The Increased Bandwidth Fallacy: Performance and Usage in Rural Zambia

Mariya Zheleva, Paul Schmitt, Morgan Vigil and Elizabeth Belding
Department of Computer Science
University of California, Santa Barbara
{*mariya, pschmitt, mvigil, ebelding*}*@cs.ucsb.edu*

ABSTRACT

Broadband Internet access has become a critical part of socio-economic prosperity; however, only 6 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 links, if available at all. As a result, the quality of the Internet access is often poor and at times unusable. In this paper we study the performance and usage implications of an Internet access upgrade, from a 256kbps satellite link to a 2Mbps terrestrial wireless link in rural Zambia. While usage did not immediately change, performance improved soon after the upgrade. By three months post-upgrade, however, subscribers began to use the faster connection for more bandwidth-hungry applications such as video-streaming and content upload. This change in usage resulted in dramatic deterioration of network performance, whereby the average round trip time doubled, the amount of bytes associated with failed uploads increased by 222% and that of failed downloads by 91%. Thus, while an Internet access upgrade should translate to improved performance and user experience, in rural environments with limited access speed and growing demand, it can bring unexpected consequences.

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 from democratic changes [4], to government [19], e-learning [21] and health care [7]. Broadband Internet access, however, is still largely unavailable in developing countries with only 6% of the population having broadband connectivity [11], the majority of which is in urban areas.

Recent efforts to bring connectivity to rural areas of the developing world utilize asymmetric satellite or other low-bandwidth wireless links [17, 23]. 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 [16]. As a result, residents of developing rural regions access the web with inadequate conectivity for the bandwidth requirements of modern content. These opposing trends in content growth and limited capacity render Internet access frustrating or even unusable [6, 14] in many developing areas.

Previous work on traffic analysis shows a "strong feedback loop between network performance and user behavior" [14], whereby residents in bandwidth-constrained environments tend to focus more on bandwidth-light applications such as web-browsing, as opposed to those in bandwidth-rich environments which enable multimedia streaming, content upload, and real-time user interaction. In the face of limited bandwidth, the failure rate of uploads is high [15], discouraging rural residents from contributing to the Internet content and resulting in consumption of largely Western content [26]. Thus, while recognizing the potential 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 [25].

The focus of our work is in Africa, where the increased fiber-optic capacity [1], coupled with higher-bandwidth, lower-latency technologies such as terrestrial microwave wireless gives hope for improved Internet access in remote areas. In this paper we study the implications of an Internet access upgrade from a satellite to a microwave terrestrial link on the performance and Internet usage in the rural community of Macha, Zambia. To the best of our knowledge, this is the first real-world comparative study of pre- and post-upgrade Internet usage and performance. As such, our dataset offers a unique opportunity to study the change in user behavior and Internet usage following an eight-fold increase in access bandwidth. We evaluate a total of three months of usage: one month before the upgrade, one month after the upgrade and one month three months later. Our results show that while usage did not change immediately, application performance improved. However, as time progressed subscribers began to change their Internet usage behavior, which ultimately resulted in network performance degradation and subsequent deterioration of user experience. The Internet access upgrade broadened users' abilities to access content, use online applications, and express themselves on the Internet. At the same time our results make a strong case

that one should not assume that advanced technologies and higher access speed grant better experience and increased adoption of the Internet in rural communities; rather one should carefully consider the evolution of usage and performance in order to assess the actual impact and adoption of Internet technologies.

2. MACHA

Economics and Demographics. Macha is a typical poor rural village located in the Southern province of Zambia. A total of 135,000 people live in the area, spread over a large radius of 35 km with average population density of 25 persons/km^2. The primary occupation in the village is maize farming. The average estimated income is \$1/person/day – 5 times less than the round-trip cost to the closest town and 30 times less than a monthly Internet subscription limited to 1GB.

Macha has been a local leader in health-care and technological innovation. Active organizations in the village 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 6 km^2, including schools, the hospital, the research institute and residential areas. LITA (LinkNet Information Technology Academy), an IT school affiliated with MachaWorks, teaches basic computer skills to local residents.

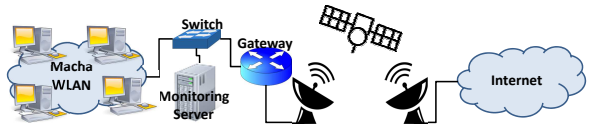
Internet usage and provisioning. While Macha is connected to the national power grid, electricity is rarely available in individual households. The lack of electricity coupled with the high prices for user equipment and Internet provisioning makes it virtually impossible for Machans to use Internet at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

Internet access is distributed from the village Internet gateway to central facilities via a local wireless mesh network (Fig. 1(a)) maintained by LinkNet. We refer the interested reader to [17] for more details about the wireless mesh in Macha. Between 2008 and April 2011, the village was connected to the Internet through a satellite connection that cost \$1200/month and provided 256kbps downlink bursting to 1Mbps, and 64kbps uplink bursting to 256kbps. In April 2011, the village Internet access was upgraded to a higher quality microwave terrestrial link (Fig. 1(b)) with speeds up to 2Mbps costing \$3600/month. At the time of the Internet link upgrade, approximately 300 residents were regular users of the Internet connectivity.

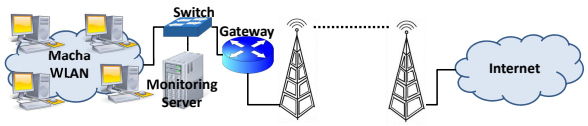
3. NETWORK ANALYSIS

We evaluate the network performance and usage for three months. We select one month immediately before (which we call Pre-upgrade) and one month immediately after the upgrade (Post-upgrade) to measure the short term impact on the network usage and performance. We also evaluate one month of traffic approximately three months Post-upgrade to determine whether performance changed as time progressed. We call this time period Long-term.

We start by describing our traffic collection methodology as well as our approach to calculating evaluation metrics. We



(a) Network architecture before the upgrade



(b) Network architecture after the upgrade

Figure 1: Network architecture and traffic monitoring.

then continue with detailed results from our network analysis. We first focus on overall network performance analysis, which indicates that the majority of traffic traversing the network is TCP (93%). We, thus, focus our analysis on TCP performance following the increased bandwidth. We describe trends in uplink and downlink performance of TCP flows, and we identify the most popular applications based on TCP port number. We then assess the success and failure rate of TCP flows. We conclude our TCP analysis by outlining performance trends in Windows and Linux machines. We then switch to evaluation of network usage focusing on popular URIs. We conclude by analyzing the "worldliness" of network flows initated in Macha in an effort to determine whether Machans started using more global services once they had better Internet access.

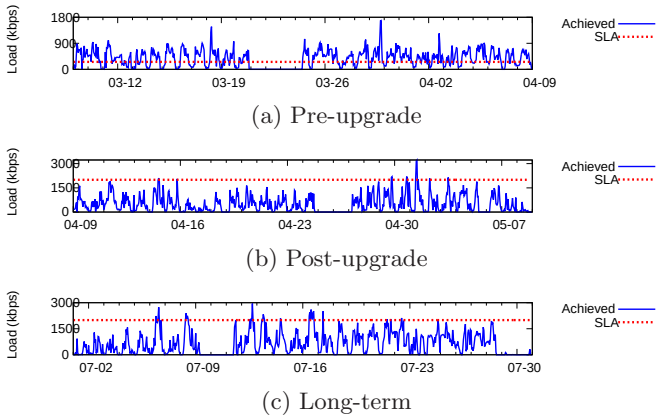
Pre-upgrade the network was typically saturated, resulting in high round trip time (RTT), congestion, and aborted sessions. Post-upgrade, we saw a decrease in the number of retransmissions and RTT due to improved network performance and movement away from the saturation point. By three months after the upgrade, the traffic had increased once again to saturation. Our analysis shows a difference in network performance and utilization Post-upgrade and Long-term: while Post-upgrade user behavior did not change, automatic programs, such as software updates, were suddenly able to complete, which resulted in an increase in traffic demand. In Long-term, subscribers utilized 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 usage and performance patterns in detail in the following sections.

3.1 Methodology

We capture traffic at the Internet gateway in Macha. As shown in Fig. 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, which allows us to capture all the traffic traversing the WLAN. With user consent, we capture packet headers and store traces on the monitoring server. During our last trip to Macha in Summer 2012, we offloaded the collected traces to an external hard drive and brought them to our research facility for offline analysis.

Table 1: General TCP statistics averaged over each time period.

	Total GB	Total packets (x10^6)	Total control packets (%)	Average Window (kB)	Average RTT (s)	Total retransmissions (%)
Pre	123	373	56.59	38	0.1436	1.12
Post	163	338	47.69	52	0.1085	1.09
LT	210	432	49.72	62	0.3190	1.16



(a) Pre-upgrade

(b) Post-upgrade

(c) Long-term

Figure 2: Traffic load over time.

We now describe our methodology for extracting metrics from the collected network traces. In our evaluation we use metrics such as TCP window size, RTT, TTL and retransmissions. We extract these metrics by running `tshark` in an offline mode on the collected traces. For our flow analysis 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. In order to obtain bidirectional flows, we then combine the unidirectional flows based on flow signature and timestamp.

3.2 Overall network performance

Traffic load. We start with evaluation of the traffic load. We calculate the 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 rate. 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*. Fig. 2 plots over time the traffic load averaged per hour in blue and the Service Level Agreement (SLA) with the Internet provider in red[1]. In the period before the upgrade, the demand frequently exceeded the SLA of 256kbps. This is less often the case for the period immediately after the upgrade, as users have not yet adapted to the increase in 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 usage patterns whereby users began to access more bandwidth-hungry applications once more bandwidth was available. The gaps in the plots correspond to time periods in which traffic captures were unavailable due to power or network outages.

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

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

A similar trend is observed in RTT. While immediately after the upgrade the average RTT decreased by about 35*ms*, it nearly tripled as time progressed. We explore the RTT dynamics in more detail in Fig. 3, which plots a 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 values of RTT for those two periods are lower than those observed Pre-upgrade. As we will see later in our analysis (in section 3.5), the long-tail distribution of RTT after the upgrade is due to changed browsing habits and tendency to use services that are physically further away (such as streaming video from servers abroad). We provide in-depth discussion of usage patterns in section 3.4 to validate our hypothesis.

We analyze payload size in Fig. 4. We see a clear bimodality [22] of payload size, which is due to the prevalence of either control packets with 0 bytes payload or data packets with payload of about 1500 bytes. Clearly, the percentage of large data packets Post-upgrade as well as Long-term increased. We also see an increase in the average TCP window size (Table 1), which allows more packets to be sent in the network before an acknowledgement is received. This increased TCP window size is critical to improved TCP performance as it translates to higher achieved throughput.

We next measure the overhead, focusing on the percentage of the total packets that are due to retransmissions and control packets (e.g. TCP control packets are ACK, SYN, FIN). The payload of control packets is zero bytes. As Table 1 and Fig. 4 indicate, the fraction of control packets decreased after the link upgrade from 56.59% to 47.69% and then slightly increased in Long-term to 49.72%. The number of retransmissions follows a similar trend. This overall decrease in control overhead can be attributed to improved network performance, which resulted in less protocol overhead from retransmissions and repeated acknowledgements, as well as fewer attempts to re-establish failed TCP sessions. The uptick in retransmissions and control packets over the Long-term can be attributed to decrease in performance due to the increase in offered load to the new saturation point.

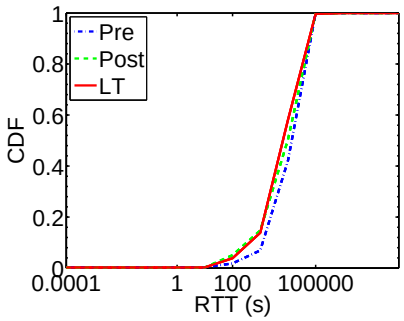


Figure 3: RTT.

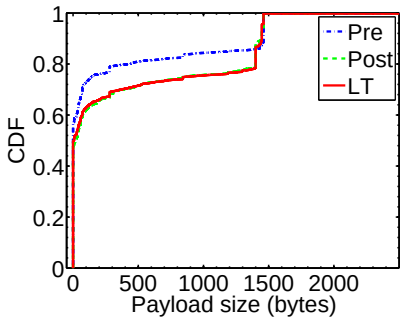


Figure 4: Payload size.

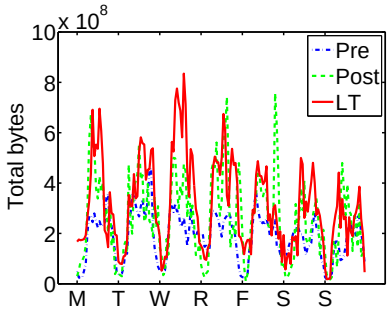


Figure 5: Bytes by day.

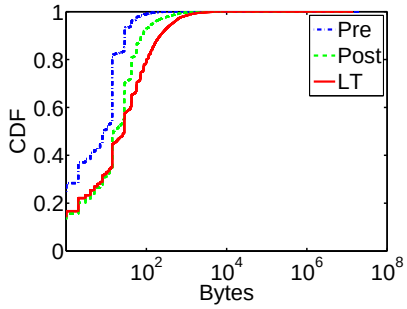


Figure 6: Bytes in flight.

Temporal trends. We now discuss performance trends over time. We evaluate byte count in Fig. 5, which plots the average on a weekly basis aggregated per hour. For example, the first data point of Fig. 5 presents an average over all occurrences of the first hour of Monday, for each of the one month time windows. As the figure shows, there is a clear diurnal pattern in link utilization. Furthermore, the amount of traffic generated during weekdays differs from that on weekends, with weekday traffic loads typically being heavier. The increase in traffic after the network upgrade is also observable in the figures.

3.3 TCP performance analysis

Our analysis shows that more than 93% of the traffic traversing the gateway in Macha is TCP. The performance of TCP improved significantly after the link upgrade. One factor indicative of this improvement is the *bytes in flight*, which is the fraction of sent data that has not yet been acknowledged. The bytes in flight is influenced by the TCP window size: the better the link performance, the larger the window size, which allows more data to be sent on the link before an acknowledgement is received. As indicated in Table 1, the TCP window size increased Post-upgrade and Long-term, allowing the amount of bytes in flight to ramp up. Fig. 6 presents a CDF of bytes in flight for the three periods. Immediately after the upgrade, the bytes in flight drastically increased and continued growing over the Long-term.

We continue our analysis by exploring TCP flow trends following this improved TCP performance. In order to extract uni-directional TCP flows from our `tshark` captures we develop a tool that examines packet signatures (sourceIP-sourcePORT-destinationIP-destinationPORT) and timestamp and groups them in flows accordingly. We start by presenting general trends of TCP flows in Table 2. The total 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.

Table 2: TCP flow analysis.

Period	Total GBytes	Flow size (B)	IAT (s)
Pre-upgrade	105	3445	1.92
Post-upgrade	145	7708	1.49
Long-term	183	8103	1.91

We next examine the average flow size across the three periods. As we can see in Table 2, the flow size doubled Post-upgrade and then continued increasing in Long-term. The increase of flow size can be attributed to different applications utilizing 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 as we explore in section 3.4. The average packet inter-arrival time (IAT) decreased Post-upgrade and then increased Long-term.

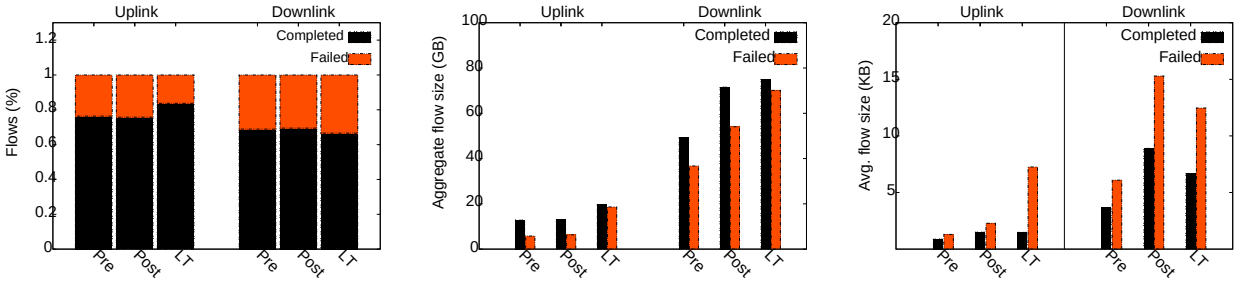
Uplink and downlink flows. Next we differentiate flows into uplink and downlink to analyze direction-specific trends. In Table 3 we first present aggregate bytes in each direction. Both uplink and downlink bytes increased after the link upgrade. While downlink increased rapidly, uplink remained almost unchanged Post-upgrade but then increased drastically over the Long-term. Average uplink packet size and flow size increased Post-upgrade and in Long-term. At the same time, downlink packet and flow sizes increased Post-upgrade and then slightly decreased over the 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, which 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 GBytes		# of Flows (x10^5)		Pkt size (B)		Flow size (B)	
	UL	DL	UL	DL	UL	DL	UL	DL
Pre	18.65	85.9	189	194	132.7	616.0	988.9	4427
Post	19.26	125.7	114	116	158.6	877.0	1691	10856
LT	38.14	145	157	168	227.7	787.6	2422	8613

Table 4: Top 10 most accessed TCP ports.

	Pre				Post				Long Term			
	Port	% Requests	% DL Bytes	Service	Port	% Requests	% DL Bytes	Service	Port	% Requests	% DL Bytes	Service
1	80	43.5	78.7	HTTP	80	42.3	83.5	HTTP	80	41.8	69.8	HTTP
2	443	27.5	10.8	HTTPS	443	23.8	5.6	HTTPS	443	26.1	6.7	HTTPS
3	5943	0.6	0.07	N/A	6346	0.7	0.04	Gnutella	25	0.7	0.04	SMTP
4	33033	0.3	0.03	Skype	6348	0.7	0.03	Gnutella	33033	0.7	0.04	Skype
5	25	0.3	0.03	SMTP	51413	0.5	0.1	BitTorrent	5943	0.3	0.02	N/A
6	12350	0.2	0.1	Skype	33033	0.4	0.01	Skype	12350	0.3	0.03	Skype
7	995	0.06	0.4	POP3	12350	0.2	0.07	Skype	445	0.1	0.001	Samba
8	13392	0.05	0.003	N/A	5943	0.1	0.008	N/A	51413	0.1	0.3	BitTorrent
9	8008	0.03	0.001	HTTP	6881	0.1	0.05	BitTorrent	4158	0.09	0.002	stat-cc
10	993	0.03	0.5	IMAP	45682	0.05	0.003	uTorrent	995	0.06	0.2	POP3



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

Figure 7: TCP flow success and failure in uplink and downlink direction.

Finally, we concentrate on the number of flows. As we can see in Table 3, the number of flows in both up- and downlink directions decreased dramatically Post-upgrade and then increased. The initial decrease can be attributed 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 flow failure rate as user demand again reaches link capacity.

Top 10 TCP ports. We analyze the top 10 applications based on TCP port popularity. Table 4 presents results for the top ten most accessed TCP ports in the uplink direction over the three periods along with information about the percentage of total downloaded traffic for these requests. As we can see, HTTP and HTTPS were persistently the most prevalent applications across the three periods. However, the other top eight ports vary over time. Pre-upgrade we see many requests associated with Skype and e-mail clients (SMTP, IMAP, POP3). We also see multiple requests to port 8008, which is commonly used by Trojans. Post-upgrade security vulnerabilities, such as the one associated with port 8008, disappeared as computers were more successful in downloading and installing critical security updates. In the rest of the ports we see prevalence of Skype and P2P (mostly torrent) networks. As time progressed, the prevalence of torrent decreased and was replaced by e-mail client ports as well as some new malware associated with port 445 and 4158, which are often used for DoS attacks and remote unauthorized access of hard disks.

In terms of fraction of download bytes, we see that HTTP and HTTPS are very high PRE and POST upgrade. Interestingly, the amount of HTTP bytes decreased by 14% over the long term. This difference in bandwidth consumption was distributed among other applications such as HTTPS, RTMP used by Adobe for streaming audio and video through Flash Player, and Torrent applications.

TCP flows success and failure. We now focus on flow completion and failure. In compliance with RFC 793 that 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 a RST (reset) packet indicates a failed flow. Fig. 7(a) presents the fraction of completed and failed flows in uplink and downlink in each period. The completion rate of uplink flows remained unchanged Post-upgrade and then slightly increased in Long-term. On the other hand, the downlink flow completion rate remained unchanged. In Fig. 7 we also analyze success and failure trends correlated with byte volume and flow size. Fig. 7(b) plots the aggregate flow size in bytes for each direction. The aggregate size of both completed and failed uplink flows remained the same Post-upgrade and then increased in the Long-term. Unfor-

tunately, the amount of bytes in failed flows approaches the amount of bytes in completed flows, which indicates that, while users were likely more successful in uploading content, over the Long-term half of the total content that users generated failed to upload. Similarly, in terms of total size of downlink flows, we see a gradual increase in successful downloads; however, over the Long-term the aggregate size of download flows that failed also increased, nearly reaching the aggregate size of successful downloads.

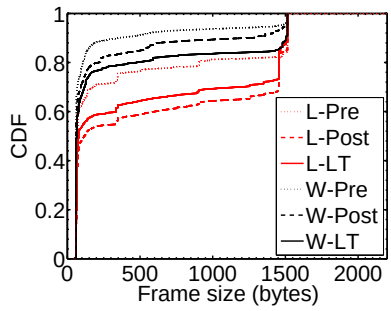
We evaluate average flow size of completed and failed flows in Fig. 7(c). We find the size of an individual flow by summing the packet sizes of all packets associated with the flow. In the uplink direction, the average size of failed flows over the Long-term is four times larger than the size of completed 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 e-mail 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.

Windows vs. Linux. Lastly, we evaluate the TCP performance of two of the most prevalent operating systems used in Macha: Windows and Linux. Using the observed TTL values, we were able to distinguish between the two operating systems [24]. Linux implements CUBIC TCP, which has optimized congestion control mechanisms for high bandwidth networks with high latency. This optimization occurs by calculating the window size according to the last congestion event. In this way, CUBIC TCP measures congestion independently from long RTTs [9]. This differs from Windows, which implements TCP Reno in Windows XP and Compound TCP in Windows Vista and subsequent Windows versions. TCP Reno and Compound TCP base window size on the RTT – window size increases with low RTT values and decreases with high RTT values. This method of congestion calculation causes Windows machines to interpret network latency as network congestion [18].

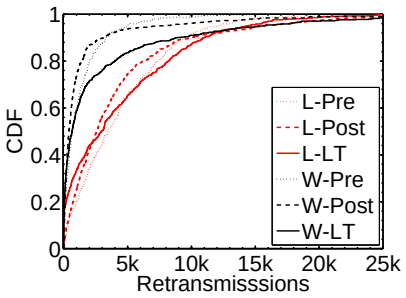
In Fig. 8 we plot frame size and retransmissions per hour for Windows (W) and Linux (L) machines. As we can see, Linux maintains higher mean frame size over all three periods. Thus, it is much more aggressive in pushing data onto the link. Naturally, this results in more retransmissions per hour in comparison with Windows. Linux's aggressive behavior, however, leads to higher achieved throughput of 487.6 Kbps in comparison with Windows, which only achieves 106.2 Kbps in the Long-term.

3.4 Network usage

The most prevalent application protocol used in Macha is web-browsing. 86.54% of the Pre-upgrade traffic in up- and downlink direction was a combination of HTTP and HTTPS. This number remained almost unchanged Post-upgrade – 85.85%, and dropped to 67.63% in the Long-term. At the same time, traffic categorized as Other, which includes services to unspecified ports (e.g. Skype and BitTorrent) increased in the Long-term. This is a



(a) Frame Size



(b) Number of Retransmits

Figure 8: Comparison of TCP performance in Windows and Linux.

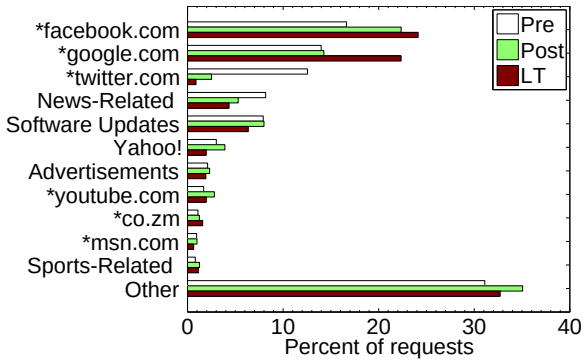
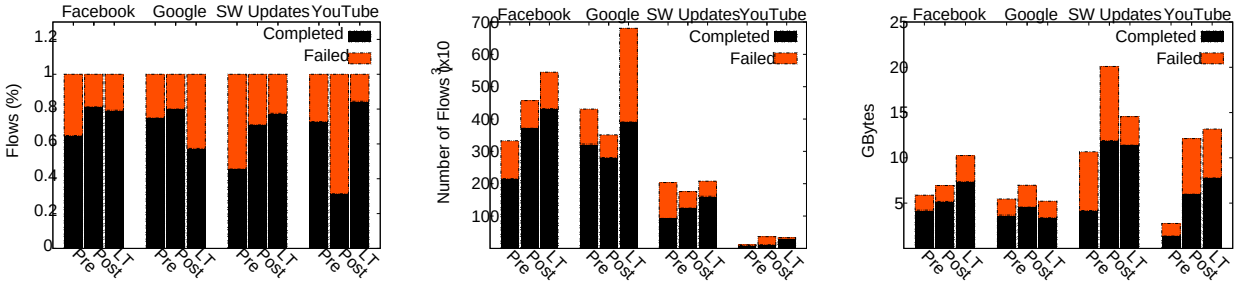


Figure 9: Popular URI Requests.

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 about popular applications with network performance and make inferences about the user experience based on this correlation.

Popular URIs. We begin our analysis by evaluating popular web services. Fig. 9 shows web URI requests classified by the destination domains and includes the top 14 requested sites. 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



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

Figure 10: TCP flow success and failure for URIs of interest.

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 [13].

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 increased Post-upgrade. While requests for multimedia-rich sites or large binary downloads remained the same across periods, the actual traffic associated with such requests increased as more requests were successfully completed. We explore TCP sessions count and size later in this section.

Advertisement-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 user experience.

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 uni-directional flows into bi-directional sessions based on flow signature and timestamp. We then extract flows of interest based on the URIs that have been accessed in the corresponding session. Fig. 10 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 and total bytes associated with Facebook flows increased over the three periods. This trend is different than the one followed by Google, which in terms of number of flows remained almost unchanged Post-upgrade, but increased over the Long-term. Similar to Facebook, YouTube also increased immediately after the upgrade both in terms of flows and aggregate bytes. Interestingly, the failure rate of YouTube flows was high Post-upgrade and then decreased. This might be due to software updates using a large fraction of the bandwidth Post-upgrade, which caused YouTube to fail more often. Of note, while only 16% of YouTube flows failed in the Long-term, those accounted for 40% of the YouTube flow bytes. This implies that large flows were most often the ones to fail. Due to the increased interest in access to real-time streaming services such as YouTube, the network quickly achieved its maximum capacity, inhibiting these services with substantial flow failures.

Lastly, we look at TCP flows from software updates. The number of such flows decreased slightly Post-upgrade and then increased 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 doubled immediately after the link upgrade. This is likely due to long-postponed software updates finally being able to complete. We see a decrease in software update bytes in the Long-term due to successful completion of updates in the period Post-upgrade.

Table 5: 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

Next, we measure HTTP response codes in an effort to find discernible differences between observation periods. We find noticeable changes in three response types: 200, 400, and 408. 200 (OK) responses indicate a valid request for which an HTTP server can correctly craft a response. As shown in Table 5 the percentage of HTTP 200 responses increases more than 10% after the link upgrade. 400 (Bad Request) 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. The table shows 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 request format. Secondly, in a similar fashion to operating system software, anti-virus software was updated to newer versions which could potentially allow for the detection and removal of malware on hosts. The final response code we investigate is 408 (Request Timeout) which indicates that the server was expecting a request from the client in some amount of time and the client failed

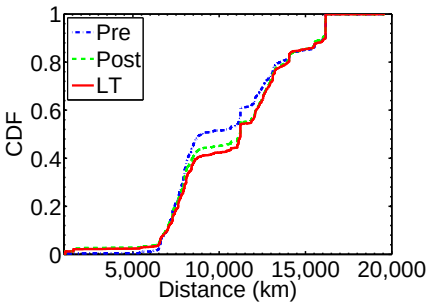


Figure 11: Flow Distance CDF.

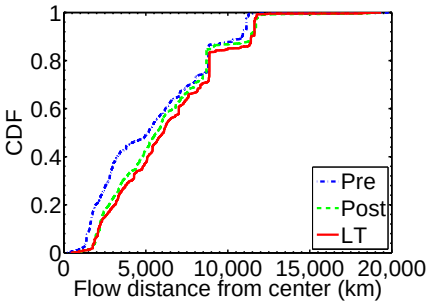


Figure 12: Flow distance from center mass CDF.

to produce such a request. Such errors occur in networks with very limited bandwidth or where multiple packets are dropped some point 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.

3.5 Flow Distance

We investigate the network traffic using geographical information to characterize usage. For each traffic flow we identify the external node IP address. Using these IP addresses, we query the MaxMind GeoIP database [2] 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 [20]. Figure 11 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 to 12,000 km range. While Long-Term flows show even longer distances as compared to Post-upgrade, the increase is not as pronounced. We posit that a potential reason for the increase in distances from Macha is the result of a better user-experience after the network upgrade, which encouraged users to access such services that are physically further 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 increased dramatically after the network upgrade. In the Pre period, Zambia ranks as the thirteenth most popular country in terms of bytes, representing 0.9% of all traffic. In the Post period, Zambia jumps up to rank second representing 23.4% of all bytes; in Long-Term it is ranked third with 12.1%.

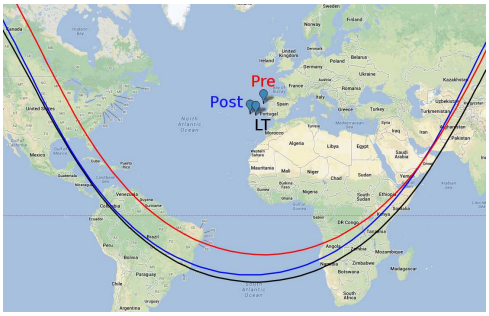


Figure 13: Center-mass points with radii of gyration.

Table 6: Measured radius of gyration.

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

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, we investigate the distribution of the geographic coordinates in order to further characterize network usage. We utilize the Radius of gyration metric to provide a value for the spread of the data. Radius of gyration has been used extensively to characterize user mobility in wireless networks [8] and provides a technique for measuring dispersion. It can be understood as the range of observed points up to time t and can be calculated by the formula:

$$r_g^a(t) = \sqrt{\frac{1}{n_c^a(t)} \sum_{i=1}^{n_c^a} (\vec{r}_i^a - \vec{r}_{cm}^a)^2}$$

Figure 14: Radius of gyration equation.

where $\vec{r}_i^a$ represents the i^{th} coordinate for period a and $\vec{r}_{cm}^a$ is the calculated center-mass for the period. $n_c^a(t)$ represents the number of points measured up to time t. Table 6 shows the results for the observation periods. We see 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 further away from Macha as seen in Figure 11, they are actually spreading out further as well. Assuming users are behind the majority of network traffic we can argue that network users are connecting to content from a larger geographic variety of the world.

We verify these results in two ways. First, we find the distances between each flow and the center-mass and plot the CDF shown in Figure 12. 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 along with the network upgrade) drastically altered 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 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 13 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 and 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.

4. RELATED WORK

Our work builds upon earlier analysis of Internet access in Macha [14]. This prior work focuses on network performance and usage during a two week measurement period in 2010. Other work analyzing rural Internet usage patterns includes [6] and [10]. Web traffic from Internet cafés and kiosks in Cambodia and Ghana is analyzed in [6]. Here the focus is on the characterization of HTTP traffic to guide caching techniques for web users in developing regions. Ihm et al. [10] focus on understanding the network traffic in developing regions as compared to their OECD counterparts. This paper characterizes national traffic patterns based on network usage, with the goal of improving caching techniques for developing regions. Anokwa et al. identify the impact of latency on network performance in developing regions and propose a flow-based prioritization scheme as a solution [5]. In contrast, our work focuses on a smaller scale and characterizes web traffic in order to ascertain the impact of a network upgrade on usage and performance.

Our analysis of TCP performance builds upon the measurements employed by Johnson et al. [14] by engaging a more in-depth analysis of TCP performance, including measurements of TCP windows, retransmissions, inter-packet arrival times, RTT, and packet sizes. Performance of CUBIC TCP, Compound TCP, and TCP Reno interactions is measured via simulation of high delay wireless networks in [3]. 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, and draw conclusions about TCP fairness of different variants found on the network.

5. NEXT STEPS

Our analysis of a network upgrade in a rural community indicates that even a small increase in access bandwidth can improve the usability of the network: for example, successful software updates and updated anti-virus protection immediately after the upgrade grant better performance in HTTP request generation and, overall, decrease the traffic due to malware activity, resulting in possibilities for better performance. While these results are encouraging, incremental increase of available bandwidth can often bring only marginal improvement of user experience, as indicated by the large volume of failed requests in the case of Macha. In the face of such increased usability but still low quality of user experience, the need of systems such as VillageShare [15] (and others [6, 12, 27]), that can intelligently manage activities in the network, is even more pronounced.

One immediate need that arises from our analysis is the one of prioritizing 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 rapid improvement in browsing experience (as indicated by the drop in HTTP Bad Request and HTTP Request Timeout messages in the network) on one hand and by the decrease of traffic associated with malware on another. This observation hints of a need for a system that can detect critical services and emulate a bandwidth increase for such services.

Such a system would be able to perform real-time detection of network traffic anomalies (for example increase in abnormal HTTP requests or traffic to ports associated with viruses) and would prioritize bandwidth assignment for software updates. Two major concerns arise with regards 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 [6, 15, 27]. Time shifting to off-peak hours, however, runs the risk of users turning off their computers, which brings us to the second challenge in such system design. To handle this, local caching techniques [12, 15, 27] can be employed which make particular content (e.g. software updates) available in the local network for use during peak hours.

6. DISCUSSION AND CONCLUSION

We utilize a unique dataset from a rural sub-Saharan village that captures usage before and after an Internet access speed upgrade. We study the effects of this upgrade on the network performance and user behavior. We find 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 were encouraged to use more bandwidth-demanding services such as YouTube video streaming. There also was a substantial increase in attempts for content sharing online, 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.

Another trend that stood out from our analysis is the stark difference in performance between Windows and Linux operating systems. As our results show, Linux outperforms Windows, achieving five times better throughput on average in the Long-term. This makes a case for careful selection of operating system and/or modifications of the network stack to facilitate better networking performance in bandwidth-constrained environments.

Internet access upgrade in the context of developing rural regions is not a trivial task. Although such upgrades are perceived to lead to overall better 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 just a small increment to the more sub-

stantial access speed that is needed to accomodate modern web content and applications. Each such increment gives users the ability to more fully utilize the modern Internet with bandwidth-intensive applications; however, it is clear that in the developing regions case, even an eight-fold increase in network capacity can be not nearly enough. Many rural communities like Macha have a long way to go before their Internet experience parallels that of users in the Western world.

7. ACKNOWLEDGEMENTS

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

8. REFERENCES

[1] African Undersea Cables, `http://manypossibilities.net/african-undersea-cables/`.

[2] GeoIP Products, MaxMind, `http://dev.maxmind.com/geoip/`.

[3] I. Abdeljaouad, H. Rachidi, S. Fernandes, and A. Karmouch. Performance analysis of modern TCP variants: A comparison of Cubic, Compound and New Reno. In *QBSC*, Kingston, ON, Canada, May, 2010.

[4] I. Allagui and J. Kuebler. The Arab Spring and the Role of ICTs. In *International Journal of Communication, Vol. 5*, pages 1435–1442, 2011.

[5] Y. Anokwa, C. Dixon, G. Borriello, and T. Parikh. Optimizing high latency links in the developing world. In *WiNS-DR*, San Francisco, CA, September, 2008.

[6] B. Du, M. Demmer, and E. Brewer. Analysis of WWW traffic in Cambodia and Ghana. In *WWW*, Edinburgh, Scotland, May 2006.

[7] H. S. F. Fraser and S. J. D. McGrath. Information technology and telemedicine in sub-Saharan Africa. *BMJ*, 321:465–466, August 2000.

[8] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, Jun 2008.

[9] S. Ha, I. Rhee, and L. Xu. CUBIC: a new TCP-friendly high-speed TCP variant. *SIGOPS Oper. Syst. Rev.*, 42:64–74, July 2008.

[10] S. Ihm, K. Park, and V. S. Pai. Towards understanding developing world traffic. In *NSDR*, San Francisco, CA, June, 2010.

[11] International Telecommunications Union. The World in 2013; Facts and Figures. `http://www.itu.int/en/ITU-D/Statistics/Documents/facts/ICTFactsFigures2013.pdf`, 2013.

[12] S. Isaacman and M. Martonosi. Low-infrastructure methods to improve internet access for mobile users in emerging regions. WWW '11, Hyderabad, India, 2011.

[13] D. L. Johnson, E. M. Belding, K. Almeroth, and G. van Stam. Internet usage and performance analysis of a rural wireless network in Macha, Zambia. In *NSDR*, San Francisco, CA, June, 2010.

[14] D. L. Johnson, V. Pejovic, E. M. Belding, and G. van Stam. Traffic characterization and internet usage in rural Africa. In *WWW*, Hyderabad, India, March, 2011.

[15] D. L. Johnson, V. Pejovic, E. M. Belding, and G. van Stam. VillageShare: Facilitating content generation and sharing in rural networks. In *ACM DEV*, Atlanta, GA, March, 2012.

[16] A. B. King. Web site optimization, `http://www.websiteoptimization.com/speed/tweak/average-web-page/`.

[17] K. W. Matthee, G. Mweemba, A. V. Pais, G. van Stam, and M. Rijken. Bringing Internet connectivity to rural Zambia using a collaborative approach. In *ICTD*, Bangalore, India, 2007.

[18] Microsoft Corporation. Microsoft Windows Server 2003 TCP/IP Implementation Details. `http://www.microsoft.com/en-us/download/details.aspx?id=13902`, 2007.

[19] V. Ndou. E-government for developing countries: opportunities and challenges. In *The Electronic Journal of Information Systems in Developing Countries, Vol 18*, pages 1–24, 2004.

[20] C. C. Robusto. The cosine-haversine formula. *The American Mathematical Monthly*, 64(1):38–40, 1957.

[21] A. S. Sife, E. Lwoga, and C. Sanga. New technologies for teaching and learning: Challenges for higher learning institutions in developing countries. In *International Journal of Education and Development using ICT, Vol. 3, Issue 2*, pages 57–67, 2007.

[22] R. Sinha, C. Papadopoulos, and J. Heidemann. Internet packet size distributions: Some observations. Technical Report ISI-TR-2007-643, USC/Information Sciences Institute, May 2007. Orignally released October 2005 as web page `http://netweb.usc.edu/~rsinha/pkt-sizes/`.

[23] S. Surana, R. Patra, S. Nedevschi, M. Ramos, L. Subramanian, Y. Ben-David, and E. Brewer. Beyond pilots: keeping rural wireless networks alive. In *NSDI*, San Francisco, CA, April 2008.

[24] Default TTL Values in TCP/IP, `http://www.map.meteoswiss.ch/map-doc/ftp-probleme.htm`.

[25] P. van Hoorik and F. Mweetwa. Use of Internet in rural areas of Zambia. In *IST Africa*, Windhoek, Namibia, May 2008.

[26] L. Vannini and H. le Crosnier. *NET.LANG: Towards the multilingual cyberspace*. C&F edition, March 2012.

[27] W. W. Vithanage and A. S. Atukorale. Bassa: a time shifted web caching system for developing regions. NSDR '11, Bethesda, Maryland, USA, 2011.