Full text available at: http://dx.doi.org/10.1561/0100000068

Coding Techniques for Repairability in Networked Distributed Storage Systems

Coding Techniques for Repairability in Networked Distributed Storage Systems

Frédérique Oggier

Nanyang Technological University
Singapore
frederique@ntu.edu.sg

Anwitaman Datta

Nanyang Technological University
Singapore
anwitaman@ntu.edu.sg

the essence of knowledge

Boston – Delft

Foundations and Trends® in Communications and Information Theory

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
USA
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is F. Oggier and A. Datta, Coding Techniques for Repairability in Networked Distributed Storage Systems, Foundations and Trends® in Communications and Information Theory, vol 9, no 4, pp 383–466, 2012

ISBN: 978-1-60198-676-4

Foundations and Trends® in Communications and Information Theory
Volume 9 Issue 4, 2012

Editorial Board

Editorial Scope

Foundations and Trends® in Communications and Information Theory will publish survey and tutorial articles in the following topics:

- Coded modulation
- Coding theory and practice
- Communication complexity
- Communication system design
- Cryptology and data security
- Data compression
- Data networks
- Demodulation and Equalization
- Denoising
- Detection and estimation
- Information theory and statistics
- Information theory and computer science
- Joint source/channel coding
- Modulation and signal design
- Multiuser detection
- Multiuser information theory
- Optical communication channels
- Pattern recognition and learning
- Quantization
- Quantum information processing
- Rate-distortion theory
- Shannon theory
- Signal processing for communications
- Source coding
- Storage and recording codes
- Speech and Image Compression
- Wireless Communications

Information for Librarians
Foundations and Trends® in Communications and Information Theory, 2012, Volume 9, 4 issues. ISSN paper version 1567-2190. ISSN online version 1567-2328. Also available as a combined paper and online subscription.

Foundations and Trends® in
Communications and Information Theory
Vol. 9, No. 4 (2012) 383–466

DOI: 10.1561/0100000068

now
the essence of knowledge

Coding Techniques for Repairability in Networked Distributed Storage Systems*

Frédérique Oggier[1] and Anwitaman Datta[2]

[1] *Division of Mathematical Sciences, School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore, frederique@ntu.edu.sg*
[2] *Division of Computer Science, School of Computer Engineering, Nanyang Technological University, Singapore, anwitaman@ntu.edu.sg*

Abstract

This survey comprises a tutorial on traditional erasure codes and their applications to networked distributed storage systems (NDSS), followed by a survey of novel code families tailor made for better repairability in NDSS.

Keywords: Distributed Storage Systems, Erasure Codes, Repair.

* Portions of this survey, particularly the second part, was originally written as personal notes when we started to work on this topic, as an attempt to understand the big picture. The big picture was accordingly summarized at a very high level in a short survey [6]. The tutorial part on networked distributed storage systems and coding theory was added later, together with one code construction that we proposed, when these personal notes became lecture notes that were provided for the Open Phd program at Warsaw University and presented at a tutorial in ICDCN 2012. The current version is an updated version of these lecture notes, including technical details and taking into account some recent developments, as well as providing background context to make the manuscript self-contained. Very recent literature is skipped on purpose: it is both too difficult to keep track of all the papers, and too early to have a clear picture of what will be the most significant ones.

Contents

Part I

Background

1

Introduction

When communicating over an erasure channel, a transmitter typically adds redundancy to the message to be sent, so that the intended recipient can reconstruct the original message despite the loss of parts of the transmitted data. The mapping from the original message to its redundant version is referred to as encoding, and the challenge is to design an efficient coding scheme, that provides good protection against erasures at a low overhead.

Analogous problems arise in the context of data storage. Damages to the physical storage medium may make some bits/bytes unreadable and redundancy is needed to protect the stored data. For instance, a compact disc (CD) can often tolerate scratches thanks to the presence of a suitable coding technique, called Reed–Solomon codes [36]. Another example at the other end of the size spectrum of storage systems is a large-scale distributed system such as a data-center or a peer-to-peer (P2P) system with many storage devices, some of which may fail or become inaccessible, e.g., due to network problems. Redundancy is again needed for fault tolerance, so that the aggregate data stored in the system can be retrieved. Though coding is a way of handling failures in the aforementioned scenarios, the design of a good code naturally

depends on the peculiarities of the setting considered — thus, codes for magnetic medium, solid state devices, CD, disk arrays or distributed systems may aim for distinct desirable properties.

The most commonly deployed multi-storage device systems are RAID (Redundant Array of Independent/Inexpensive Disks) systems [32], which store the data across multiple disks, some of which containing the actual information, while the others provide fault-tolerance by storing redundancy. Furthermore, distributing the data over multiple storage disks may also help increase the throughput of reading data, thanks to the parallelization of disk accesses. RAID systems traditionally put the multiple storage disks within a single computing unit, making the internal distribution transparent both logically as well as physically for the end users. Currently, typical RAID configurations allow for two failures within a RAID unit, though configurations tolerating more failures have also been studied.

The idea of distributing data across multiple disks has been naturally extended to multiple storage nodes which are interconnected over a network, as we witness in data-centers, and some P2P storage systems. We call such systems networked distributed storage systems (NDSS), where the word "networked" insists on the importance of the network interconnect. It is worth recalling that the individual storage nodes in an NDSS may themselves be comprised of multi-disk RAID systems, whose storage disks may themselves employ some redundancy scheme for fault-tolerance of their physical medium. Thus, while redundancy is present at several layers of a large storage system, this survey only looks at redundancy through coding techniques at the highest level of abstraction, namely for NDSS — and do so in a manner agnostic of the lower layer details.

At the NDSS level, data stored in individual storage nodes may become unavailable due to various reasons. As pointed out earlier, either a storage node or the communication link to this node may fail, but these are not the only cases. In P2P settings, a user operating a storage node may just decide to make it offline temporarily, or leave the system permanently. Irrespective of the nature of the failure, redundancy is needed to ensure data availability. Depending on the nature of failure, the lost redundancy may also need to be replenished in order to

ensure long-term data durability. The simplest form of coding, namely replication, has been and still is a popular way to ensure redundancy in NDSS, due to its simplicity. However, given the sheer volume of data that needs to be stored and the overheads of replication, there has been in recent years an immense interest in using coding for NDSS among major commercial players such as Google and Microsoft [19] to name a few, an interest which has also been mirrored in the academic world.

The aim of this survey is to look at coding techniques for NDSS, which aim at achieving (1) fault tolerance efficiently and (2) good repairability characteristics to replenish the lost redundancy, and ensure data durability over time. We will like to make the following disclaimer about the scope of this survey. There are many other criteria (than repair) that may guide the design of codes for NDSS. There are also many other kind of performance issues (than repair) that still need to be studied for many of the codes that we summarize in this survey. We will however confine our discussions mainly to codes providing good repairability. Also, while we have tried to provide an overview of the most prominent code techniques representing different points in the code design space, our treatment of the subject is by no means exhaustive. We have both deliberately as well as out of our ignorance given the rapid pace of developments in the area, left out many works.

This survey is organized into two parts. The first part gives an overview of some basic concepts related to NDSS and provides a quick introduction to classical coding theory, concluding with a discussion of the pros and cons of using classical erasure codes for NDSS. Such a discussion leads us to the second part, where several new families of codes tailor made for NDSS repairability are described and reviewed. Since it is impossible to keep track of every single code construction proposed, we instead identify prominent design choices, which are described and illustrated respectively in Section 6 for a network coding approach, in Section 7 for combining two layers of erasure codes, and in Section 8 for codes aiming at local repairability.

References

[1] R. Alshwede, N. Cai, S.-Y. R. Li, and R. W. Yeung, "Network information flow," *IEEE Transactions on Information Theory*, vol. 46, no. 4, 2000.

[2] D. Beaver, S. Kumar, H. C. Li, J. Sobel, and P. Vajgel, "Finding a needle in Haystack: Facebooks photo storage," *OSDI*, 2010.

[3] R. Bhagwan, K. Tati, Y.-C. Cheng, S. Savage, and G. M. Voelker, "Total recall: System support for automated availability management," *USENIX Symposium on Networked Systems Design and Implementation (NSDI)*, 2004.

[4] R. E. Blahut, *Algebraic Codes for Data Transmission*. Cambridge.

[5] A. Datta and F. Oggier, "Redundantly grouped cross-object coding for repairable storage," *APSys*, 2012.

[6] A. Datta and F. Oggier, "An overview of codes tailor-made for better repairability in networked distributed storage systems," *SIGACT News*, vol. 44, no. 1, Available at http://arxiv.org/abs/1109.2317, March 2013.

[7] A. Datta, L. Pamies-Juarez, and F. Oggier, "On data insertion and migration in erasure-coding based large-scaled storage systems," to appear in *ICDCIT 2013*.

[8] A. G. Dimakis, P. B. Godfrey, Y. Wu, M. Wainwright, and K. Ramchandran, "Network coding for distributed storage systems," *IEEE Transactions on Information Theory*, vol. 56, no. 9, September 2010.

[9] A. G. Dimakis, K. Ramchandran, Y. Wu, and C. Suh, "A survey on network codes for distributed storage," *The Proceedings of the IEEE*, vol. 99, no. 3, March 2011.

[10] A. Duminuco and E. Biersack, "Hierarchical codes: How to make erasure codes attractive for peer-to-peer storage systems," *Eighth International Conference on In Peer-to-Peer Computing, P2P*, 2008.

[11] P. Elias, "Error-free coding," *Transactions on Information Theory*, vol. 4, no. 4, September 1954.
[12] K. S. Esmaili, L. Pamies-Juarez, and A. Datta, "The CORE storage primitive: Cross-object redundancy for efficient data repair & access in erasure coded storage," preprint, available at http://arxiv.org/abs/1302.5192.
[13] P. Gopalan, C. Huang, H. Simitci, and S. Yekhanin, "On the locality of codeword symbols," *Electronic Colloquium on Computational Complexity (ECCC)*, vol. 18, 2011.
[14] hadoop.apache.org/.
[15] H. D. L. Hollmann, "Storage codes — coding rate and repair locality," International Conference on Computing, Networking and Communications, available at http://arxiv.org/abs/1301.4300 (ICNC 2013).
[16] http://wiki.apache.org/hadoop/HDFS-RAID.
[17] Y. Hu, Y. Xu, X. Wang, C. Zhan, and P. Li, "Cooperative recovery of distributed storage systems from multiple losses with network coding," *IEEE Journal on Selected Areas in Communications*, vol. 28, no. 2, February 2010.
[18] C. Huang, M. Chen, and J. Li, "Pyramid codes: Flexible schemes to trade space for access efficiency in reliable data storage systems," *IEEE International Symposium on Network Computing and Applications*, NCA 2007, 2007.
[19] C. Huang, H. Simitci, Y. Xu, A. Ogus, B. Calder, P. Gopalan, J. Lin, and S. Yekhanin, "Erasure coding in windows azure storage,".
[20] Y. H. K. W. Shum, "Cooperative regenerating codes," preprint, available at http://arxiv.org/abs/1101.5257.
[21] A.-M. Kermarrec, N. Le Scouarnec, and G. Straub, "Repairing multiple failures with coordinated and adaptive regenerating codes," in *the Proceedings of the 2011 International Symposium on Network Coding (NetCod 2011)*.
[22] R. Koetter and M. Medard, "An algebraic approach to network coding," *IEEE/ACM Transactions on Networking*, vol. 11, 2001.
[23] D. Leong, A. G. Dimakis, and T. Ho, "Distributed storage allocations," *IEEE Transactions on Information Theory*, vol. 58, no. 7, pp. 4733–4752, July 2012.
[24] W. K. Lin, D. M. Chiu, and Y. B. Lee, "Erasure code replication revisited," *P2P*, 2004.
[25] F. Oggier and A. Datta, "Homomorphic self-repairing codes for agile maintenance of distributed storage systems," http://arxiv.org/abs/1107.3129.
[26] F. Oggier and A. Datta, "Byzantine fault tolerance of regenerating codes," *P2P*, 2011.
[27] F. Oggier and A. Datta, "Self-repairing codes for distributed storage — a projective geometric construction," *ITW*, 2011.
[28] F. Oggier and A. Datta, "Self-repairing homomorphic codes for distributed storage systems," *INFOCOM*, 2011.
[29] L. Pamies-Juarez, A. Datta, and F. Oggier, "RapidRAID: Pipelined erasure codes for fast data archival in distributed storage systems," http://arxiv.org/abs/1207.6744.
[30] L. Pamies-Juarez, F. Oggier, and A. Datta, "An empirical study of the repair performance of novel coding schemes for networked distributed storage systems," http://arxiv.org/abs/1206.2187.

[31] D. S. Papailiopoulos and A. G. Dimakis, "Locally repairable codes," *IEEE International Symposium on Information Theory (ISIT)*, Available at http://arxiv.org/abs/1206.3804, 2012.

[32] D. A. Patterson, G. Gibson, and R. H. Katz, "A case for redundant arrays of inexpensive disks (RAID)," *ACM SIGMOD International Conference on Management of Data*, 1988.

[33] S. Pawar, S. El Rouayheb, and K. Ramchandran, "Securing dynamic distributed storage systems against eavesdropping and adversarial attacks," *IEEE Transactions on Information Theory (Special Issue on Facets of Coding Theory: from Algorithms to Networks)*, vol. 57, no. 9, September 2011.

[34] K. V. Rashmi, N. B. Shah, P. Vijay Kumar, and K. Ramchandran, "Explicit construction of optimal exact regenerating codes for distributed storage," *Allerton*, 2009.

[35] A. S. Rawat and S. Vishwanath, "On locality in distributed storage systems," *ITW*, 2012.

[36] I. Reed and G. Solomon, "Polynomial codes over certain finite fields," *Journal of the Society of Industrial and Applied Mathematics*, 1960.

[37] K. W. Shum, "Cooperative regenerating codes for distributed storage systems," *ICC*, available at http://arxiv.org/abs/1101.5257, 2011.

[38] E. Stefanov, M. van Dijk, A. Oprea, and A. Juels, "Iris: A scalable cloud file system with efficient integrity checks," Cryptology ePrint Archive, Report 2011/585, 2011.

[39] C. Tian, "Rate region of the (4, 3, 3) exact-repair regenerating codes," preprint.

[40] Y. Wu and A. G. Dimakis, "Reducing repair traffic for erasure coding-based storage via interference alignment," *ISIT*, 2009.

[41] www.cleversafe.com.

[42] www.wuala.com/.

[43] R. W. Yeung, S.-Y. R. Li, N. Cai, and Z. Zhang, "Network coding theory: Single sources," *Foundations and Trends in Communication and Information Theory*, vol. 2, no. 4.