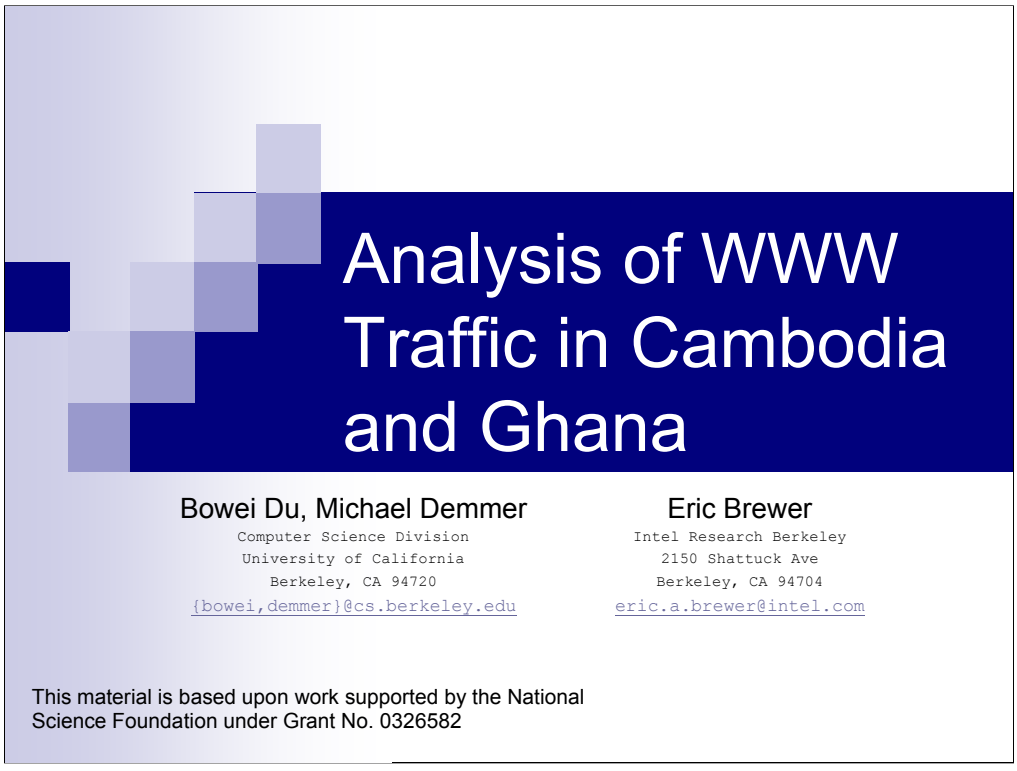
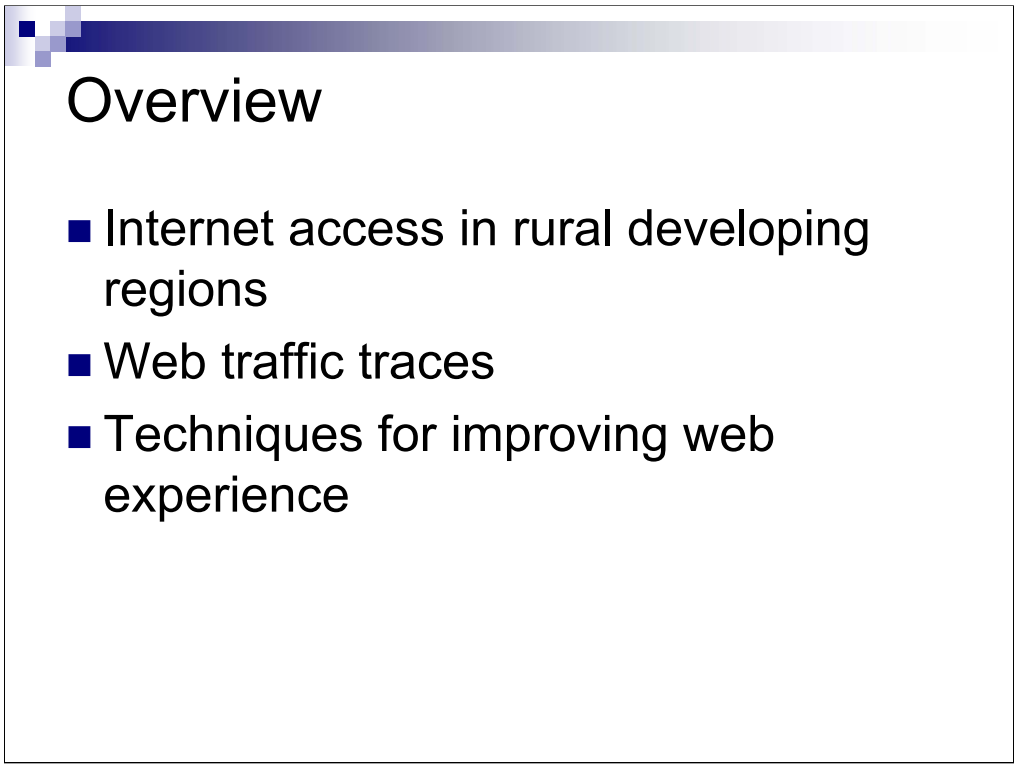


(1) This is joint work with

(1) **Mike Demmer** @ UC Berkeley

(2) **Eric Brewer** @ Intel Research Berkeley

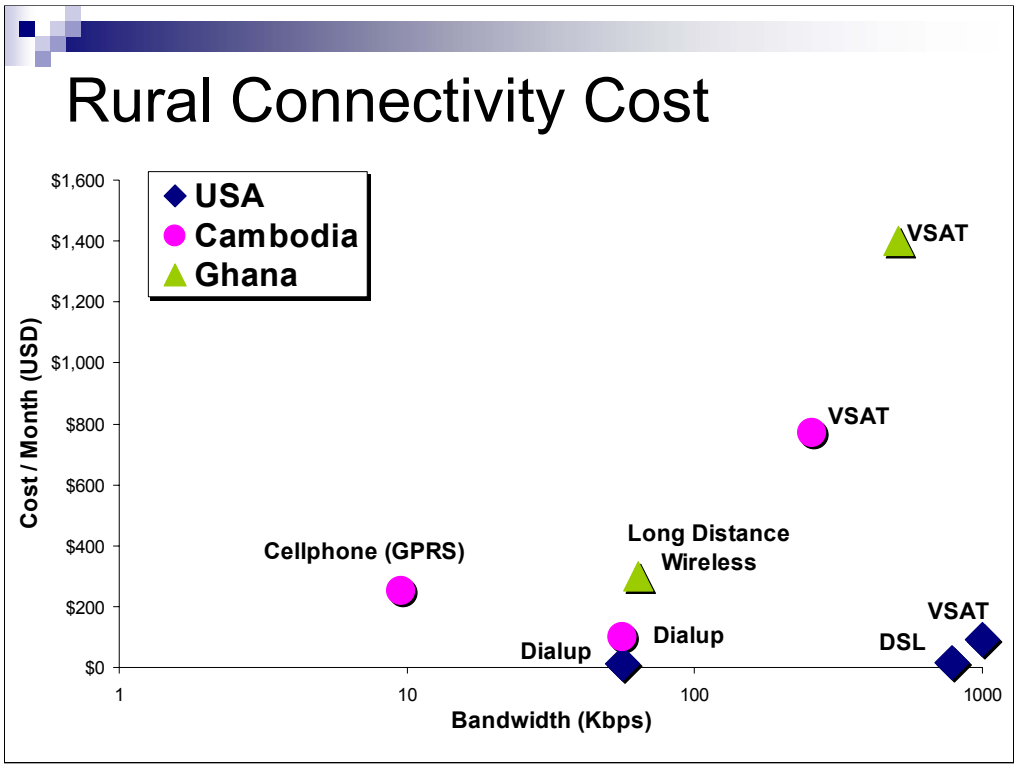


- Today I will be giving a talk:
  - Characteristics of Internet connections Cambodia and Ghana
  - Properties of web traces gathered from two rural developing regions, Cambodia and Ghana.
  - Talk about techniques that can be used for improving web experience

Rural connectivity

- **Quality is poor**:
  - Non-trivial latency and loss
  - Rural connections in Cambodia:
    - 1 – 2 second roundtrip time, up to ~10% packet loss
  - TCP doesn't behave well
- **Bandwidth is low**:
  - Sharing
  - Cost

- Rural web access in these two countries is challenged by quality and cost of connectivity.
- **There is non-trivial latency and loss** – enough – to significantly affect web experience
  - Measured data from Cambodia.
  - TCP/HTTP does not do very well under these situations.
- **Internet access is usually shared among many users through an Internet kiosk.**
  - Picture illustrates such a usage case
  - Bandwidth available for a single is a small fraction of the bandwidth of the connection itself.
- Not only is connectivity bad but it is also **expensive in terms of price** … in the next slide…



- Chart that is a comparison cost of connectivity in Cambodia and Ghana compared with prices in the United States.
- X axis is bandwidth on a log scale
- Y axis is cost in US dollars per month
- Several important things to note:
  - **In the two countries we examine, bad infrastructure limits the types of connectivity that can be used.**
    - Cell phone or VSAT only option when ground infrastructure is unavailable
    - As can see, their cost is at least an order of magnitude higher than dialup/DSL
  - **Bandwidth is not directly related to cost**
    - Cell very bad at 9.6 kbps but expensive at $250/month
  - **Cost and ability to connect can vary with time**
    - Dialup costs depend on load
    - Operators know when what times are good for connecting

Overview

- Internet access in developing regions
- Web traffic properties
- Techniques for improving user experience

•Having given you a brief description of the underlying network transport, we now move onto a discussion of the web traffic we captured.

Web Traffic Logs

- **Cambodia:**
  - Community Information Centers (CICs)
  - 6 month web proxy log (~12 million URLs, 110 GBs web objects)
  - ~16k users total, average of 85 users/day
  - 64 – 128 broadband kbps with VSAT uplink at ISP
- **Ghana:**
  - Busy Internet Café
  - 1 month web proxy log (~14 million URLs, 106 GBs web objects)
  - ~100 users/day
  - VSAT uplink
- Internet blocked by firewall, all traffic through proxy

- Captured two sets of web proxy log data
  - Both sets of data were from shared use Internet Kiosks
  - Cambodia
    - 6 months trace
    - 12 million URLs
    - Representing 110 GBs of web objects
    - 85 users/day
  - Ghana
    - 1 month trace
    - 14 million URLs
    - Representing 106 GBs of web objects
    - 100 users/day
- Internet blocked by firewall, IPs anonymized by Network Address Translation

Web Traffic Classification

- Classification based on URL path into general categories:

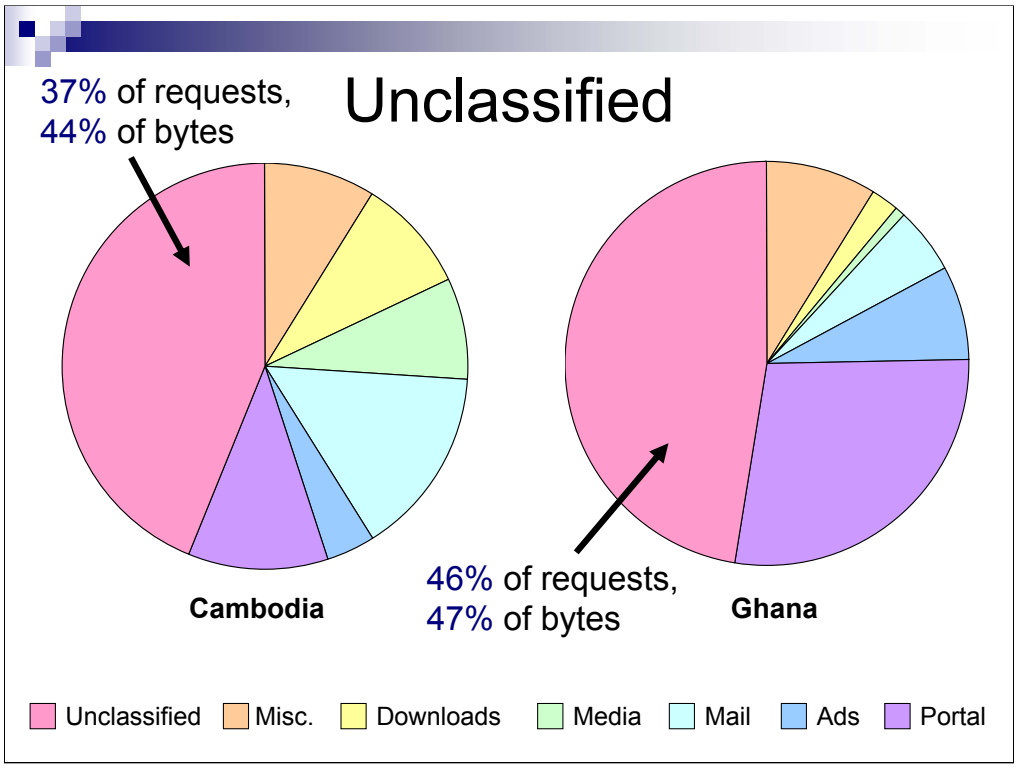
URL	Category
http://mail.yahoo.com/ym/ShowFolder	E-mail
http://www.yahoo.com/	Portal

- Advertising identified by ad-blocking software blacklist (http://www.pierceive.com/)
- Not exhaustive, but does show larger content trends

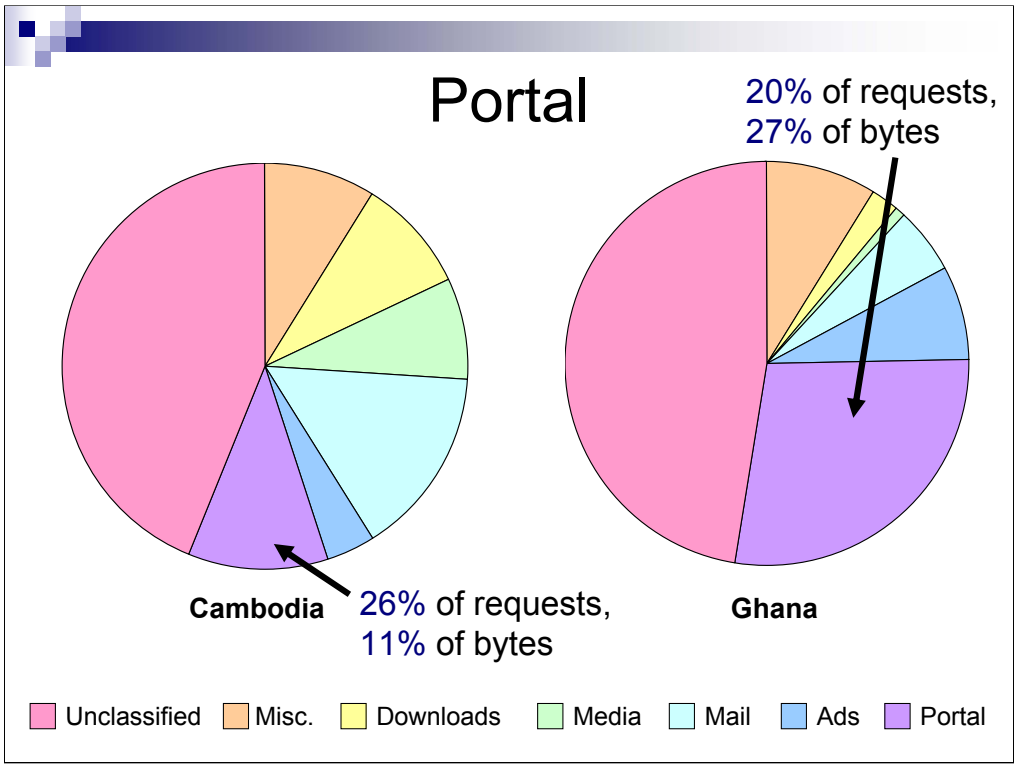
- **Question we wanted to look at was: What kind of content was viewed by the user**
- We classified websites viewed into broad categories based on the URL
- Advertising sites were identified using popular ad-blocking software blacklists
- I am going to say up front that this is by far a rough cut and not at all exhaustive, but does reveal some of the large trends in the content.

- The following are pie charts of the number of bytes in each category in each country. **I will be going over the highlights of the data for conciseness.**

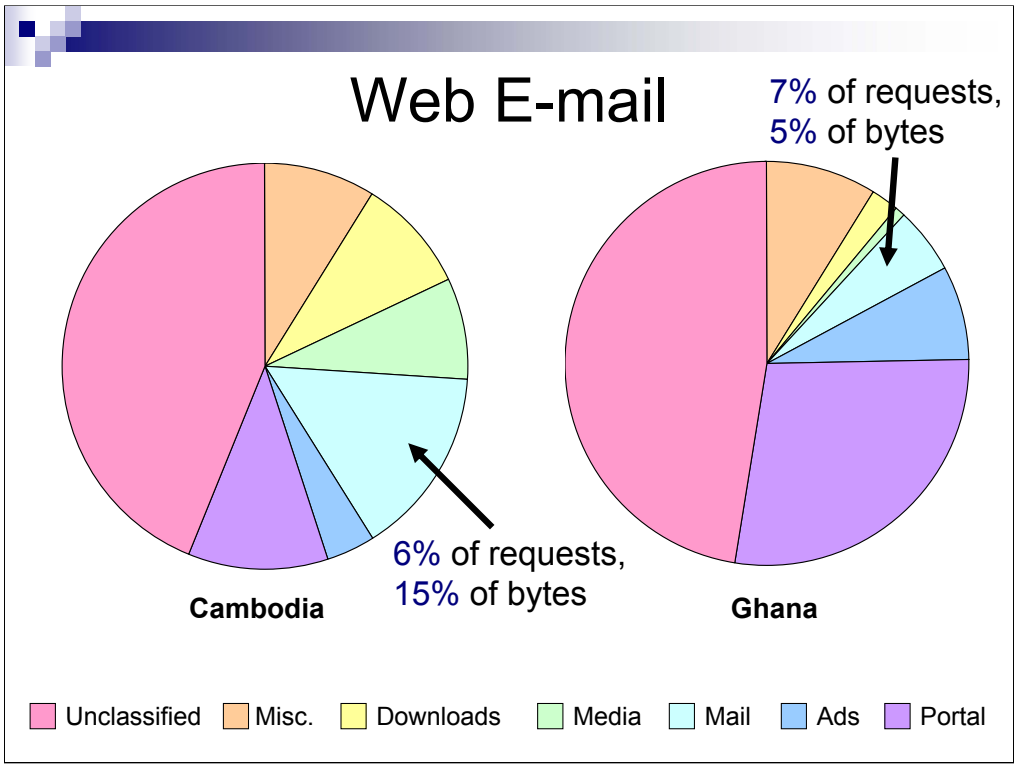
- Both # of requests and bytes have an impact
  - # requests because of long latency and connection quality. A page with many objects on it loads much slower than a page as a single object
  - Bytes because of bandwidth



- As I said before, the classification is a rough estimate of content viewed. As you can see little less than half could not be easily classified into large general categories.
    - Diminishing returns.
    - Most likely there is no good general group of websites hiding in the unknown chunk.



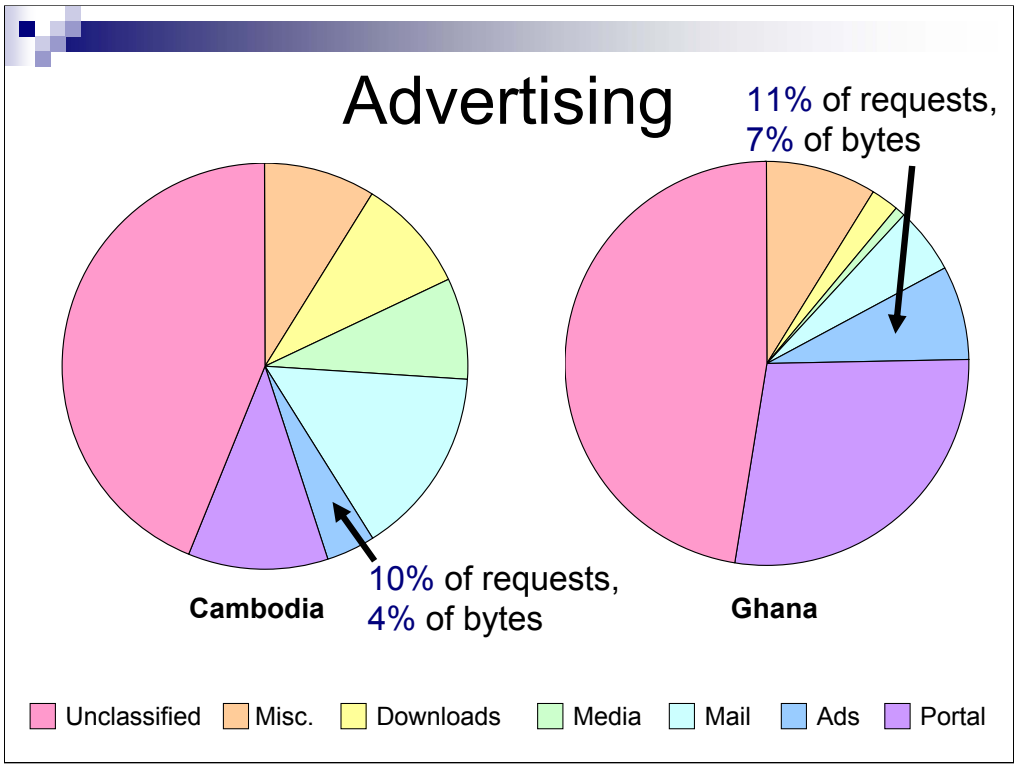
- Portals
- Front page websites such as Yahoo!, MSN
- Found some localized sites, but by far the most were portal sites for the United States
- Localized portals would be very well received.



•Web e-mail is a the popular web application in Cambodia and Ghana.

•E-mail style application itself is well suited for badly connected user

•Most operations are local, and then data is sent in batch to the server

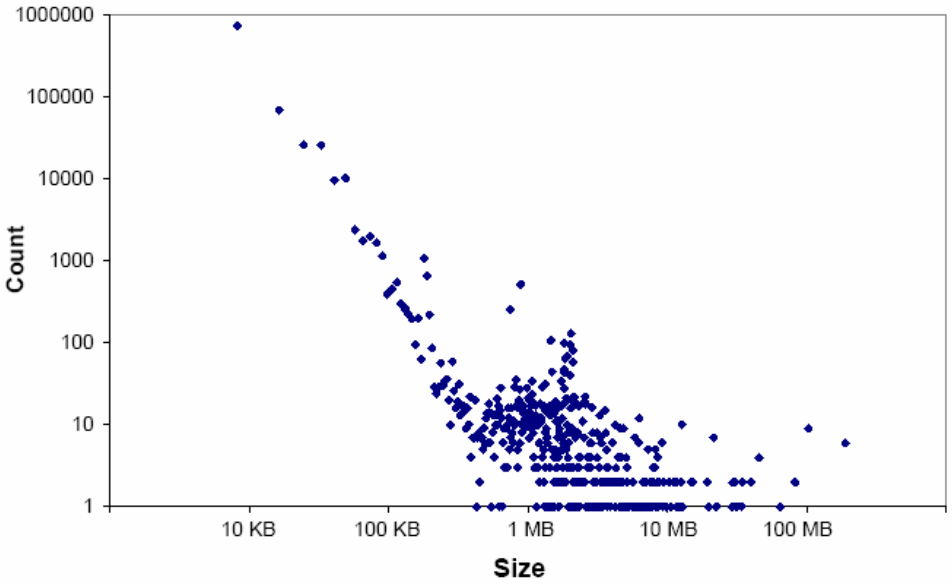


- Finally, the last category of URLs I will talk about is advertising.

- Advertising is completely irrelevant for users in their home countries.
- Many advertisements for services, such as Vonage, not available in country.
- Wasted bandwidth for the users.

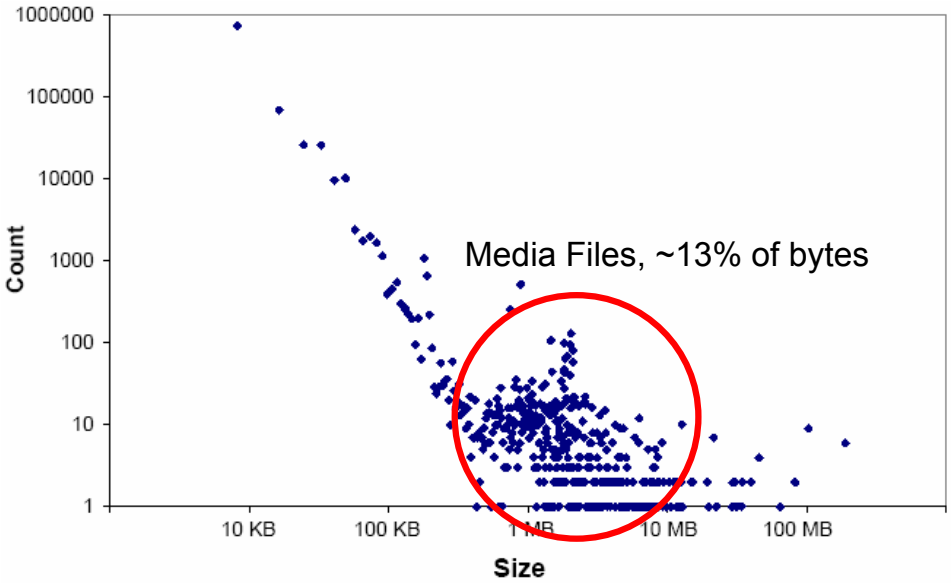
- Classification aside, we also looked in more detail at characteristics of the web objects requested. We only show the data collected from Cambodia in this talk for succinctness.

Traffic Size, CIC Data

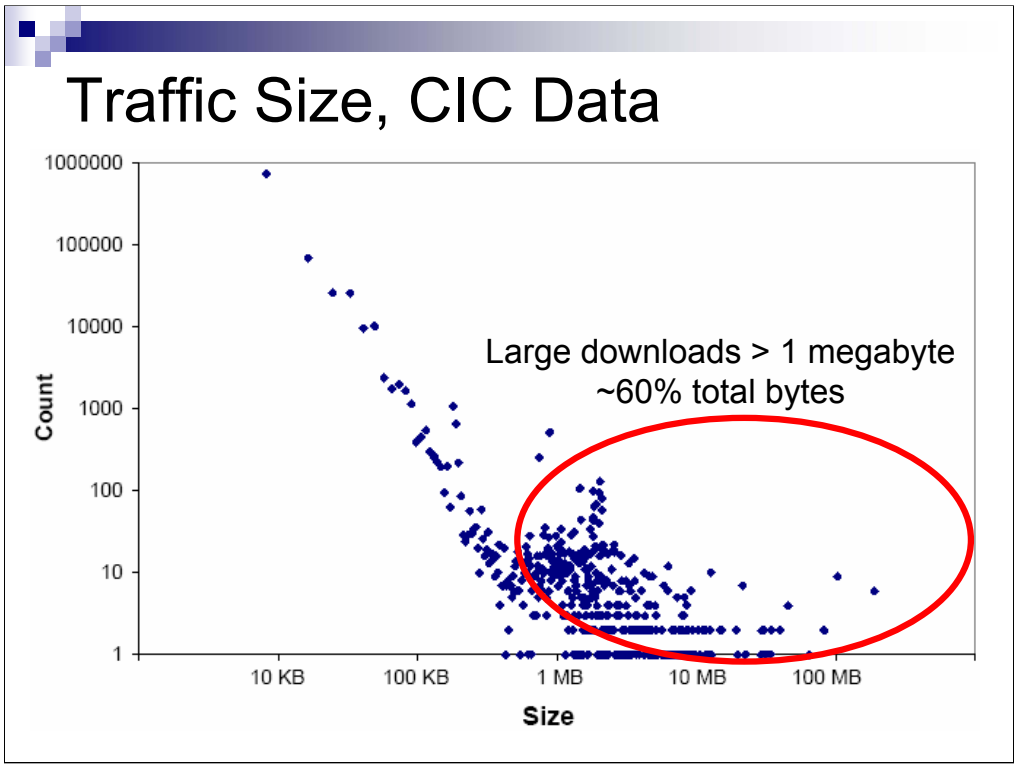


- Here is a plot of the size of the HTTP objects in bytes
- X-axis is the size
- Y-axis is the count, where we grouped the data into 8 KB buckets
- Log/Log scale

Traffic Size, CIC Data

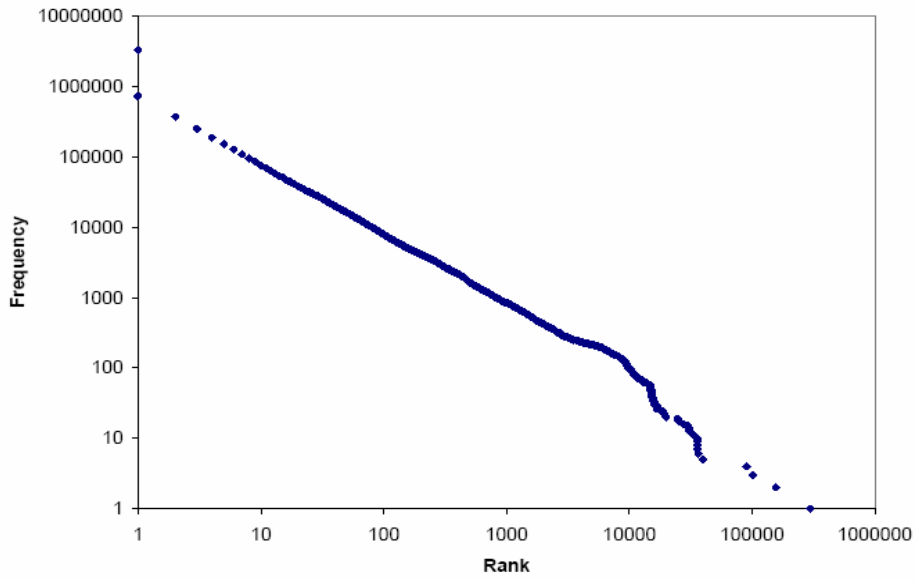


- When we examined MIME type, we found quite a few media files
- Media is video, music and flash animations.
- Somewhat surprising given how long it takes to download in the centers



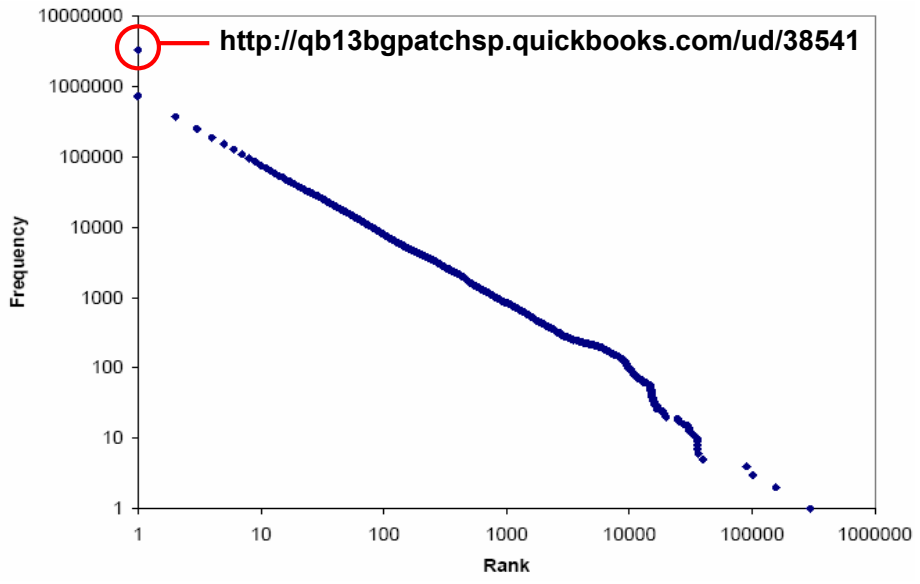
- Presence of large downloads
  - 60% of downloads greater than 1 MB – this takes a long time
  - Extreme outliers benefits very few users

Traffic Frequency, CIC Data



- Frequency – number of times a specific URL is accessed
- Rank – ordering of popularity of URL, 10^{th} rank is the 10^{th} most popular URL
- Log/Log scale
- Power law style distribution

Traffic Frequency, CIC Data



- One big exception, auto updater for an piece of accounting software.
- In this particular case, the URL was downloaded 125 times/hour.
  - Misconfiguration or too aggressive
  - Kill cost saving schemes such as dial on demand connections such as VSAT.

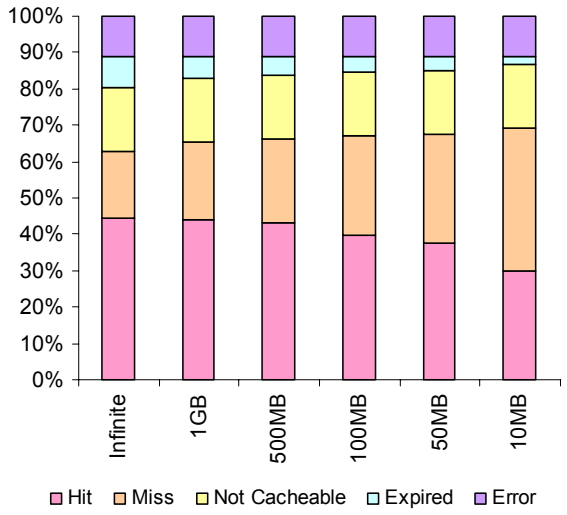
Overview

- Internet access in developing regions
- Web traffic properties
- Techniques for improving user experience

•Given the bandwidth constraints and web traffic properties, we now examine some techniques for improving user experience.

Caching

- Crawled 1 week of URLs in CIC trace
  - HTTP header pragmas
  - Errors in crawl treated as uncachable
- LRU cache simulation
  - Simple model – no browser caches
- Even small caches work well



•One question is how much plain caching would help in the situation

•We took a random week of CIC trace data

- •Recrawled the web objects and obtained their HTTP headers
- •Ran a LRU cache simulation
  - •Cache hits are **PINK at the bottom**
  - •Simple model – no multilevel caching due to interactions of browser and proxy caches

•Even small caches seem to work well

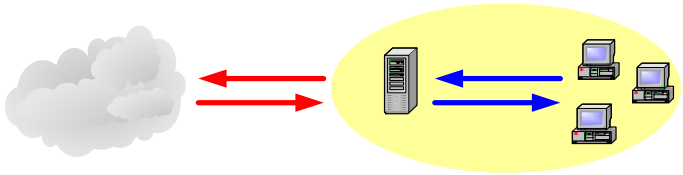
- •Perhaps due to small user population coming back to the centers

Offline Caching

- Use local cache to serve:
  - Items when disconnected
  - Expired items
- Spectrum of "Availability" versus "Freshness"
- Advantages:
  - Cost
  - Still useful when Internet is unavailable

- Serve items out of cache of the local proxy server without connecting to the Internet.
- Also, continue serving item out of cache even if it has from the pragma standpoint expired.
- **Brings up the "availability" vs. "freshness" of web content.** Generally speaking because of the bad link
  - Should be favoring availability of information vs.
  - Guarantee that its fresh
  - When link quality is good, availability is not a problem, however when link bad.
  - **Users are able to choose where on the spectrum they care about.**
- Can take advantage of the diversity of cost schedules to lower cost
- **Why may work well:**
  - Generally speaking, many websites update very frequently (e.g. portals large portion of traffic)
    - rotate advertising
    - provide hot off the wire news
    - At a time granularity that is not deemed useful by the user.

Offline Caching Experiment

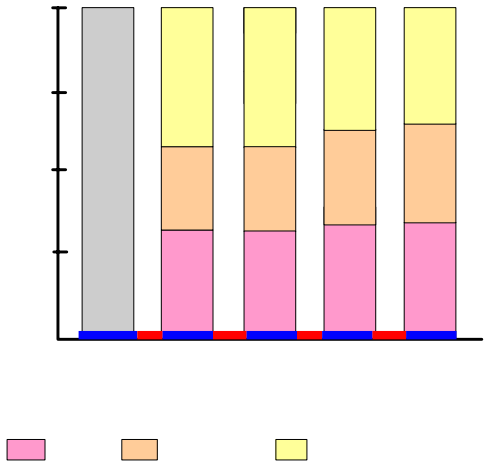


- **Connection Schedule:**
  - Operate disconnected during day, queuing requested URLs (**blue arrows**)
  - Fetch requested URLs at night (**red arrows**)
- Ran CIC 1 week of trace URLs
- Infinite cache size

- Operate disconnected during the day
- Synchronize the requested URLs at night.
- Local cache serves URLs available in cache, **even** if it is expired or otherwise invalid
- Ran CIC trace data for 1 week period of time with the above connection schedule
- Assumed an infinite cache size.

Offline Caching Experiment

- **Hit** – Item is fresh and up to date
- **Stale hit** - URL contents could be satisfied from cache ignoring all pragmas and expiration times
- **Miss/Uncacheable** – Item not in cache or labeled uncachable



•These were the results of the offline caching experiment

•X-axis is the time of day

•Y-axis is percentage of URLs and the responses from the cache

•Note the connection pattern illustrated by the alternating

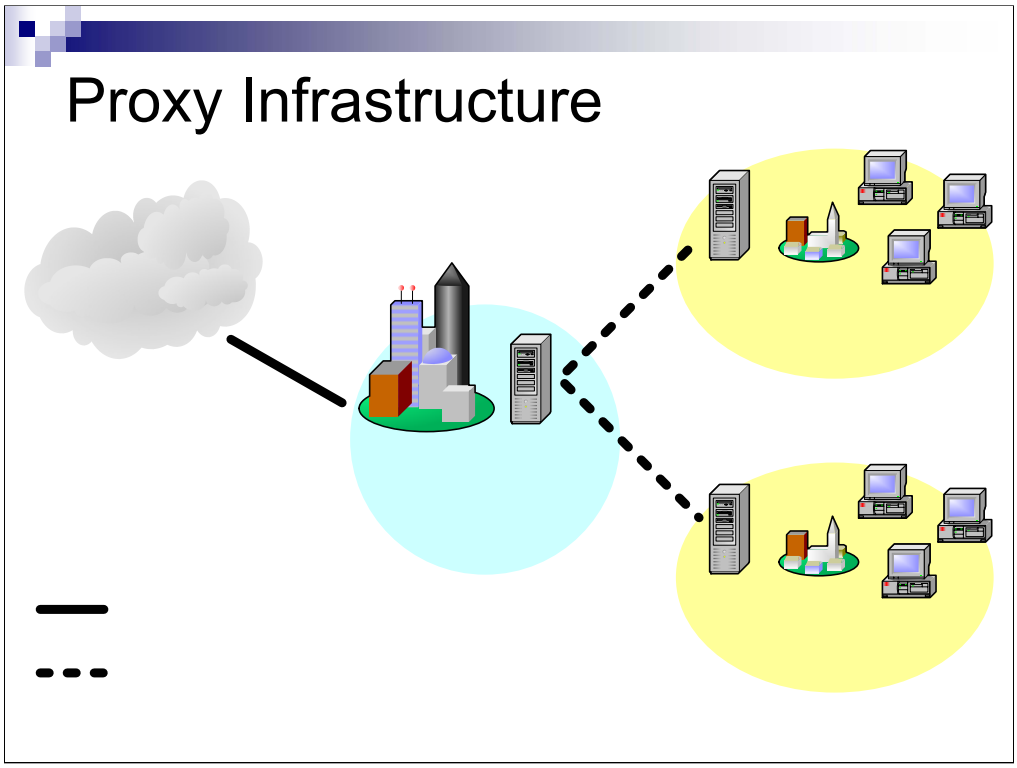
- •Disconnected portion (blue)
- •Synchronization portion (red)

•Stale hit

- •URL that could be satisfied from cache but under ordinary Internet caching rules is deemed to be invalid

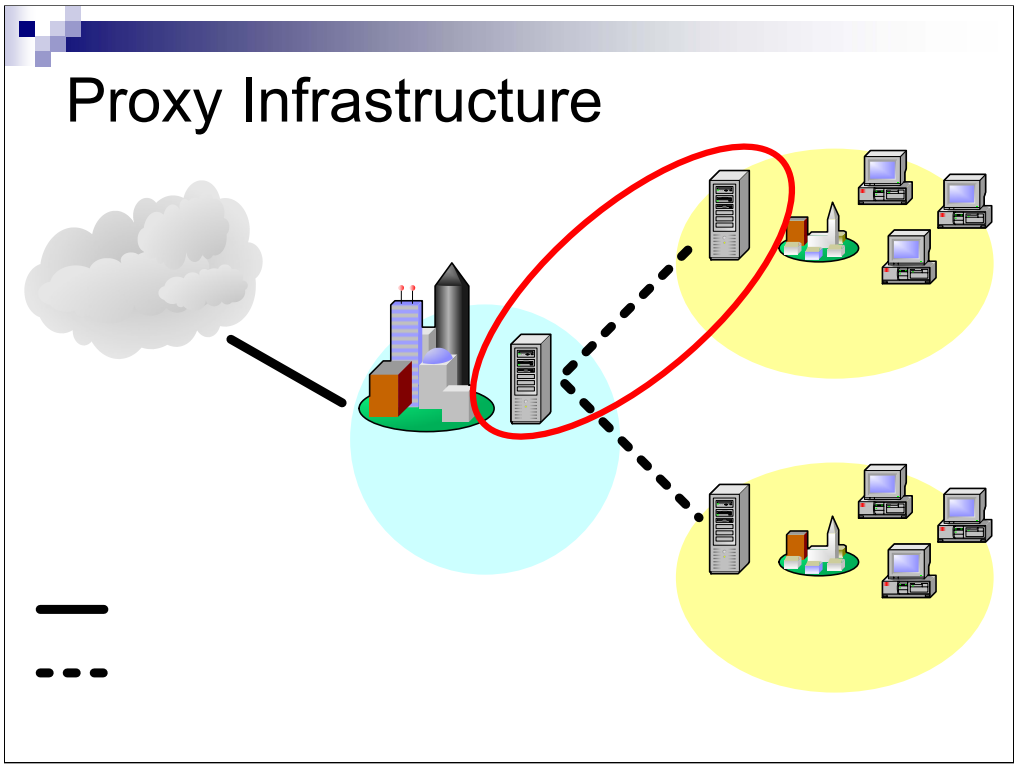
•2 take away points:

- •**Offline caching satisfies 70%.** Caveat is that websites are not designed for offline browsing.
- •**Using the stale hits (orange) under such an operating environment, we can just about double the amount of hits (pink) to the cache.**



- Next I am going to talk about more application specific optimization
- This is a generalized schematic of the structure of Internet access in Cambodia and Ghana,
- Rural internet center behind some constrained link to some well connected point urban access point
    - **Note urban area (in cyan) may not have to be located in country**. This schematic also applies to a rural center proxy communicating with a well connected server in the US

Internet



- Many of the techniques we describe next depend on a cooperating proxy infrastructure, circled in red.
  - Proxies servers lie on either end of bad link
  - Cooperation proxies can mask the bad link using different better adapted protocols other than HTTP
- Important to note that the link circled here is the important one
  - Bottleneck for access to the web.
  - This link is the one we want to optimize.

Internet

Application Specific Improvements

- **Web based e-mail**
  - Inefficient – ballpark overhead of most popular provider - 52 KB/message
  - Content not helped by caching
  - Better protocol can be used
- **Convert polling to push**
  - Data center polls and pushes content to proxies when changes occur
- **Replace irrelevant advertising**
  - Services and goods not available in local markets

•Web based e-mail can be replaced with a more efficient e-mail protocol

- •We measured the e-mail overhead of the most popular mail provider in the traces 52 kb of overhead per message view
- •For ex: **just html overhead is 3 gb of data over the lifetime of the Cambodian trace**
- •One thing to note is that web applications such as e-mail don't cache that well because of private data

•Another application specific improvement that can be done is for the proxy servers to convert polling to push behavior

- •Essentially the well connected data center will poll for the rural center
- •Push content when it changes

•Finally, it would be great to be able to replace the irrelevant advertising in websites

- •Scheme to do so at the local proxy cache

Proxy Infrastructure

- **Time shifting**
  - Move large downloads to cheaper connection times
- **Compression**
  - LZ results in 1/3 reduction of data
  - Delta coding
- **Transcoding**
  - Increase information content per byte
  - TEK (http://tek.sourceforge.net)
  - LoBand (http://www.loband.org)

•Other optimizations enabled by the proxy infrastructure.

- •Time shifting is traffic shaping for downloads
- •Compression
- •Semantic level compression: transcoding

Conclusion

- Web in developing countries is limited by cost and quality of connectivity
- Many traditional approaches can be used
- More novel approaches:
  - Availability vs. Freshness
  - Cooperating Proxies

- Thanks to Asia Foundation, Busy Internet, Pauline Tweedie and R.J. Honicky for their help in obtaining the data.


TIER research group website:
http://tier.cs.berkeley.edu