's presence significantly skews the results. For the remainder of this section we present results from the 2½ weeks analyzed, referring to this traffic as the November traffic, or simply November.

In November, 158 Gb traverse the network, consisting of 373 million packets. This is about 1 Gb/day more than observed in July. This increase in the amount of data traversing the network is due to a combination of improved TCP performance and larger download attempts. The improved TCP performance is indicated by factors such as control overhead and bytes in flight. The control overhead in November decreases to 45.2% (compared to 49.7% in July), and the bytes in flight increases from 10.5 Kb to 36.4 Kb. We postulate that the large value of bytes in flight is due to longer TCP flows, which allows enough time for TCP to ramp up the number of bytes sent and achieve higher throughput. Lastly, we examine retransmissions as an indicator of TCP performance. We observe an increase of retransmissions from 1.16% in July to 3.68% in November. While this increase contradicts the TCP performance improvement, a closer look at retransmissions over time indicates several short periods when the network experiences poor performance, resulting in several flows with millions of retransmissions. By omitting those flows, the retransmission performance is comparable to July's flows.

We now examine uplink and downlink trends to assess whether user experience in sharing and accessing content persists throughout November. The number of uplink bytes in the November traffic is 11.7 Gb, or approximately 650 Mb/day. This is a decrease in the uploaded traffic compared to July, where uploads are 1.3 Gb/day. In fact, November's uploaded bytes/day is comparable to those before the upgrade (620 Mb/day). At the same time, the amount of downloaded traffic increases from 4.8 Gb/day in July to 6.3 Gb/day in November, indicating a clear trend of more Internet traffic consumption than generation.

Table 6. TCP Flow Success and Failure in Uplink and Downlink.

	Percent of flows		**Aggregate Gb/day**		**Average flow size, Kb**	
	July	**November**	**July**	**November**	**July**	**November**
Failed downloads	38	42	2.33	2.08	12	6.2
Successful downloads	62	58	2.5	4.21	6.5	9
Failed uploads	19	21	0.633	0.147	7	0.9
Successful uploads	81	79	0.667	0.504	1.5	0.8

For a better understanding of how success and failure of uploads and downloads influence users' inclination to share and consume content, we separate the November traffic into completed and failed TCP flows and analyze these flows in the uplink and downlink directions. As Table 6 shows, there is a slight increase in the percentage of successful uplink and downlink flows. At the same time, the traffic volume measured in aggregate bytes changes more substantially in both directions. The aggregate bytes in failed uploads decrease by 77%, reflecting users' decreasing interest in uploading online content. Similarly, the aggregate bytes of successful uploads decrease from 667 Mb to 504 Mb/day. In line with this trend, the average flow size of both completed and failed uplink flows decreases in November. The aggregate of successfully downloaded bytes increases by 60% compared to July, indicating a rapid increase in the amount of consumed traffic.

To better understand this shift in usage characterized by rapid decrease in uploads and growth in downloads, we examine November's most popular layer-7 services. We observe a notable change in usage compared to July, whereby P2P traffic decreases and HTTP increases. The fraction of November's P2P bytes drops to 9% compared to 25% in July. This decrease is shared between uploads and downloads, but is more pronounced in the upload direction: P2P bytes drop from 52% to 7%. Because the P2P uploads constitute a large portion of the overall uploads in the network (51%) in July, their decrease has a significant impact on the overall upload volume observed in November. Simultaneously, the percentage of HTTP downloads increases from 63% to 78%.

The goal of the analysis presented in this section is to evaluate the persistence of network performance and usage in the post-upgrade period. While network performance improves compared to July, the use of network services changes. In particular, we observe a rapid decrease in the volume of upload flows, while more emphasis is put into downloads. We observe a shift of usage from HTTP immediately after the upgrade to P2P file transfers in July and back to HTTP in November, indicating that users are eager to try new services but ultimately revert to HTTP, likely due to poor user experience.

4.7 Benchmark

In this section we answer this question: Given the limited gateway capacity in Macha, how well can the network perform? To this end, we provide a benchmark of network performance in November and compare this benchmark performance with the actual observed network performance. In particular, we analyze parts of the trace when fewer users access the network in order to ensure minimal contention on the link. Our analysis focuses on TCP performance and shows that the benchmark traffic performs significantly better than November's average performance.

To generate the benchmark trace we divide the November capture into one-minute chunks and consider the minutes with three or fewer users. This method generates 2,853 minutes (nearly 48 hours), which we use as our benchmark. The website access distribution during those hours is comparable to that during busy periods. The benchmark traffic amounts to 5.88 Gb. Of 9,556,817 packets, 16% were control packets, a significant reduction in the amount of protocol overhead compared to the more general scenario. The average RTT decreases from 157 ms to 119 ms. Because the same websites are accessed, this likely indicates less queuing delay as the ISP gateway and/or core network is less overloaded.

With high-delay, low-speed connections, rural users are often limited in their ability to use the Internet. Furthermore, the benchmark presented in this section clearly shows that when multiple users share a limited Internet connection, the user's experience can be further deteriorated. In less overloaded scenarios, however, the TCP performance is comparable to that of the Western world, whereby the RTT declines and the amount of control overhead drops significantly. Thus, if a single 2 Mbps link such as the one in Macha is employed by several users, there is a hope that the Internet experience can be on par with that in the Western world. When a single 2 Mbps connection is shared among tens or hundreds of users, however, the network performance continues to negatively impact the user experience.

5. Next Steps

Our analysis of a network upgrade in a rural community indicates that even a small increase in access bandwidth can improve network usability. For example, successful software updates and updated antivirus protection immediately after the upgrade grant better performance in HTTP request generation and, overall, decrease the traffic due to malware activity, resulting in the possibility for better performance. While these results are encouraging, incremental increase of available bandwidth often brings only a marginal improvement in user experience, as indicated by the large volume of failed requests in Macha's case. In the face of such increased usability but still a poor-quality user experience, the need for systems such as VillageShare (Johnson et al., 2012) and others (Du et al., 2006; Isaacman & Martonosi, 2011; Vithanage & Atukorale, 2011) that can intelligently manage network activities is even more pronounced.

One immediate need arising from our analysis is the prioritization bandwidth allocation to critical services. For example, as usage patterns in Macha did not immediately change Post-upgrade, critical software updates were finally able to complete. This, in turn, resulted in a rapid improvement in browsing experience (as indicated by the drop in HTTP Bad Request and HTTP Request Timeout messages) on one hand and by the decrease of traffic associated with malware on the other. This observation hints at a need for a system able to detect critical services and allocate higher bandwidth for such services.

Such a system would be able to detect network traffic anomalies (for example, increase in abnormal HTTP requests or traffic to ports associated with viruses) and prioritize bandwidth assignment for software updates. Two major concerns arise with regard to such a system. First, to ensure that such bandwidth prioritization does not compromise the user experience, this functionality can be embedded in time-shifted proxies such as discussed by Du et al. (2006), Johnson et al. (2012), and Vithanage and Atukorale (2011). Time shifting to off-peak hours, however, runs the risk that users would turn off their computers, bringing us to the second challenge in such a system design. To handle offline computers, local caching techniques (Isaacman & Martonosi, 2011; Johnson et al., 2012; Vithanage & Atukorale, 2011) can be employed, making particular content (e.g., software updates) available in the local network for use during peak hours.

6. Discussion and Conclusion

We investigated a unique dataset from a rural sub-Saharan village that captured usage before and after an Internet access speed upgrade. We studied the effects of this upgrade on network performance and user behavior. We found that performance improved immediately after the upgrade, whereby automatic services that were previously failing due to slow access speed were finally able to complete. With improved network performance, subscribers attempted more bandwidth-demanding services such as YouTube video streaming. There also was a substantial increase in attempts to share online content, whereby the uplink byte volume doubled in the Long-term. Unfortunately, with the increase of upload attempts, the failure rate of uploads grew as well, resulting in a drastic decrease in the number of uploaded bytes. While our data did not allow further analysis of the reasons for such upload decreases, one likely reason was that users became discouraged from attempting uploads, a trend observed in previous research (Johnson et al., 2012; Vannini & le Crosnier, 2012). Part of the decrease in uploads could have been caused by BitTorrent users who learned to disable seeding when using torrent trackers. As discussed in Section 4.3, however, only 28% of the uploads were due to BitTorrent. This means that more than 900 Mb/day in July were intentional uploads, which is larger than all the uploads attempted in November (i.e., 650 Mb/day). The latter indicates that November's intentional uploads were substantially fewer than July's, suggesting that users deliberately stopped uploading files to the Internet.

An Internet access upgrade in the context of developing rural regions is not a trivial task. Although such upgrades are perceived to lead to improved performance and user experience, this is not always the case for communities that are largely bandwidth impaired. In such communities, an Internet upgrade can be only a small increment to the more substantial access speed needed to accommodate modern web content and applications. Each such increment gives users the ability to more fully use the modern Internet with bandwidth-intensive applications; however, it is clear that in some developing regions, even an eight-fold increase in network capacity is insufficient. Many rural communities such as Macha have a long way to go before their Internet experience parallels that of users in the Western world. In bandwidth-deprived communities there will always exist the closed loop of *bandwidth increase* followed by adoption of new *bandwidth-intensive services*, which may lead to *deteriorated performance and user experience*. The question to ask, then, is: How much bandwidth is enough for this cycle to break? Or, What can be done to make the available bandwidth sufficient?

Acknowledgments

This work was funded through NSF Network Science and Engineering (NetSE) Award CNS-1064821. We are thankful to our sponsors for enabling this work. We are also thankful to our partners in Macha and at LinkNet for their cooperation.

Mariya Zheleva, Visiting Assistant Professor, University at Albany SUNY.
mzheleva@albany.edu
Paul Schmitt, PhD student, University of California, Santa Barbara.
pschmitt@cs.ucsb.edu
Morgan Vigil. PhD student, University of California, Santa Barbara.
mvigil@cs.ucsb.edu
Elizabeth Belding, Professor, University of California, Santa Barbara.
ebelding@cs.ucsb.edu

References

Abdeljaouad, I., Rachidi, H., Fernandes, S., & Karmouch, A. (2010, May). *Performance analysis of modern TCP variants: A comparison of Cubic, Compound and New Reno*. Paper presented at the 25th Biennial Symposium on Communications, Kingston, ON. doi:10.1109/BSC.2010.5472999

Allagui, I., & Kuebler, J. (2011). The Arab Spring and the role of ICTs. *International Journal of Communication, 5,* 1435–1442. Retrieved from http://ijoc.org/index.php/ijoc/article/viewFile/1392/616

Anokwa, Y., Dixon, C., Borriello, G., & Parikh, T. (2008, September). *Optimizing high latency links in the developing world*. Paper presented at the ACM Wireless Networks and Systems for Developing Regions Workshop, San Francisco, CA. doi:10.1145/1410064.1410076

Du, B., Demmer, M., & Brewer, E. (2006, May). *Analysis of WWW traffic in Cambodia and Ghana*. Paper presented at the 15th International World Wide Web Conference, Edinburgh, Scotland. doi:10.1145/1135777.1135894

Fraser, H. S. F., & McGrath, S. J. D. (2000). Information technology and telemedicine in sub-Saharan Africa. *BMJ, 321*, 465–466. doi:http://dx.doi.org/10.1136/bmj.321.7259.465

Gonzalez, M. C., Hidalgo, C. A., & Barabasi, A.-L. (2008). Understanding individual human mobility patterns. *Nature, 453*(7196), 779–782. doi:10.1038/nature06958

Ihm, S., Park, K., & Pai, V. S. (2010, June). *Towards understanding developing world traffic*. Paper presented at the 4th ACM Workshop on Networked Systems for Developing Regions, San Francisco, CA. doi:10.1145/1836001.1836009

International Telecommunication Union (ITU). (2013). *The world in 2013: Facts and figures*. Retrieved from http://www.itu.int/en/ITU-D/Statistics/Documents/facts/ICTFactsFigures2013-e.pdf

Isaacman, S., & Martonosi, M. (2011, March–April). *Low-infrastructure methods to improve Internet access for mobile users in emerging regions*. Paper presented at the 20th International World Wide Web Conference, Hyderabad, India. doi:10.1145/1963192.1963361

Johnson, D. L., Belding, E. M., Almeroth, K., & van Stam, G. (2010, June). *Internet usage and performance analysis of a rural wireless network in Macha, Zambia*. Paper presented at the 4th ACM Workshop on Networked Systems for Developing Regions, San Francisco, CA. doi:10.1145/1836001.1836008

Johnson, D. L., Pejovic, V., Belding, E. M., & van Stam, G. (2011, March–April). *Traffic characterization and Internet usage in rural Africa*. Paper presented at the 20th International World Wide Web Conference, Hyderabad, India. doi:10.1145/1963192.1963363

Johnson, D. L., Pejovic, V., Belding, E. M., & van Stam, G. (2012, March). *VillageShare: Facilitating content generation and sharing in rural networks*. Paper presented at the 2nd ACM Symposium on Computing for Development, Atlanta, GA. doi:10.1145/2160601.2160611

King, A. B. (2012). *Web site optimization*. Retrieved from http://www.websiteoptimization.com/speed/tweak/average-web-page

Matthee, K. W., Mweemba, G., Pais, A. V., van Stam, G., & Rijken, M. (2007, December). *Bringing Internet connectivity to rural Zambia using a collaborative approach*. Paper presented at the International Conference on Information and Communication Technologies and Development, Bangalore, India. doi:10.1109/ICTD.2007.4937391

Ndou, V. (2004). E-government for developing countries: Opportunities and challenges. *Electronic Journal of Information Systems in Developing Countries, 18,* 1–24. Retrieved from http://www.ejisdc.org/ojs2/index.php/ejisdc/article/view/110

Robusto, C. C. (1957). The cosine-haversine formula. *The American Mathematical Monthly, 64*(1), 38–40. Retrieved from http://www.jstor.org/discover/10.2307/2309088?uid=3739832&uid=2134&uid=2&uid=70&uid=4&uid=3739256&sid=21105430901623

Sife, A. S., Lwoga, E., & Sanga, C. (2007). New technologies for teaching and learning: Challenges for higher learning institutions in developing countries. *International Journal of Education and Development Using ICT, 3*(2), 57–67. Retrieved from http://ijedict.dec.uwi.edu/viewarticle.php?id=246

Sinha, R., Papadopoulos, C., & Heidemann, J. (2005, May). *Internet packet size distributions: Some observations* (Technical Report ISI-TR-2007-643). Los Angeles, CA: USC/Information Sciences Institute. Retrieved from http://www.isi.edu/~johnh/PAPERS/Sinha07a/

Song, S. (2014, November). African undersea cables [Web log post]. Retrieved from http://manypossibilities.net/african-undersea-cables

Surana, S., Patra, R., Nedevschi, S., Ramos, M., Subramanian, L., Ben-David, Y., & Brewer, E. (2008, April). *Beyond pilots: Keeping rural wireless networks alive*. Paper presented at the 5th USENIX Symposium on Networked Systems Design and Implementation, San Francisco, CA. Retrieved from http://tier.cs.berkeley.edu/docs/wireless/nsdi-surana.pdf

van Hoorik, P., & Mweetwa, F. (2008, May). *Use of Internet in rural areas of Zambia*. Paper presented at IST Africa, Windhoek, Namibia. Retrieved from http://publications.tno.nl/publication/104213/OnyGJD/hoorik-2008-use.pdf

Vannini, L., & le Crosnier, H. (2012). *NET.LANG: Towards the multilingual cyberspace*. Caen, France: C&F éditions.

Vithanage, W. W., & Atukorale, A. S. (2011, June–July). *Bassa: A time shifted web caching system for developing regions*. Paper presented at the 5th ACM Workshop on Networked Systems for Developing Regions, Bethesda, MD. doi:10.1145/1999927.1999944