Web Caching and Replication

by Michael Rabinovich and Oliver Spatscheck
Addison Wesley; 1st edition (2002)
392 pages, list price US$ 49.99
ISBN 0-201-61570-3

Review by:
Qiang Wang and Brian D. Davison
Department of Computer Science and Engineering, Lehigh University (PA, USA)
qiw3@lehigh.edu, davison@lehigh.edu

Introduction

As the Internet has become an essential part of everyday life, hundreds of millions of users now connect to the Internet. At the same time, more resource-hungry and performance-sensitive applications have emerged. Expectations of scalability and performance have made caching and replication common features of the infrastructure of the Web. By directing the workload away from possibly overloaded origin Web servers, Web caching and replication address Web performance and scalability from the client side and the server side, respectively. Caching stores a copy of data close to the data consumer (e.g., in a Web browser) to allow faster data access than if the content had to be retrieved from the origin server. Replication, on the other hand, creates and maintains distributed copies of content under the control of content providers. This is helpful because client requests can then be sent to the nearest and least-busy server. Moreover, value-added services, such as virus checking, Web page language translation, and automatic content adaptation for small handheld devices can be developed upon this platform. This book, by two leading researchers at AT&T Labs, Michael Rabinovich and Oliver Spatscheck, presents the state-of-the-art in Web caching and replication from and for both industry and academic research.

A Walk Through the Contents

Consisting of 331 pages of main text, the book is organized into four parts (Background, Web Caching, Web Replication, and Future Directions). These four parts are covered in 16 chapters, each concluding with a summary of the most important concepts.

The Introduction section at the beginning of the book is essentially a primer (only 11 pages) for the basics of the Web, Web caching, and replication. The Glossary at the end is a comprehensive collection of terminologies needed, with precise and easy-to-understand definitions. It is followed by a bibliography and an index. The index is particularly helpful since cross-referencing is frequently needed for serious readers.

The first four chapters of the Background (Part I) cover the basic concepts of the Internet model, IP and routing in the network layer, TCP in the transport layer, and DNS and HTTP protocols in the application layer. Chapter 5 elaborates on the most important HTTP features pertaining to caching and replication, making it easier to understand how HTTP supports caching and replication without consulting the whole complex protocol. Chapter 6 summarizes the "rules of thumb" of Web characteristics, which are fundamental for understanding how to optimize Web performance. Although inclusion of these chapters makes the book self-contained, some experience with networking and HTTP is particularly helpful. Many good books, for example, Kurose and Ross's *Computer Networking: A Top-Down Approach Featuring the Internet* (2nd ed., 2003) can serve this purpose.

Web caching and replication can be viewed as two different solutions that also share many similar concepts and technologies. The authors present the topic in two parts instead of organizing around technologies, making it easier to embrace the whole picture of the system. Part II (151 pages) is devoted to Web caching and Part III (71 pages) to Web replication. Cross-references are extensively used throughout the book for readers to consult related materials in different places.

Part II (Web caching) concentrates on forward proxies. Chapter 7 discusses the realistic benefits from forward proxies in latency reduction and bandwidth savings. Chapter 8 describes transparent and non-transparent deployments of forward proxies in ISPs and enterprise networks. Security and access control issues of proxy deployment is also discussed. Chapter 9 talks about different methods of

cooperative proxy caching, in which a set of proxies share their cached content with each other's clients. Location management and proxy pruning problems are discussed in detail, exemplified by an overview of existing platforms that have already been deployed on the Internet. Chapter 10 presents different validation and invalidation mechanisms used to make distributed copies of objects consistent with the origin object. Chapter 11 introduces replacement polices, which are used to decide which pieces of content should be replaced from the proxy's cache when no space is available to store additional objects. Chapter 12, contributed by Daniel Duchamp of Stevens Institute of Technology, concentrates on prefetching, a mechanism to perform work in anticipation of future needs. Popularity-based predictions, Markov modeling, and algorithms exploiting document structure are introduced. Chapter 13 discusses uncacheable content, delta encoding, active proxies, and cache-friendly Web development techniques.

Part III is dedicated to three issues surrounding Web replication: request distribution, content distribution, and server selection. Chapter 14 is devoted to basic mechanisms for request distribution: how to transparently distribute client requests to one among a set of servers that hold the content. Methods include content-blind and content-aware request distribution. Chapter 15 applies these mechanisms in a real-world platform for scalable content delivery—the content delivery network (CDN). Topics include DNS request distribution, streaming content delivery, secure content access, and data consistency. Chapter 16 focuses on how to select a server based on performance metrics (e.g. proximity metrics, load metrics, and aggregate metrics) of individual servers when redirecting a request.

Even in today's post-bubble economy, advances in Web caching and replication continue. Part IV (chapters 17 and 18) discusses future directions in value added services and content distribution internetworking. Examples of value-added services are presented, stimulating the reader to creatively consider other potential applications.

Targeted Audiences

The authors state "this book should be of interest to IT professionals, engineers at companies providing Internet services or equipment, and to researchers and graduate students in such fields as computer and information systems and networking." While not attempting to survey all related work, this book covers well the essential topics within Web caching and replication. As a result, it provides a solid foundation for researchers starting in the area, as well as for practitioners desiring an introduction to the field. The book can be used for advanced students as a text for upper-level and graduate courses in Internet technologies, Web performance and measurement, etc.

What's Missing?

Rabinovich and Spatscheck have succeeded in putting lots of content into a readable text with only 392 pages. However, given the breadth of topics covered, we'd love to see a few hundred additional pages to elaborate on some of them in a future edition. For example, chapter 11 (replacement polices) deserves more than five pages. Similarly, chapter 12 (prefetching) is an area closely related to data mining and machine learning, making full coverage of the topic a challenge in two dozen pages.

This book's primary focus is on distributed aspects of caching and replication, rather than individual components comprising the Web and the Internet. As a result, we recommend *Web protocols and Practice: HTTP/1.1, Networking Protocols, Caching, and Traffic Measurement* (2001) by Balachander Krishnamurthy and Jennifer Rexford, as a second resource for graduate level Web technology courses. A collection of recent papers is also a helpful supplement for such a graduate course.

Summary

In summary, this book is well organized and easy to read. Written by "one of the few researchers who are qualified to write this book" (Pei Cao, Cisco Systems), the information contained in the book is comprehensive, authoritative, and informative. It is particularly beneficial that the book discusses many aspects for a problem, as well as implications of and trade-offs between alternative approaches, leading to a deeper understanding of both concepts and applications. Illustrations and examples are used throughout, making complex technologies easier to understand. This is an excellent introductory book on Web caching and replication for students, engineers, and researchers.