On-the-fly Inter-proxy Data Compression for Web Access

Pradeepa Gurusinghe,
Gihan Dias,
Vishaka Nanayakkara,

University of Moratuwa
Sri Lanka.
Overview

- Motivation
- Overview of the system
- Analysis
- Content-type Compression
- HTTP Header Compression
- Permanent TCP connections
- Further Work
Motivation for a new system

- Need to access the Web with a satisfactory performance level
  - Fast web access
  - At a lower cost

![Diagram showing network architecture and bandwidth limitations.](image-url)
How to solve the problem?

- Increase the international bandwidth
- Per-network and per-user Bandwidth allocation
- Dynamic bandwidth negotiation
- Bandwidth allocation on user request
- Data compression
Overview – System

In this link, there is a limited and very expensive bandwidth compared to local area network.
Overview – System Operations

Browser Request

- Open a TCP Connection to Upstream Proxy

Downstream Proxy

- Decompression

- Decompressed Data

Upstream Proxy

- Open a TCP Connection to Web server

Compression

- Compressed Data

- Normal Data

Response
# The Content Types

<table>
<thead>
<tr>
<th>Content-type</th>
<th>Extension</th>
</tr>
</thead>
<tbody>
<tr>
<td>text/html</td>
<td>html, htm</td>
</tr>
<tr>
<td>text/plain</td>
<td>txt, c, c++, pl, cc, h</td>
</tr>
<tr>
<td>text/css</td>
<td>css</td>
</tr>
<tr>
<td>image/gif</td>
<td>gif</td>
</tr>
<tr>
<td>image/jpeg</td>
<td>jpeg, jpg, jpe</td>
</tr>
<tr>
<td>audio/x-wav</td>
<td>wav</td>
</tr>
<tr>
<td>video/mpeg</td>
<td>mpeg, mpg, mpe</td>
</tr>
<tr>
<td>application/pdf</td>
<td>pdf</td>
</tr>
</tbody>
</table>
Distribution of Content-types

Text Content-types

- text/html: 30%
- text/plain: 5%
- application/javascript: 2%
- text/css: 1%
- application/octet-stream: 17%
- image/jpeg: 12%
- image/gif: 5%
- application/vnd.ms-powerpoint: 5%
- video/mpeg: 4%
- application/x-zip-compressed: 3%
- application/msword: 2%
- text/html: 2%
- application/octet-stream: 2%
- application/x-shockwave-flash: 1%
- other: 1%

Image Content-types

- image/jpeg: 38%
- image/gif: 29%
- application/odt: 17%
- application/pdf: 12%
- application/octet-stream: 1%

University of Moratuwa – Web traffic
Compression Techniques

- **Lossless**
  - The Shannon-Fano
  - Lempel-Ziv Coding
  - Huffman Coding

- **Lossy**
  - Fractal compression
  - JPEG
  - Wavelet compression

- *for text content*
- *for image content*
Compression Ratio – Text Contents

Compression Ratio – avg. 70%

File Size ---->>

gzip  bzip2  lzop
Compression Time – Text Contents

File Size -->

Time/Size (s/kB)

- gzip
- bzip2
- lzop

Pradeepa Gurusinghe
Text content-type compression

- gzip, bzip2 and lzop
- All use lossless compression algorithms
- The best – gzip
  - Compression Ratio
  - Time
- zlib Compression Library
Image/JPEG

- Compressed file format – Lossy method
- Stands for Joint Photographic Experts Group
- Millions of colors, while GIF supports 256 colors
- Better for photographs, nature sceneries
- Allows users to control the quality by varying compression
Further compressing JPEG

- Using JPEG Algorithms
- Quality reduction is not noticeable to human eye.
- Better to compress using original file to a low quality level
- At congested time
  - JPEG files can be compressed dropping certain non-essential headers, If it is acceptable to have a reduction of image quality
Original 11628B

Q-60%  - 9016B

Q-50%  - 8047B

Q-30%  - 5925B

Q-10%  - 3345B

Q-02%  - 1808B
Image/gif

- Compressed file format – Lossless method
- Stands for Graphics Interchange File format
- Better for cartoons, line drawings
  - Low colour combinations
- Further compressing image/gif
  - Lossy compression
    - By reducing colors
- Yet to be implemented
HTTP Header

- Two types
  - Request Header (Send by web browser to web server)
  - Response Header (Send by web server to web browser)
- Text data records
HTTP Header Compression

- More than 400 bytes in size
- But, large number of requests * 400 Bytes
  - Significant!!
  - Eg. www.cmb.ac.lk 30 requests * 400 Bytes * 2
    24Kb for the one user request
- Text → highly compressible
- Same set of words
- A preset dictionary
  - Leads to a high compression ratio (avg. 70%)
- Improved dictionary – dynamic change
  - Yet to be implemented
Statistics for Header Compression

- Improved Dictionary (Static)
- Normal Pre-set Dictionary
- Without a Dictionary

Header Sequence Number

Compression Ratio
Inter-proxy Permanent Connections

- Time critical
- Slow link between two proxies - Take considerable amount of time
  - To open a connection
- HTTP 1.1 provides persistency
  - Only for few number of HTTP requests
- By keeping permanent connections between proxies, and using them as required, better performance can be gained.
Summary

- Selective compression
  - Text/plain, text/html → zlib
  - HTTP Headers → zlib with preset dictionary
  - JPEG → To a low quality jpeg using libjpeg

- Header compression

- No involvement of browser or web server
Further Work…

- Compressing next major content-type, image/gif
  - By reduction of colors
- The method to retrieve the original image
  - Refresh of the browser meant to the proxy that “I want the original image”
- Eliminating the Inverse Discrete Cosine Transform (IDCT) and DCT operations beyond the jpeg image compression
- Performance analysis by running the system in a production environment
Thank You!

Email: pradeepa@cse.mrt.ac.lk