Network Caching Resource Allocation for Multimedia Objects

Michael Kozuch, Wayne Wolf, and Andrew Wolfe
Princeton University
{kozuch, wolf, awolfe}@ee.princeton.edu

With the rapid growth of the World Wide Web, demand for internet bandwidth has risen sharply [1]. This demand for data naturally creates network congestion and tends to increase retrieval latency. In order to reduce the demand for bandwidth and to improve retrieval latency, several agencies have begun to implement network caching for the internet (examples may be found in [2], [3], and [4]).

The main idea in network caching is to create copies of network accessible documents which are “closer” (either physically, with respect to transmission cost, or with respect to latency) than the home location of the document. The current efforts have largely been based on demand placement techniques such as LRU. That is, when a user requests a document, a copy of that document is cached in the closest cache to the user. While this may be the proper approach for moderate size data objects such as a typical HTML document, a more conservative approach may be warranted for large objects such as video data. Indeed, in this work, the authors propose an algorithm for allocating network storage and communication resources for large multimedia objects.

The size of an object is significant for two related reasons. The first reason is that miscaching a large object incurs a greater real cost than caching a smaller object. When one considers that a real cost is associated with disk space ($/MB/month), the penalty incurred by miscaching a gigabyte size object becomes obvious. The second reason is that

the costs associated with “intelligent caching” become more tolerable when considering large objects such as video objects rather than small objects such as HTML pages. For example, the space required to maintain access history for an object is relatively independent of the object size, and therefore, the meta-data space required for n large objects represents a smaller fraction of the usable storage space than the meta-data space required for n smaller objects.

The utility of network caching becomes clear through straightforward cost analysis considerations. The cost of internet-class bandwidth has been reported to be decreasing at a rate of 30% per year [5]. Hard disk prices are also decreasing at an exponential rate [6]. However, in the case of disk prices, the rate is 50% per year. Considering the current cost of bandwidth and disk prices, we see that caching an object is beneficial if it is accessed just a handful of times per month [7]. Furthermore, because disk storage cost is decreasing more rapidly than network cost, network caching will become more beneficial in the future.

In order to derive maximum utility from network caching however, the network system must determine both the appropriate objects to cache and also *where* to cache those objects. Determining the location of cached copies has been termed the File Allocation Problem and shown to be NP-complete for general networks [8]. However, the authors have developed a solution for tree networks whose worst-case execution time is $\Theta(N^2)$. An important feature of this algorithm is that it is *distributed*. That is, a globally optimal solution is achieved without centralized computation. This is possible because each node makes a number of local calculations and then forwards the results to its parent node. We have instituted a parent-child hierarchy to mimic the routing techniques employed in the internet. Each multimedia object considered is assigned to a home node. Each client, then,

may query caches on the path from the client to the home node. If a cached copy is found along that path, that copy is returned. Otherwise, the client fetches the object from the home node.

In conclusion, the authors have developed an optimal network caching algorithm. Further, this algorithm is distributed and behaves with worst-case execution time which is $\Theta(N^2)$.

1.0 References

1. Kimberly C. Claffy, Hans-Werner Braun, and George C. Polyzos, “Tracking Long-Term Growth of the NSFNET,” *Communications of the ACM*, Vol. 37, No. 8, pp. 34-45, August 1994.
2. “A Distributed Testbed for National Information Provisioning,” http://www.nlanr.net/Cache/.
3. Neil Smith, “The UK National Web Cache - The State of the Art,” *Fifth International World Wide Web Conference*, http://www5conf.inria.fr/fich_html/papers/P45/Overview.html, May 6-10, 1996.
4. Donald Neal, “The Harvest Object Cache in New Zealand,” *Fifth International World Wide Web Conference*, http://www5conf.inria.fr/fich_html/papers/P46/Overview.html, May 6-10, 1996.
5. Jeffrey K. MacKie-Mason and Hal R. Varian, “Pricing the Internet,” *Public Access to the Internet*, May 1993.
6. David A. Patterson and John L. Hennessy, *Computer Architecture: A Quantitative Approach*, Second Edition, 1996, Morgan Kaufmann Publishers, Inc.
7. Michael Kozuch, Wayne Wolf, and Andrew Wolfe, “An Approach to Network Caching,” Princeton University Computer Engineering Technical Report CE-W96-39.
8. Kapali P. Eswaran, “Placement of Records in a File and File Allocation in a Computer Network,” *Information Processing ‘74*, pp. 304-307, 1974.