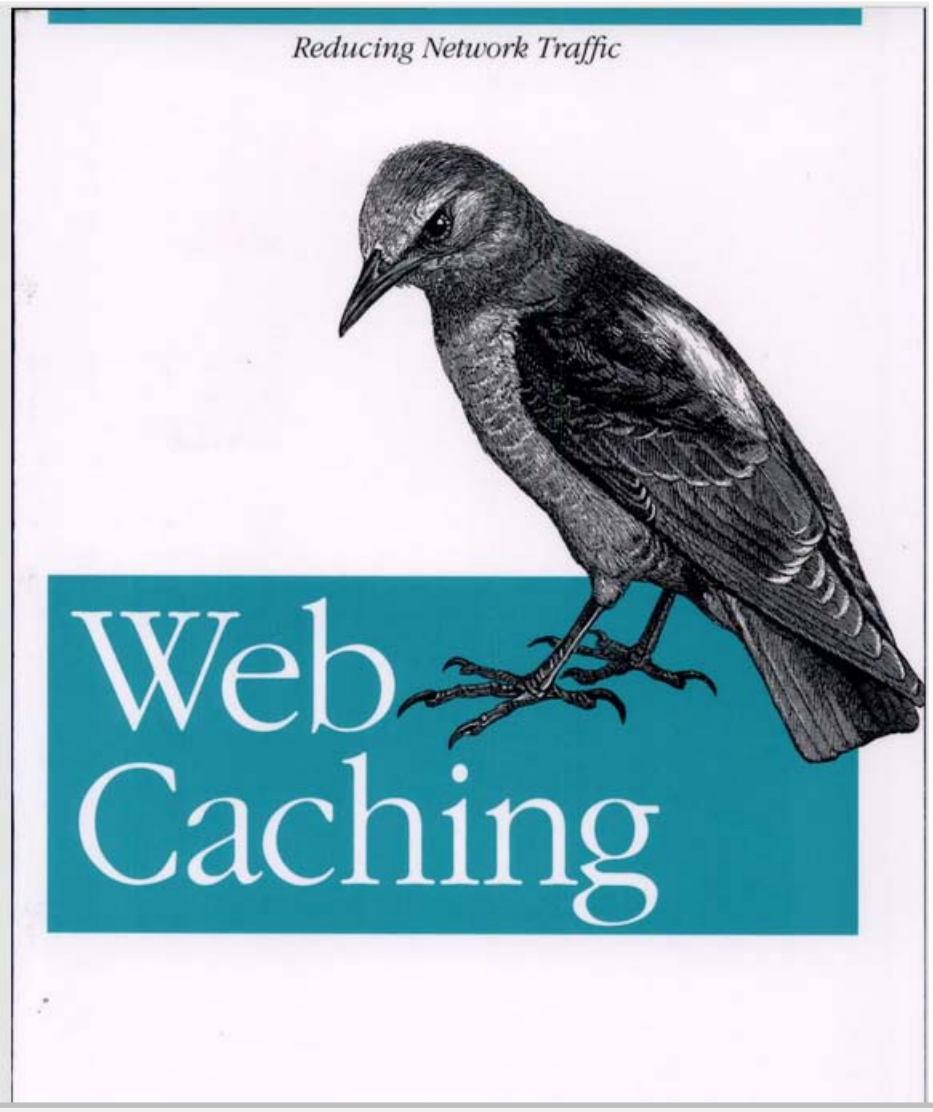

Web Caching

Duane Wessels

O'REILLY®

Web Caching
by Duane Wessels

Published by O'Reilly & Associates, Inc., 101 Morris Street, Sebastopol, CA 95472.

Editors: Nathan Torkington and Paula Ferguson

Production Editor: Leanne Clarke Soylemez

Cover Designer: Edie Freedman

Printing History:

June 2001: First Edition.

Library of Congress Cataloging-in-Publication Data

Wessels, Duane.
Web Caching/Duane Wessels
p. cm.
ISBN 1-56592-536-X
1. Cache memory. 2. Browsers (Computer programs) 3. Software configuration management. 4. World Wide Web. I. Title.

TK7895.M4 W45 2001
004.5'3--dc21 2001033173

ISBN: 1-56592-536-X

they access it. It's much more efficient to transfer the page once, cache it, and then serve future requests directly from the cache.

In order for caching to be effective, the following conditions must be met:

- Client requests must exhibit locality of reference.
- The cost of caching must be less than the cost of direct retrieval.

We can intuitively conclude that the first requirement is true. Certain web sites are very popular. Classic examples are the starting pages for Netscape and Microsoft browsers. Others include searching and indexing sites such as Yahoo! and Altavista. Event-based sites, such as those for the Olympics, NASA's Mars Pathfinder mission, and World Cup Soccer, become extremely popular for days or weeks at a time. Finally, every individual has a few favorite pages that he or she visits on a regular basis.

It's not always obvious that the second requirement is true. We need to compare the costs of caching to the costs of not caching. Numerous factors enter into the analysis, some of which are easier to measure than others. To calculate the cost of caching, we can add up the costs for hardware, software, and staff time to administer the system. We also need to consider the time users save waiting for pages to load (latency) and the cost of Internet bandwidth.

Let's take a closer look at the three primary benefits of caching web content:

- To make web pages load faster (reduce latency)
- To reduce wide area bandwidth usage
- To reduce the load placed on origin servers

1.3.1 Latency

Latency refers to delays in the transmission of data from one point to another. The transmission of data over electrical or optical circuits is limited by the speed of light. In fact, electrical and optical pulses travel at approximately two-thirds the speed of light in wires and fibers. Theoretically, it takes at least 25 milliseconds to send a packet across the U.S. In practice, it takes a little longer, say about 30 milliseconds. Transoceanic delays are in the 100-millisecond range.

Another source of latency is network congestion. When network links are close to full utilization, packets experience queuing delays inside routers and switches. Queuing, which can occur at any number of points along a path, is occasionally a source of significant delay. When a device's queue is full, it is forced to discard incoming (or outgoing) packets. With reliable protocols, such as TCP, lost packets