VillageShare: Facilitating content generation and sharing in rural networks

David L. Johnson, Veljko Pejovic,
Elizabeth M. Belding,
University of California, Santa Barbara
davidj,veljko,ebelding@cs.ucsb.edu

Gertjan van Stam
Linknet, Macha, Zambia
gertjan.vanstam@machaworks.org

ABSTRACT

While broadband Internet connectivity has reached a significant part of the world's population, those living in rural areas of the developing world suffer from poor Internet connectivity over slow long distance links, if they even have connectivity at all. While this has a general negative impact on Internet utilization, our social survey of users in the community of Macha, Zambia shows that the severest impact is in the area of content generation and sharing. To this end, our work describes VillageShare, an integrated time-delayed proxy server and content-sharing Facebook application. Through these two components, VillageShare facilitates localization of traffic, protecting the bandwidth-limited Internet link from content shared between local users, and minimizes upload abortions by time-shifting large uploads to periods when the gateway link is under-utilized. In this work we analyze traffic traces from Macha to discern opportunities for improvement of connection utilization, and then describe and evaluate the VillageShare architecture.

Categories and Subject Descriptors

C.2.2 [**Computer-Communications Networks**]: Network Protocols-*Applications*; C.4 [**Performance of Systems**]: Performance attributes

General Terms

Measurement, Performance

Keywords

Rural network, Social network, Traffic analysis, Developing regions, Localization

1. INTRODUCTION

Internet connectivity is a key driver of development of any nation [15]. Access to information influences numerous aspects of development, from education to health care and economy [18, 21, 1]. Through the Internet, communities with unique cultures and customs can establish a worldwide digital presence. Rapidly expanding social networks enable quick and easy idea dissemination and have already proved to be an invaluable tool for democratic change facilitation, such as during the 2011 Arab spring.

However, many challenges remain in rural developing areas which stand to gain the most from the positive social impact of the Internet. In rural areas, connectivity is often supplied using slow, costly asymmetric satellite links. In addition, the landscape of the web has changed drastically in the last decade, and user-generated, dynamic audio and video content has been integrated into static client-server content. The average web page size in 2011 is 48 times larger than the average size in 1995 [17]. As a result of these changes, bandwidth-poor rural area networks have experienced performance degradation leading to user frustration.

Existing solutions for rural areas concentrate on improving web browsing by enabling local caching and providing alternatives to the media-rich world wide web. These solutions only improve content consumption and are inappropriate in the context of Web 2.0 and social networking. Without a significant change in the way rural connectivity is managed, the web threatens to lose its egalitarian nature. User-generated content, such as Wikipedia articles, is virtually non-existent in local African languages[1]. Thus, on the web, rural Africans have to revert to a major language, e.g. English. Similarly, online social networks thrive due to abundance of user-shared content. Because of their centralized structure, OSNs require good outside connectivity, which is rarely supported by satellite-enabled rural area networks [23].

To identify major hurdles in content generation and sharing in rural Africa, we investigate Internet usage in the village of Macha, Zambia. Our investigation consists of network trace analysis and a social survey. We collected a two month network trace in early 2011 from which we extract information on upload behavior, online social network interaction and file sharing. In addition, we performed an online survey of Internet users in Macha with the aim of identifying key inhibitors of online content creation. Our survey points out that content generation depends on perceived Internet availability; our interviewees report restricted bandwidth as the main obstacle towards their greater Internet participa-

[1]For example, there are no Wikipedia articles in Chi-Tonga, Ila and Lozi languages spoken by more than two million people in (rural) Zambia, whereas approximately the same number of Slovenians enjoy a bounty of more than 100,000 articles in their native language.

DEV '12 March 11-12, Atlanta, GA

tion. When it comes to communication patterns, we find that a higher fraction of social network communication occurs between local users in the villages than between local and outside users. However, this communication between local users still utilizes the costly, bandwidth-limited gateway to route this traffic through cloud computing services on the Internet, significantly deteriorating user experience.

In rural communities, bandwidth is inherently limited by the Internet gateway link. In Macha, the local 802.11 network has an order of magnitude more capacity than the gateway. As a result, there is sound motivation for building solutions that attempt to redirect file sharing traffic between local users away from the Internet gateway such that the traffic can stay in the local network. Compounding the limited bandwidth, previous work has found that large objects sent or received during peak traffic times often suffer from repeated time-outs. This has a negative impact on shorter-lived flows in a bandwidth-constrained network by consuming bandwidth before they abort [12]. These network characteristics provide strong motivation for a solution that attempts to time-shift larger flows to low-usage periods. Such an approach will improve the efficiency of the bandwidth usage and the overall usability of the network.

In this work we propose a novel architecture for file sharing and deliberate delaying of large file uploads/downloads with the goal of improving bandwidth utilization, and thereby more readily facilitating social networking. Our solution, dubbed VillageShare, consists of a time-delayed proxy server and a content-sharing Facebook application. The proxy intelligently schedules uploads for times when the Internet gateway is underutilized. The Facebook application allows local storage of locally generated, locally consumed content. The architecture can be implemented using services installed on one application server and two file servers: the application server is placed in a city with well-provisioned Internet, while the file servers are placed in both the city and in the community connected through the bandwidth constrained gateway. The Facebook application installed on the city server tracks the home location of registered users and orchestrates the storage and synchronization of files on the file servers sent between Facebook users in the village and outside the village.

We carry out traffic analysis to establish content generation and file sharing patterns in the network. We find that outgoing traffic follows a diurnal pattern and off-peak periods are not utilized to upload content. Two thirds of these uploads fail during peak-hours. Further, many people do not have access to computers at home during off-peak hours when uploading content would be more successful. There is strong locality of interest in Macha with at least 25% of all shared Facebook photos generated by local users in the village. These results provide strong support for our VillageShare solution.

We evaluate the benefits of our solution through a set of calculations that predict the amount of bandwidth that could be saved using VillageShare as well as the impact of time-shifting flows. Our findings show that between 11% and 22% of the large outbound flow traffic, normally utilizing the satellite gateway, can be sent between local users using a file server hosted locally by VillageShare. The time-delay proxy creates enough off-peak capacity to fulfill the current traffic load with enough capacity remaining to accommodate additional large upload requests. We hope that, as a result, users who previously abandoned the idea of uploading media due to frustration will be encouraged to generate and share digital content, becoming active, full-fledged participants in social networking and content generation and distribution.

2. METHODOLOGY

Content generation and sharing in rural areas can be influenced by a complex mix of technological and social factors [13]. To investigate opportunities for improved content generation and sharing, we perform a holistic investigation of Internet usage in rural Macha, Zambia. Our study incorporates both network performance profiling and traffic analysis as well as a social survey among local Internet users. The network in Macha shares common characteristics with a larger group of networks in rural Africa mentioned in the literature [5] and can be used, at a high level, as representative of rural Internet deployments.

2.1 Macha background

Macha is a resource-limited rural area in Africa with scattered homesteads, very little infrastructure, and people living a subsistence lifestyle; the primary livelihood is maize farming. Like many sub-Saharan rural communities, Macha has a concentrated central area, and a large, geographically dispersed rural community with a sparse population [22]. Macha Works, through the LinkNet project, has deployed a wireless network that provides connectivity to approximately 300 community workers and visitors living around a mission hospital and medical research institute using a satellite-based Internet connection [14].

2.2 Network analysis

Connectivity in Macha is provided over a 6 km^2 area using a combination of 802.11 point-to-point links, hotspots and bridged 802.11 routers. Over the past few years, Internet connectivity has spread to a growing community of users living near the Macha Mission hospital and a community center, which provides public access through an Internet café. The Internet connection in Macha is provided through a satellite gateway that has a committed download speed of 128 kbps bursting to 1 Mbps and a committed upload speed of 64 kbps bursting to 256 kbps with no monthly maximum. The total monthly cost of the connection is $1200 (USD).

To measure traffic usage patterns in the network, we placed a measurement point at the gateway and captured full packet headers of all Internet traffic on the interfaces to the satellite and to the wireless network. All IP addresses were anonymized in order to protect the privacy of the users. We captured approximately 400 GB of traffic, corresponding to two months of network usage from February 2011 to April 2011. Network traces were downloaded on external hard drives and moved to areas with good Internet connectivity to avoid affecting our trace with in-band uploads of the traces.

2.3 Social survey

To complement and explain findings from the network traffic analysis, we administered an online survey of Internet users in Macha. The survey was conducted in June and July 2011 and broadly focuses on usage of Web 2.0 applications and services. The questionnaire consisted of 89 questions and was implemented using the SurveyMonkey tool[2]. Invi-

[2]www.surveymonkey.com

tations for participation were sent via email and Facebook links, and participation was on a voluntary basis. A total of 66 users responded; 41 completed all the questions. We restrict analysis to the latter group. There is significant gender, age and Internet skills diversity within the sample.

Besides descriptive statistics about the fraction of Macha Internet users that generate content, we are interested in associations that may exist between content generation/sharing and technology properties and availability. For the questions we posed we provided two possible types of predetermined answers: a binary "yes/no", and an ordinal "(strongly) agree/disagree" or "daily/weekly/monthly/yearly/never" set of answers. For the first group, we use the independent samples χ^2 test, while for the second group we use Kendall's tau b test of association. We report test results for which the two-tailed significance is lower than .1. We feel that this slightly looser requirement can be justified for the sample size and the domain in which we are working[3].

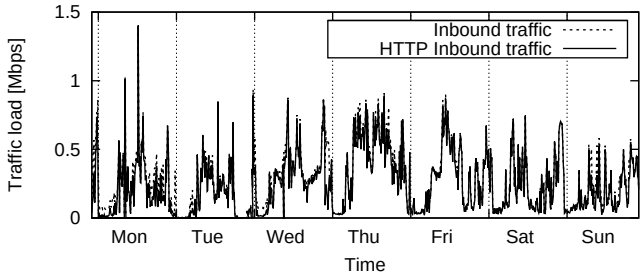
3. CONTENT GENERATION IN RURAL AFRICA

Online presence has great potential to facilitate the preservation of intangible cultural heritage within the local community and its sharing with the wider society. If only access to web-browsing is available, but content generation is not supported, local customs and way of life threaten to be replaced by a lifestyle that is observed online, which mostly originates from urban developed areas [19]. In addition, if the content coming from a rural community is created only by a subset of users, this micro digital divide can even further polarize the community. We contrast content generation with the demographic of Internet users in Macha we surveyed. We find that neither gender, age, nor reported level of IT competence impacts content generation. This promising result emphasizes the importance of facilitating content generation as a tool to tackle the existing inequalities in the developing world [8].

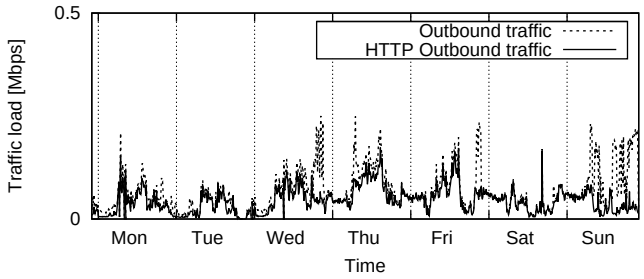
The results of our social survey show that content generation is strongly associated with user-reported accessibility of Internet connection (Kendall's tau-b $(N = 41) = .249, p = .071$). This is well aligned with a previous study [3] that identifies perceived ease of use as one of the key factors for user acceptance of IT. Limited bandwidth is the main hurdle for full fledged Internet access in rural Zambia - 41.5% of the interviewees mentioned it as one of the reasons why they do not spend more time online. Limited bandwidth was a strong differentiator between those who do and those who do not generate content in the US during the transition period from dial-up to broadband connections [9]. Thus, to stimulate content generation and sharing in rural areas, it is crucial to provide support for bandwidth-hungry applications.

4. TRAFFIC ANALYSIS

A user that is limited to content consumption does not experience a full level of utility from the Internet. In rural areas, the lack of bandwidth prevents users from generating and sharing content. The essence of the problem lies in both technology that is used for access (i.e. long distance terrestrial wireless or satellite links), as well as the way online applications are designed (centralized architecture). In this section we investigate specifics of rural area network traffic and identify opportunities for improvement of connection utilization. As described in section 2.2, Macha is connected to the Internet through a satellite. Hence, in the remainder of this paper we focus on a satellite gateway link. However, our work is generalizable to any low bandwidth gateway link.



(a) Download traffic.



(b) Upload traffic.

Figure 1: Network usage over a week (the x-axis ticks mark starts of days). There is a clear diurnal pattern.

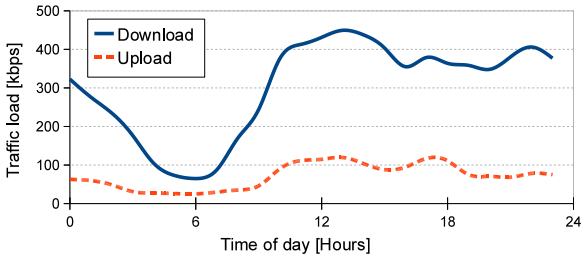
4.1 Temporal characteristics of network traffic

A diurnal pattern of web browsing has been reported in both developed and developing world networks [20, 4, 7]. Daily periodicity, however, is much more important in bandwidth limited networks where the capacity can easily be reached during peak usage times [20]. In this work we examine not only incoming, but also outgoing network traffic, since our goal is improved content generation and sharing.

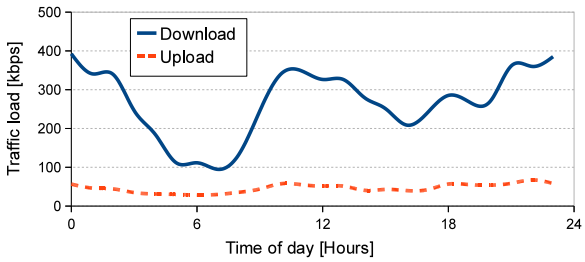
Network traffic distribution in time has been explored in the context of HTTP traffic, through web proxy log analysis [20, 4]. Here we analyze all TCP flows in the full network trace. Thus, the resulting graphs (shown for a week long period in Figure 1(a) for inbound and in Figure 1(b) for outbound traffic) present, besides HTTP, other protocols that utilize TCP. Moreover, this method of analysis allows us to capture those HTTP uploads that have not been completed successfully[4]. Figures 1(a) and 1(b) show a typical daily pat-

[3]We report statistics according to the American Psychological Association standards: χ^2 statistics are reported with degrees of freedom and sample size in parentheses, the Pearson chi-square value, and the significance level; Kendall's tau test results are reported like χ^2, but only the degrees of freedom are in parentheses.

[4]HTTP POST messages are sent out after the actual content is successfully transferred.



(a) Weekdays



(b) Weekends.

Figure 2: Network usage depending on the time of week. Traffic load differs between weekends and weekdays.

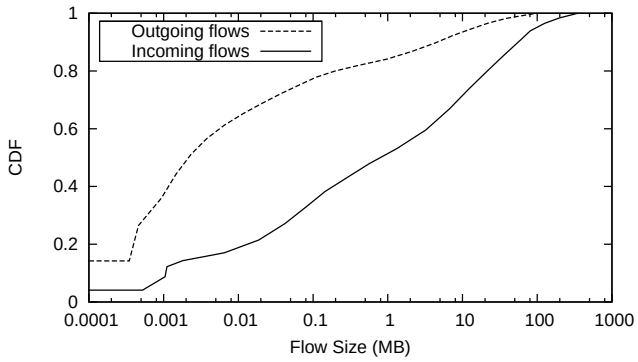


Figure 3: Size distribution of incoming and outgoing flows.

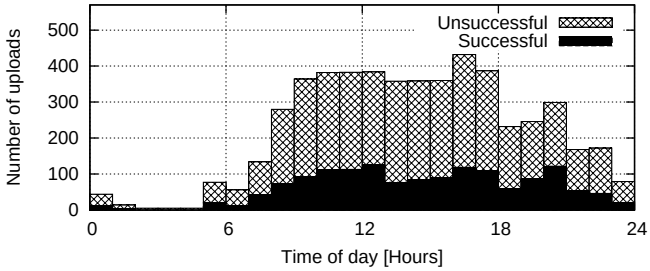


Figure 4: TCP upload attempts sorted in 24 hour bins.

tern. There is little activity late in the evenings and early in the mornings. While HTTP traffic remains dominant, in the upload there are periods when non-HTTP traffic prevails, such as Wednesday, Friday and Sunday evening in the plotted sample.

A closer analysis reveals that the majority of non-HTTP traffic is peer-to-peer (P2P) traffic; in total 8.2% of all traffic is P2P. While it is less pronounced than in the developed world [16], P2P can be highly problematic in rural area networks. P2P networking is a bandwidth costly technique for sharing content, one which networks using an asymmetric satellite gateway connection can ill afford.

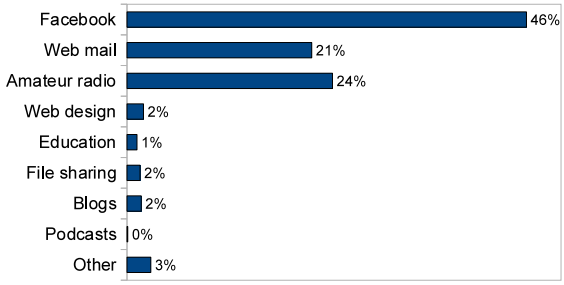
We investigate the difference between total Internet usage in Macha during weekdays and weekends. We take the two-month trace and average observed upload and download traffic over twenty four hours, separately for weekdays (figure 2(a)) and weekends (figure 2(b)). The usage patterns differ significantly. The point of access can play a major role in Internet usage [12], and at-home access is still rare in Macha. Thus, network load is lower in weekends. In addition, the typical diurnal pattern observed on a weekday is less pronounced on weekends. This implies that a solution that relies on low usage periods, such as our proposed proxy (section 5.1), needs to adapt to changing traffic patterns.

4.2 Upload patterns

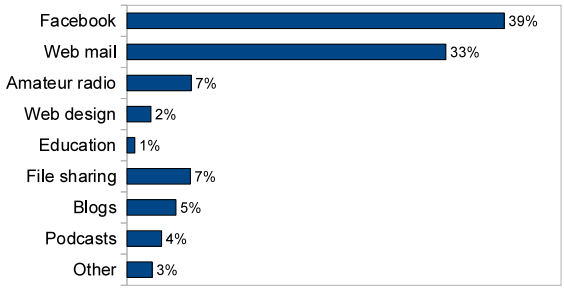
We now concentrate on content generation and in figure 3 plot the size distribution of all TCP flows in either the upload or download direction. According to the figure, there is a significant number of small flows. These correspond to short, often automated messages (i.e. TCP ACKs). User-generated media, such as photos, podcasts and videos, are larger. We take 100KB as a rough size limit over which an upload flow is considered to carry a user-generated content. We first note that the fraction of outbound flows greater than our threshold of 100KB is only 22% by volume. This is in stark contrast to incoming flows where those greater than 100KB account for 80% by volume. Hence there is already a weak tendency to upload content, most likely due to both frustration and a lack of a content-generation culture. Our previous work [11] did show, however, that there is an active messaging culture that is not reflected in flow volume statistics due to their small size.

We inspect HTTP uploads larger than 100KB according to the request time. We take the full two-month trace and put each of the outgoing TCP flows in one of 24 bins according to the time of day when each began. These flows are additionally differentiated based on their completion. In Figure 4 we show both successfully completed and aborted flows. Aborted flows may have timed out or otherwise be incomplete. During the peak usage time the number of upload attempts rises, but the success rate is still very poor.

To identify types of content that users in Macha generate, we inspect the HTTP uploads according to the application. We concentrate on those flows that are larger than 100KB; a total of 1.2GB of such flows, sent via 2260 upload requests, was detected. We extract HTTP POST requests from the trace and inspect the "host" field. We group requests by domain into eight categories based on either frequent occurrences or significant traffic load. The most popular types of traffic are: "Facebook", "Web mail" (Gmail, Yahoo mail and Hotmail), "File sharing" (4share and Skydrive), "Blogs", "Podcasts", "Education" (LinkedIn, Mendeley, JoVE), "Web design" and "Amateur radio" (Hrdlog.com). The traffic classified as "Other" corresponds to online dating services, ICQ, and automated software crash reports, among others. In figures 5(a) and 5(b) we show the distribution of upload requests and traffic volume, respectively.



(a) By number of requests.



(b) By upload size.

Figure 5: Upload web traffic by domain. Distribution of 2260 HTTP POST requests (approx. 1.2GB) observed in the two month trace.

More than two thirds of requests, both by number and volume, correspond to Facebook and web mail. Photo sharing via OSNs and email attachments are common, as can be seen from the trace. However, video and audio uploads are still rare in rural Zambia. We saw only two podcast uploads and no video uploads, though watching (download) videos is common in Macha [12]. Video and audio content can be significantly large; the main reason for the lack of upload media content most probably stems from poor network performance during large uploads. File sharing is present with 7% of total upload traffic, but these files are not large, with the median size slightly larger than 2MB. Blogs account for 5% of total traffic; however, these are often written by foreign visitors that are temporarily stationed in Macha. To our surprise 7% of the uploaded content corresponds to an amateur radio application. We suspect that these are automated updates.

4.3 Locality of online interaction

Outbound traffic from a rural area network can be divided in two groups based on the locality of interest: traffic that is consumed by users outside of the network and traffic that is both produced and consumed by local population. An example of the first type of traffic is an automated software crash report sent to a central application server. The second traffic group includes email attachments sent from one local user to another, or a Facebook photo shared with a number of uploader's local friends.

The locality of interaction in OSNs has been investigated [23], and the findings suggest that the majority of interaction happens among users that reside within a relatively small geographic region. In [11] we analyze OSN interaction locality on an even smaller scale - within a single village. We find that more than half of the interaction is among users who reside in the same local network. While these results open up opportunities for inventive network solutions that can save satellite bandwidth by keeping local traffic local, they are drawn from light-weight Facebook message and/or wall post analysis. In order to evaluate the potential for traffic localization solutions, we investigate the locality of Facebook photo sharing.

Facebook photos are downloaded via an HTTP GET request to Facebook CDN servers. Facebook encodes a user's unique Facebook ID into the filename of any user-generated image. The ID used in the filename will always be the Facebook ID of the user who uploaded the image. We exploit this feature to establish the quantity of Facebook image content that was locally generated, filtering out profile thumbnails as they have no bearing on locality of interest. This is done using a list of local user IDs extracted in the social graph analysis work in [11] and comparing these to the originator IDs of each Facebook user generated image. The results are summarized in table 1. It is important to note that local Facebook user IDs were extracted using instant message triangulation (a method to detect locality by matching outgoing and incoming flows) and only represent a subset of the total local user population. Hence these results represent a minimum bound for locality. The demographics of the local Facebook users, including whether they are foreign or local, is not known but social graph analysis in [11] revealed that 23% of the 182 local users seen in the trace logged in from both inside and outside the village. This represents an upper bound on the number of potential foreign users in Macha.

From our data set, we extract 6066 unique Facebook IDs from user generated images, of which 182, or 3%, are from local users. Although only a small fraction of Facebook users are local, these users generate a significant portion of image content observed in the trace. Our analysis reveals that 9% of the unique user generated images are local. However, 24% of the images viewed, including some viewed repeatedly, are local images. Further analysis shows that a local user views a local image 5.2 times on average, as opposed to 1.5 times on average for an externally generated image. This demonstrates the strong interest in local content among Facebook users in Macha. However, we cannot make any claims about whether or not the same content from Macha is also popular outside of the village. Thus, our localization solution also has to allow distribution of local content to the global Internet.

Table 1: *Locality of Facebook image generation and sharing in Macha, Zambia. A large part of the actual content observed in the trace comes from the village itself, indicating high locality of interest.*

Sample information	
Total unique Facebook IDs	6066
Total local Facebook IDs	182 (3%)
Image counts	
Total unique user images	27403
Total local unique user images	2370 (9%)
Total images viewed	50308
Total local images viewed	12273 (24%)

Facebook uploads represent 39% of the total upload traffic volume (Figure 5(b)). Much of this traffic could be saved by using a localised Facebook image and file sharing applica-

tion; we propose such a solution in Section 5.2. Due to link encryption, we were not able to evaluate the same properties for web mail (33% of the traffic) and file sharing (7% of the traffic); however, we believe that the social structure of the online community in Macha, extracted using Facebook analysis, is applicable in other messaging and file sharing applications. More importantly, the locality of interest displayed in the current network, hampered by a severely congested gateway, reveals a critical opportunity to develop new applications that encourage a much larger degree of content sharing than currently occurs.

5. VILLAGESHARE ARCHITECTURE

In this section we describe our architecture for time-shifting large uploads to low-usage periods in the network and localization of traffic when files are shared between users in the village. We focus on time-shifting uploads as time-shifting downloads has already been addressed in previous literature [20, 10]. In addition, supporting uploads meets our objective of facilitating content generation.

5.1 Time-shift upload proxy

The goal of the time-shift upload proxy is to "equalize" traffic burden on the satellite link. Due to congestion, uploads often fail (figure 4). One way of relieving the congestion is to reschedule large uploads for periods when the network is underutilized, such as nights and weekends. Two important benefits arise from such a policy: first, uploads are likely to be completed successfully, and second, short interactive flows, that are crucial for local communication via IM/VoIP, for example, perform better with large flows out of their way.

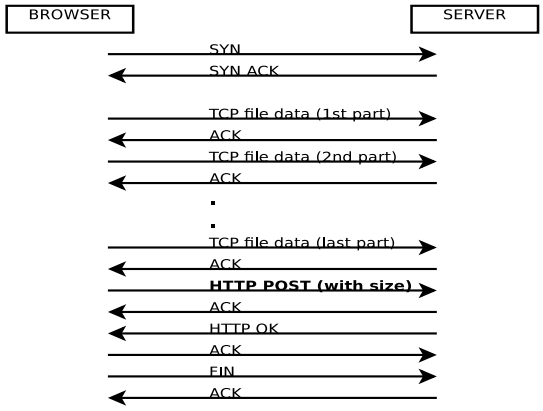


Figure 6: TCP flow messages during a file upload.

To distinguish between short and long flows we first examine the HTTP upload behavior. A file upload using the HTTP protocol involves an ordered set of TCP messages shown in figure 6. This message exchange reveals that the size of the content is only known after the content has been uploaded. The size of the content is contained in an HTTP POST message at the end of the flow. Hence the only information that can be used to determine a size of the flow, before the entire file is uploaded, is the actual real-time TCP traffic flow as the upload is in progress.

Another challenge for shifting large uploads is to determine a suitable time to reschedule traffic. While exact prediction of network usage at any specific moment is impossible, we can use daily and weekly trends in network usage, as they exist, to more accurately predict future utilization. A 24 hour running average of outgoing traffic is shown in figure 2(a) for Mondays to Fridays and figure 2(b) for Saturdays and Sundays. We estimate the maximum average capacity in the network using the maximum values in these figures. Since satellite connectivity uses a time-shared link, in which many other users within the satellite's coverage area occupy the same frequency, the capacity may fluctuate over time as users come and go. Even if that fluctuation happens, one would expect that all satellite users within the same timezone of the satellite coverage area would exhibit the same usage pattern (similar to figures 2(a) and 2(b)) and have peak usage periods at the same time. Thus, our capacity estimation represents the lower bound, and the actual capacity during off-peak times might be even higher, which provides strong motivation for our time-shifting proxy.

We consider two approaches for identifying time periods when spare capacity is available and in which we can reschedule large uploads. The first is termed "dynamic" and attempts to scavenge any spare capacity, defined as the area between bandwidth consumed by non-rescheduled, current outgoing traffic and the capacity of the satellite. The second strategy identifies "off-peak" periods in which the traffic load is below a pre-determined threshold. This threshold value can vary such that the time-shifting proxy is able to service all the delayed uploads within 24 hours. To harness short periods of extra capacity, "dynamic" requires real-time usage estimation, for example, via TCP round trip time measurements; "off-peak" relies on a moving average of usage, and can only capture trends in capacity usage. Due to a significant variation in traffic load (figure 2), the threshold can be calculated separately for weekends and weekdays.

The final piece of our proposed proxy is a set of policies that manage upload queuing and scheduling. Our scheduling policy has to balance between upload success and interactivity. When a large upload is observed during a peak period, the proxy captures the file upload in its local cache and notifies the user that the upload will be delivered to its final destination when a sufficient capacity is available. When such a moment arrives, queued upload requests are served in a FIFO manner and the cached content is transferred via the satellite link. As a part of our future work we plan to investigate more complex approaches that take into account upload size (similarly to Bassa [20]) and real-time satellite capacity measurements [2].

The full architecture of the time-shifting proxy is shown in figure 7. A gateway (labeled "village gateway") is placed between the client and the Internet in order to intercept an upload. Once the outgoing flow exceeds a threshold size and the satellite link usage is high, the gateway redirects the user to a local upload page (labeled "virtual server") and the connection via the satellite link is disabled. The upload page will capture the user's authentication information required to upload the content to a content server at a later stage. An example of a common protocol used to authenticate a user with a content server, such as Facebook or YouTube, is OAuth. The gateway stores the URI details of the upload in a FIFO queue (labeled "upload request queue") and will reuse this information to upload the content when spare capacity is available.

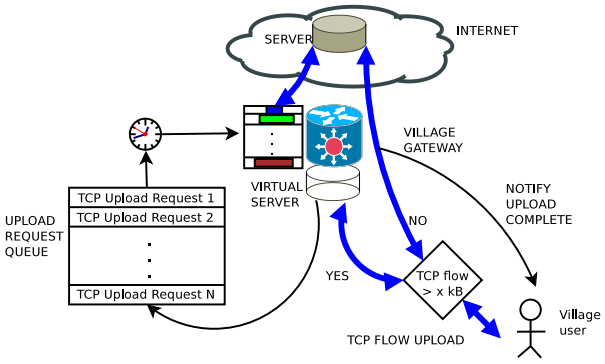


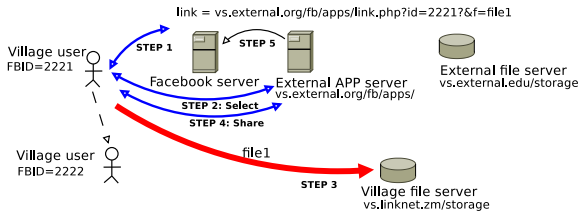
Figure 7: Architecture of a time-shift proxy for upload.

Figure 8: Main VillageShare application screen.

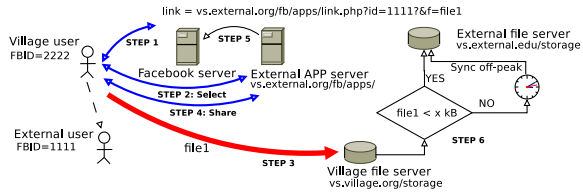
5.2 Localized and time-shifted file sharing

Web 2.0 introduced numerous applications that allow users to generate and share content: Facebook, Picassa, Youtube, and Google documents, to name a few. However, these applications store content on servers that may be geographically distant from where the users are located. While this does not present a major issue for well connected broadband users, it significantly deteriorates Web usability for those who have low bandwidth connections [23].

To minimize local traffic on the Internet gateway and thereby improve the user experience, we develop the VillageShare Facebook application shown in figure 8. VillageShare ameliorates expensive satellite link traversal by storing locally generated content on a server located within the local network. A user is able to upload a file to the file server and then share the content with their friend list. As a Facebook application, VillageShare inherits all the useful functionality of Facebook. Figure 8 shows the main screen that is presented to the user, where a list of uploaded files and file operations is presented. Each uploaded file can be shared, downloaded or deleted. When the share option is selected, the user is taken to the file sharing screen where features inherited from Facebook's functionality are used to search for friends to share a file. Content can either be shared using a direct message to a friend or by posting to the owner's wall or a friends wall. The message contains a unique URI, containing the content creator's Facebook ID and the File



(a) Sharing between two village users.



(b) Sharing between a village user and an external user.

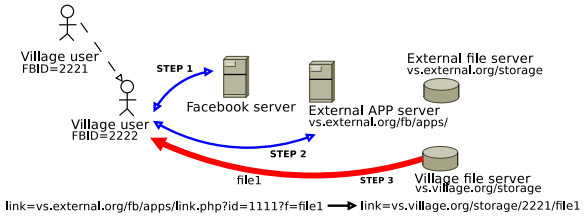
Figure 9: Architecture of VillageShare Facebook application for uploading files.

ID, which will launch VillageShare and display the shared file to the user.

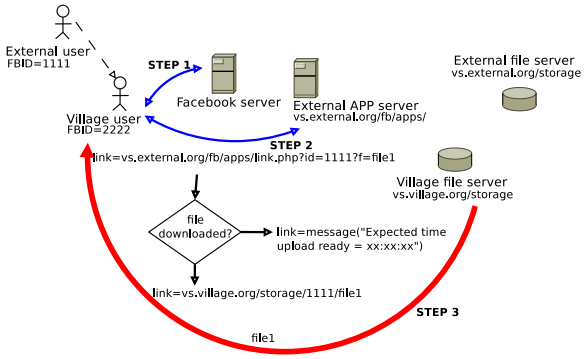
The architecture of the system is shown in figure 9(a). To support VillageShare functionality, a number of servers work together: (1) the Facebook server; (2) a VillageShare Facebook application server hosted externally; (3) a local file storage server hosted in the local village; and (4) a file storage server hosted externally. A key component of VillageShare is determining whether a user is inside or outside the village. This can easily be achieved by checking the IP address of the user using the VillageShare application. All users from the village are routed through a satellite gateway using NAT, which is allocated a single IP address by the satellite service provider. When this IP address is detected, VillageShare knows that the user is in the village; all other IP addresses are outside the village. Note that user roaming is implicitly supported, as we do not use Facebook IDs, but IP addresses to establish the locality.

The following steps, highlighted in figure 9(a), are followed when sharing files between local village users: (1) The user connects to Facebook and logs into their account. (2) The user launches the VillageShare application from Facebook. (3) The user uploads a file to the local file storage server through VillageShare. (4) The user selects one or more local friends with whom to share the file. (5) A message is posted on the recipient's wall or inbox with a link to the file using embedded PHP fields. All interactions with the Facebook server and the External Application server are lightweight HTTP messages; media is only sent to the Village file server.

It is also possible for a user in the village to use this application to share a file with a user who is outside the village. From a user perspective the process is the same as when sharing a file with a local user; however, in the back-end new techniques are required to prevent the satellite gateway from being used during peak usage periods. Figure 9(b) shows the case where a village user shares a file with an outside user. The first five steps are the same, with an additional synchronization step (6) added. The synchronization engine checks to see whether the file exceeds a threshold size. If so it delays synchronization between the local village file server and external file server, as described in Section 5.1.



(a) Sharing between two village users.



(b) Sharing between a village user and an external user.

Figure 10: Architecture of VillageShare Facebook application for downloading files.

In the download direction, when sharing a file between local village users, the following steps, shown in figure 10(a), takes place: (1) Recipient logs into Facebook. (2) User notices that they have a message or wall post with a link to a shared file and selects the file. VillageShare is automatically launched. This is not a hard-coded URL but rather a link to VillageShare with parameters specifying the source user and file being shared. (3) VillageShare knows that the user attempting to retrieve the file is local, using the IP address method discussed, and also knows that the file was generated by a local user using a location tag stored with the file which stores the location of the user who uploaded the file. Using this information it generates a URL pointing to the source file on the local village server and the file is downloaded. In the case of a local user downloading a file posted by an external user, shown in figure 10(b), the link will either resolve to a message saying that the file is not downloaded to the local store yet, or it will resolve to a URL pointing to the file on the local village sever. The synchronization status will be known as the Facebook application server will be notified when synchronization is complete.

6. VILLAGESHARE EVALUATION

Because we have not yet been able to deploy it in a rural network, it is difficult to evaluate the impact of VillageShare. We are planning to install VillageShare in a rural Macha Works/LinkNet location in the coming months. Hence, here our evaluation methodology is to utilize intuitive calculations based on our traffic analysis to predict bounds for performance improvement provided by VillageShare. We focus on outgoing traffic as our goal is to facilitate file uploads.

6.1 Traffic localization using VillageShare

VillageShare relieves Internet link congestion by bypassing the gateway and keeping local traffic within the local network. We investigate how much new capacity would be made available based on the amount of local traffic. We perform the evaluation assuming user behavior as observed in the trace. We focus our evaluation on the dominant upload traffic from Facebook and email, as these applications are most likely to benefit from VillageShare.

We infer the amount of locally produced and locally consumed traffic from the locality of Facebook photos calculated in Section 4.3 and from other recent work [11]. The former finds about 25% locality in photo sharing over a bandwidth constrained satellite link; the latter investigates less constrained Facebook message sharing in the same network and finds 50% locality of interest. Thus, we take 25% and 50% as the lower and the upper bound on Facebook locality, respectively. If we assume that this trend is consistent in other forms of content sharing, we can apply this fraction to emails with attachments sent to local users.

Based on this assumed amount of local traffic, figure 11 shows the bandwidth saved over an averaged 24 hour cycle with VillageShare. Much of the bandwidth savings occurs during the day, which helps relieve the congested gateway and improve interactive application performance. The remainder of the large uploads, not shared with local users, will utilize VillageShare's time-shift proxy. Table 2 summarizes the upper (50%) and lower (25%) bound of bandwidth savings due to VillageShare over a two month period. All traffic volume percentages are given as a fraction of large outgoing flows excluding P2P traffic. We exclude P2P traffic as it does not capture explicit user upload activity. We anticipate that, as upload performance improves, upload traffic volume will increase. Hence, these savings, by volume, are likely to escalate when much larger files, such as videos and software, are shared locally. Moreover, VillageShare leads to improved user experience, due to fewer failed uploads and faster upload times.

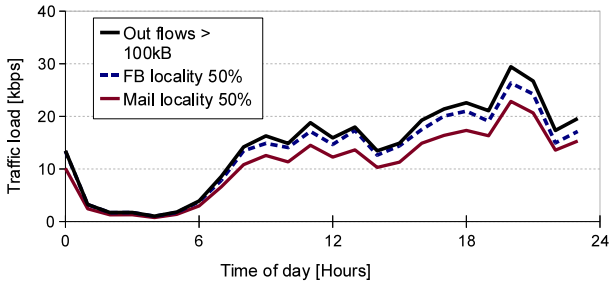


Figure 11: Capacity savings with large upload handling through VillageShare.

6.2 Time-shift upload proxy

We next estimate the amount of spare uplink capacity the time-shift proxy will be able to utilize in Macha. The satellite data bundle used in Macha claims to provide a 64kbps committed rate and a 256kbps bursting rate. From our traffic analysis, we found an average peak capacity of approximately 128kbps (see figure 2(a)); we use this as our baseline uplink capacity. This provides a maximum daily outgoing capacity of 1328MB. We assume that traffic between local users harnesses the VillageShare file-sharing application. As a result this traffic does not figure in the flows we consider for rescheduling. We concentrate on the remaining large

Table 2: *Summary of outgoing bandwidth used and potential bandwidth savings due to VillageShare.*

Traffic types	MB
Total outgoing	45287
Total outgoing > 100KB (excl. P2P)	6366
email clients	1568 (24.63%)
Web-email	611 (9.60%)
Facebook	744 (11.69%)
VillageShare(VS) savings	
Facebook 25% locality	186 (2.92%)
All email 25% locality	545 (8.56%)
Facebook 50% locality	372 (5.84%)
All email 50% locality	1090 (17.11%)
Total saved due to VS (25%)	731 (11.48%)
Total saved due to VS (50%)	1462 (22.96%)

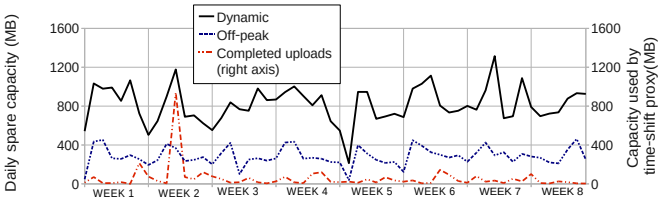


Figure 12: The capacity available each day for the time-shift proxy using dynamic and off-peak allocation strategy. An estimate of the bandwidth used by the proxy for completed uploads is also shown.

uploads (greater than 100kB) which are rescheduled by the time-shift proxy; hence this traffic is initially removed from the current set of outgoing flows.

We evaluate the two spare capacity estimation strategies elaborated in Section 5.1: "dynamic" and "off-peak". In the "dynamic" strategy the proxy automatically dequeues any enqueued flows when capacity becomes available. The threshold for the "off-peak" strategy is set to a third of the satellite capacity (42.6kbps). This results in an off-peak daily period from 3 to 8 A.M. with an additional weekend period from 2 to 5 P.M. The proxy begins to dequeue flows at the beginning of the off-peak period.

Figure 12 shows the spare capacity available using these two strategies over the 8 week measurement period, in intervals of days. Clearly, the dynamic strategy provides far more capacity, with an average of 700MB per day; however, it has the disadvantage of negative interaction between upload traffic (long-lived flows) and browsing traffic (short-lived flows) [6]. The off-peak strategy provides an average of 278MB extra capacity per day, which is sufficient for all but one day in our 2 months network measurement. The large dip in capacity at the beginning of week 5 is due to a full-day power failure.

We now reinsert the removed uploads, destined for external consumption, back into the network. We conservatively assume that 75% of all large email, Facebook and non-P2P large outbound flows are these externally consumed uploads. They are then reinserted using one of the two strategies, "dynamic" or "off-peak". We assume that the measured size of these flows in our traces was artificially small due to the aborted transfers. To make the sizes of these flows more realistic, we double the length of these flows. The bottom line in figure 12 shows the amount of capacity required by these flows over time. Both proxy strategies can handle the majority of the current outbound flows greater than 100kB within a 24 hour period. The sudden increase in demand in the middle of week two is due to an unusual email client anomaly, where large attachments were transmitted continuously throughout the day. For this single anomaly, only the dynamic bandwidth allocation strategy would have been able to service the uploads within a single day.

It is likely that the improved network performance, due to local caching, will cause users to increase the quantity of uploads. The evaluation shows that there is enough capacity to handle a substantial increase in uploads, but we acknowledge that only an in-situ deployment will be able to evaluate the efficacy of the time-shift proxy as outgoing load increases beyond its current pattern.

7. RELATED WORK

The lack of local content in rural Africa was noted by Van Hoorik and Mweetwa [19]. According to their observations, rural Africans do not find a representation of their customs and culture online, thus they may perceive the Internet as a "foreign body". In [13] models of online content generation are explored. The authors conclude that social and cultural factors have to be considered for successful implementation of content generation tools. In our Internet usage survey (Section 4), we investigate why content generation is lacking even in an existing network with experienced Internet users. We find that the lack of easy access to computer terminals and the lack of bandwidth inhibit content generation.

To cope with limited bandwidth, we propose a time-delayed proxy and an OSN-based local file sharing system. Time-delayed proxy for bandwidth-limited networks was first proposed in [4]. A Collaborative cache approach was proposed by Isaacman and Martonosi [10] in which cached content on client devices is made available to all clients in the local network. This solution acknowledges the value of local stored content as well as the importance of distributed access in unreliable networks. A system similar to the proposed one was implemented by Vithinage and Atukorale [20]. We build upon this idea and extend it so that both uploads and downloads are handled in a time-delayed manner. Locality of online interactions, on the other hand, was first observed with the advent of online social networks. Wittie et al. [23] exploit locality of interest in order to improve OSN usability for remote areas. Their work is geared towards larger geographic regions (e.g. whole countries). In this paper we find a micro-locality within a single village can be used as a basis for the improved file sharing scheme we propose in Section 5.

8. DISCUSSION AND CONCLUSION

VillageShare is not a method that transparently solves the problem of bandwidth shortage in rural area networks. Rather, it is an intervention that facilitates content generation and sharing by providing an alternative to the existing paradigm of interactive file upload, which clearly does not work well in bandwidth-poor networks. We believe that the poor performance experienced in uploading mail attachments, podcasts, videos or other files shared with local village users will convince users to converge on usage of the VillageShare application. Although the system is closely tied to Facebook, we will also explore the option of provid-

ing an API for the synchronization engine so that other new applications could reuse this functionality.

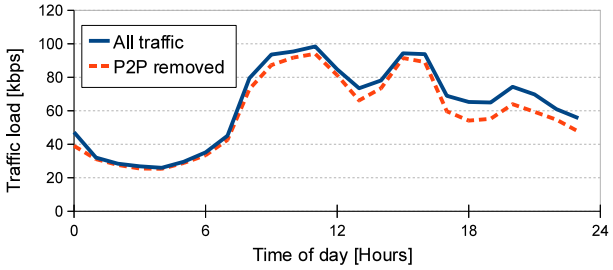


Figure 13: Effect of removing all P2P upload traffic on the network (averaged two week trace).

A tradeoff between impacting user behavior and improving network performance exists on multiple levels. VillageShare exploits locality of interaction and time shifting; traffic shaping is yet another aspect. For example, we find that peer-to-peer traffic utilizes 8.2% of the total bandwidth, while exhibiting very poor performance over an asymmetric satellite link. To examine the impact of this traffice, we show in figure 13 the upload capacity savings when all P2P traffic is blocked at the gateway through a simple firewall rule. Although this could seem like an attractive way to relieve congestion, there are cases when content is only available through a P2P network; it would be unwise to completely remove the opportunity to use this mechanism. A tradeoff solution could be a combination of a complete blockade of P2P networking in the network combined with a service on a dedicated server outside the network to forward P2P file requests. The files can then be downloaded to the local server using the time-shift proxy method that we described earlier.

Bandwidth is not the only restriction on content generation. Cost associated with Internet access is extremely important for low-income rural residents. Our interviews reveal that when Internet usage is limited by cost, it significantly impacts the frequency of content generation ($\chi^2(N = 41) = 3.475, p = .006$). Moreover, the number of hours a user spends using the Internet (irrespective of whether one has to pay for it) determines the user's affinity towards content generation ($\chi^2(N = 41) = 5.218, p = .035$). Access at work, school or any other public location comes with restricted hours and leads to "the deliberate interaction" model [24]. Only at-home access allows leisurely content generation and sharing; we plan to tackle this aspect of the problem in our future work. In our future work we also plan to carry out a long term longitudinal study to capture the content generation and content sharing behaviour of users in the village before and after the VillageShare system is implemented.

In conclusion, we hope that additional off-peak capacity enabled by the VillageShare time-delay proxy and improved local traffic exchange through the VillageShare Facebook application will enable rural users in developing regions to be more active producers and sharers of online content.

9. ACKNOWLEDGEMENTS

This work was supported in part by NSF Network Science and Engineering award CNS-1064821. The authors would also like to thank Linknet in Macha, Zambia for their continued cooperation on this project.

10. REFERENCES

[1] R. Abraham. Mobile Phones and Economic Development: Evidence from the Fishing Industry in India. In *ICTD'06*, Berkeley, CA, May 2006.

[2] L.-J. Chen, T. Sun, G. Yang, M. Y. Sanadidi, and M. Gerla. End-to-End Asymmetric Link Capacity Estimation. In *NETWORKING'05*, Waterloo, Canada, May 2005.

[3] F. D. Davis. Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *MIS Quarterly*, 13(3):319–340, September 1989.

[4] B. Du, M. Demmer, and E. Brewer. Analysis of WWW Traffic in Cambodia and Ghana. In *WWW'06*, Edinburgh, UK, May 2006.

[5] R. Flickenger, C. Aichele, C. Fonda, J. Forster, I. Howard, T. Krag, and M. Zennaro. *Wireless Networking in the Developing World*. Limehouse Book Sprint Team, first edition, January 2006.

[6] L. Guo and I. Matta. The war between mice and elephants. In *Network Protocols Ninth International Conference on ICNP 2001*, pages 180–188, 2001.

[7] T. Henderson, D. Kotz, and I. Abyzov. The Changing Usage of a Mature Campus-Wide Wireless Network. In *MobiCom'04*, Philadelphia, PA, September 2004.

[8] M. Hilbert. Digital gender divide or technologically empowered women in developing countries? A typical case of lies, damned lies, and statistics. *Women's Studies International Forum*, 34(6):479–489, November 2011.

[9] J. Horrigan. Home broadband adoption 2006. Technical report, Pew Internet & American Life Project, May 2006.

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

[11] D. Johnson, E. Belding, and G. van Stam. Network Traffic Locality in a Rural African Village. In *ICTD'12*, Atlanta, GA, March 2012.

[12] D. Johnson, V. Pejovic, E. Belding, and G. van Stam. Traffic Characterization and Internet Usage in Rural Africa. In *WWW'11*, Hyderabad, India, March 2011.

[13] M. A. J. Manschont and C. Stroek. Stimulation of Local Content Generation in Rural Africa. In *IST-Africa'09*, Lake Victoria, Uganda, May 2009.

[14] 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'07*, Bangalore, India, December 2007.

[15] U. Nations. *Wireless Internet Opportunity for Developing Countries (Ict Task Force Series)*. United Nations, 2004.

[16] H. Schulze and K. Mochalski. ipoque Internet Study 2008/2009, 2009.

[17] S. Sounders. The HTTP archive. http://httparchive.org.

[18] 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'08*, San Francisco, CA, April 2008.

[19] P. van Hoorik and F. Mweetwa. Use of Internet in Rural Areas of Zambia. In *IST-Africa'08*, Windhoek, Namibia, May 2008.

[20] W. W. Vithanage and A. S. Atukorale. Bassa: a time shifted web caching system for developing regions. In *NSDR'11*, Washington DC, June 2011.

[21] L. Waverman, M. Meschi, and M. Fuss. The Impact of Telecoms on Economic Growth in Developing Countries. *The Vodafone Policy Paper Series*, 2:10–23, 2005.

[22] Wikipedia. Macha, Zambia. http://en.wikipedia.org/wiki/Macha,_Zambia.

[23] M. Wittie, V. Pejovic, L. Deek, K. Almeroth, and B. Zhao. Exploiting Locality of Interest in Online Social Network. In *CoNEXT'10*, Philadelphia, PA, December 2010.

[24] S. P. Wyche, T. N. Smyth, M. Chetty, P. M. Aoki, and R. E. Grinter. Deliberate Interactions: Characterizing

Technology Use in Nairobi, Kenya. In *CHI'10*, Atlanta, GA, April 2010.