Annotated Bibliography for the Tutorial on Erasure Codes for Storage Systems

James S. Plank and Cheng Huang

Usenix FAST, February 12, 2013

General Coding Theory (not specific to storage systems)

I'm sorry that there is not a good general-purpose tutorial writeup on coding for storage. These books are about classic coding theory, and at times it's difficult to figure out how to apply them to storage.

- J. H. van Lint, **Introduction to Coding Theory,** Springer-Verlag, New York, 1982.
- F. J. MacWilliams and N. J. A. Sloane, **The Theory of Error-Correcting Codes, Part I,** North-Holland Publishing Company, Amsterdam, New York, Oxford, 1977.
- T. K. Moon, **Error Correction Coding: Mathematical Methods and Algorithms,** John Wiley & Sons, New York, 2005.
- W. W. Peterson and E. J. Weldon, Jr., **Error-Correcting Codes, Second Edition,** The MIT Press, Cambridge, Massachusetts, 1972.
- J. L. Hafner, V. Deenadhayalan, K. K. Rao and A. Tomlin, "*Matrix Methods for Lost Data Reconstruction in Erasure Codes,*" **FAST-2005: 4th Usenix Conference on File and Storage Technologies**, San Francisco, December, 2005, pp. 183-196. I include this because it gives a nice treatment of recovering from bit-matrix codes which is aimed at storage people rather than coding theorists.

Galois Field Arithmetic

Look for a complete treatment of Galois Field Arithmetic by Plank, Greenan, Miller and Houston, coming to a journal near you (posted on my web site), probably around April of this year. Until that comes out, the following papers are very helpful:

- J. S. Plank, "*A Tutorial on Reed-Solomon Coding for Fault-Tolerance in RAID-like Systems,*" **Software -- Practice & Experience**, 27(9), September, 1997, pp. 995-1012. http://web.eecs.utk.edu/~plank/plank/papers/CS-96-332.html. Of course, I'm partial to my work, but the tutorial walks you through Galois Field arithmetic, and how to implement it using discrete logarithm tables.
- K. Greenan, E. Miller and T. J. Schwartz, "*Optimizing Galois Field Arithmetic for Diverse Processor Architectures and Applications,*" **MASCOTS 2008: 16th IEEE Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems**, Baltimore, MD, September, 2008.
- J. Luo, K. D. Bowers, A. Oprea and L. Xu, "*Efficient Software Implementations of Large Finite Fields $GF(2^n)$ For Secure Storage Applications,*" **ACM Transactions on Storage**, 8(2), February, 2012.

- J. S. Plank, K. M. Greenan and E. L. Miller, "*Screaming Fast Galois Field Arithmetic Using Intel SIMD Instructions,*"**FAST-2013: 11th Usenix Conference on File and Storage Technologies**, San Jose, February, 2013.

- H. P. Anvin, "*The Mathematics of RAID-6,*", http://kernel.org/pub/linux/kernel/people/hpa/raid6.pdf. The first version of this was posted in 2004. It gets revised periodically, and the latest revision is December, 2011. This walks you through the technique to multiply a region of bytes by two extremely fast.

Reed Solomon Coding

- J. S. Plank, "*A Tutorial on Reed-Solomon Coding for Fault-Tolerance in RAID-like Systems,*" **Software -- Practice & Experience**, 27(9), September, 1997, pp. 995-1012. http://web.eecs.utk.edu/~plank/plank/papers/CS-96-332.html. Please note the correction in http://web.eecs.utk.edu/~plank/plank/papers/CS-03-504.html.

- I. S. Reed and G. Solomon, "*Polynomial codes over certain finite fields,*" **Journal of the Society for Industrial and Applied Mathematics**, 8, 1960, pp. 300-304. (No one ever reads this, but when do you get to cite papers from 1960?).

- M. O. Rabin, "*Efficient Dispersal of Information for Security, Load Balancing, and Fault Tolerance,*" **Journal of the Association for Computing Machinery**, 36(2), April, 1989, pp. 335-348.

Cauchy Reed Solomon Coding

- J. Blomer, M. Kalfane, M. Karpinski, R. Karp, M. Luby and D. Zuckerman, "*An XOR-Based Erasure-Resilient Coding Scheme,*" Technical Report TR-95-048, International Computer Science Institute, August, 1995.

- J. S. Plank and L. Xu, "*Optimizing Cauchy Reed-Solomon Codes for Fault-Tolerant Network Storage Applications,*" **NCA-06: 5th IEEE International Symposium on Network Computing Applications**, Cambridge, MA, July, 2006. http://web.eecs.utk.edu/~plank/plank/papers/NCA-2006.html.

- J. Luo, L. Xu and J. S. Plank, "*An Efficient XOR-Scheduling Algorithm for Erasure Codes Encoding,*" **DSN-2009: The International Conference on Dependable Systems and Networks**, IEEE, Lisbon, Portugal, June, 2009. http://web.eecs.utk.edu/~plank/plank/papers/DSN-2009.html.

Linux RAID-6

- H. P. Anvin, "*The Mathematics of RAID-6,*", http://kernel.org/pub/linux/kernel/people/hpa/raid6.pdf. The first version of this was posted in 2004. It gets revised periodically, and the latest revision is December, 2011.

EVENODD

- M. Blaum, J. Brady, J. Bruck and J. Menon, "*EVENODD: An Efficient Scheme for Tolerating Double Disk Failures in RAID Architectures,*" **IEEE Transactions on Computing**, 44(2), February, 1995, pp. 192- 202.

RDP

- P. Corbett, B. English, A. Goel, T. Grcanac, S. Kleiman, J. Leong and S. Sankar, "*Row Diagonal Parity for Double Disk Failure Correction,*" **3rd Usenix Conference on File and Storage Technologies**, San Francisco, CA, March, 2004.

- L. Xiang, Y. Xu, J. C. S. Lui and Q. Chang, "*Optimal Recovery of Single Disk Failure in RDP Code Storage Systems,*" **ACM SIGMETRICS**, June, 2010.

- G. Wang, X. Liu, S. Lin, G. Xie and J. Liu, "*Generalizing RDP Codes Using the Combintorial Method,*" **NCA-08: 7th IEEE International Symposium on Network Computing Applications**, Cambridge, MA, 2008. This is not an easy read, but these guys are really clever.

X-Code and some other Vertical Codes

- L. Xu and J. Bruck, "*X-Code: MDS Array Codes with Optimal Encoding,*" **IEEE Transactions on Information Theory**, 45(1), January, 1999, pp. 272-276.

- C. Jin, H. Jiang, D. Feng and L. Tian, "*P-Code: A New RAID-6 code with optimal properties,*" **23rd International Conference on Supercomputing**, New York, June, 2009. This is another vertical code with the same properties as the X-Code. It's just a different construction.

- L. Xu, V. Bohossian, J. Bruck and D. Wagner, "*Low Density MDS Codes and Factors of Complete Graphs,*" **IEEE Transactions on Information Theory**, 45(6), September, 1999, pp. 1817-1826. And another vertical RAID-6 code.

- V. Bohossian and J. Bruck, "*Shortening Array Codes and the Perfect 1-Factorization Conjecture,*" **IEEE International Symposium on Information Theory**, Seattle, 2006, pp. 2799-2803. This paper is super-cool, because it shows how to shorten a vertical code. Not for the feint of heart!

- J. L. Hafner, "*WEAVER Codes: Highly Fault Tolerant Erasure Codes for Storage Systems,*" **FAST-2005: 4th Usenix Conference on File and Storage Technologies**, San Francisco, December, 2005, pp. 211-224. These are non-MDS, vertical codes that have some nice properties.

- J. L. Hafner, "*HoVer Erasure Codes for Disk Arrays,*" **DSN-2006: The International Conference on Dependable Systems and Networks**, IEEE, Philadelphia, June, 2006. These are a combination of vertical and horizontal codes. Not MDS, but close. Jim gives nice and complete treatemnt of these codes.

Minimum Density RAID-6 Codes

- J. S. Plank, A. L. Buchsbaum and B. T. Vander Zanden, "*Minimum Density RAID-6 Codes,*" **ACM Transactions on Storage**, 6(4), May, 2011. This paper describes the three codes: Blaum-Roth, Liberation and Liber8tion.

- M. Blaum and R. M. Roth, "*On Lowest Density MDS Codes,*" **IEEE Transactions on Information Theory**, 45(1), January, 1999, pp. 46-59. This is the original Blaum-Roth code paper. You gotta bring your A-game when you try to read this paper.

Generalized EVENODD and RDP

The first two papers are a bit hard to read if you're not a coding theorist. Jerasure 2.0 (which should be released in August) will have an explanation of these codes and support for them so that you can use them without really understanding them.

- M. Blaum, J. Bruck and A. Vardy, "*MDS array codes with independent parity symbols,*" **IEEE Transactions on Information Theory**, 42(2), February, 1996, pp. 529- 542. (This is generalized EVENODD)

- M. Blaum, "*A Family of MDS Array Codes with Minimal Number of Encoding Operations,*" **IEEE International Symposium on Information Theory**, Seattle, September, 2006. (This is generalized RDP).

- C. Huang and L. Xu, "*STAR: An Efficient Coding Scheme for Correcting Triple Storage Node Failures,*" **IEEE Transactions on Computers**, 57(7), July, 2008, pp. 889-901. This paper describes the *m=3* version of the EVENODD code, paying particular attention to decoding.

- C. Huang and L. Xu, "*Decoding STAR Code for Tolerating Simultaneous Disk Failure and Silent Errors,*" **DSN-2010: The International Conference on Dependable Systems and Networks**, IEEE, Chicago, June, 2010.

Heuristics for reducing the XOR's in bit-matrix coding and decoding

- J. S. Plank, C. D. Schuman and B. D. Robison, "*Heuristics for Optimizing Matrix-Based Erasure Codes for Fault-Tolerant Storage Systems,*" **DSN-2012: The International Conference on Dependable Systems and Networks**, IEEE, Boston, MA, June, 2012. http://web.eecs.utk.edu/~plank/plank/papers/DSN-2012.html.

- J. L. Hafner, V. Deenadhayalan, K. K. Rao and A. Tomlin, "*Matrix Methods for Lost Data Reconstruction in Erasure Codes,*" **FAST-2005: 4th Usenix Conference on File and Storage Technologies**, San Francisco, December, 2005, pp. 183-196.

- C. Huang, J. Li and M. Chen, "*On Optimizing XOR-Based Codes for Fault-Tolerant Storage Applications,*" **ITW'07, Information Theory Workshop**, IEEE, Tahoe City, CA, September, 2007, pp. 218-223.

- J. S. Plank, "*XOR's, Lower Bounds and MDS Codes for Storage,*" **IEEE Information Theory Workshop (ITW)**, Paraty, Brazil, October, 2011. This is a small paper that makes the observation that you can encode and decode with fewer than *(k-1)* XOR's per encoded and decoded symbol.

Open Source Support

- Onion Networks, "*Java FEC Library v1.0.3,*" Open source code distribution:

http://onionnetworks.com/fec/javadoc/, 2001.

- L. Rizzo, "*Erasure codes based on Vandermonde matrices,*" Gzipped **tar** file posted at http://planete-bcast.inrialpes.fr/rubrique.php3?id_rubrique=10, 1998.

- Z. Wilcox-O'Hearn, "*Zfec 1.4.0,*" Open source code distribution: http://pypi.python.org/pypi/zfec, 2008. This puts python wrappers around Rizzo's code.

- A. Partow, "*Schifra Reed-Solomon ECC Library,*" Open source code distribution: http://www.schifra.com/downloads.html, 2000-2007.

- J. S. Plank, K. M. Greenan, E. L. Miller and W. B. Houston, "*GF-Complete: A Comprehensive Open Source Library for Galois Field Arithmetic,*" Technical Report UT-CS-13-703, University of Tennessee, January, 2013. http://web.eecs.utk.edu/~plank/plank/papers/CS-13-703.html.

- J. S. Plank, S. Simmerman and C. D. Schuman, "*Jerasure: A Library in C/C++ Facilitating Erasure Coding for Storage Applications - Version 1.2,*" Technical Report CS-08-627, University of Tennessee, August, 2008. http://web.eecs.utk.edu/~plank/plank/papers/CS-08-627.html.

- Jerasure version 2 should be done in August.

- J. S. Plank, "*Uber-CSHR and X-Sets: C++ Programs for Optimizing Matrix-Based Erasure Codes for Fault-Tolerant Storage Systems,*" Technical Report CS-10-665, University of Tennessee, December, 2010. http://web.eecs.utk.edu/~plank/plank/papers/CS-10-665/index.html.

Non-MDS Codes with Efficient Repair

- J. L. Hafner, "*WEAVER Codes: Highly Fault Tolerant Erasure Codes for Storage Systems,*" **FAST-2005: 4th Usenix Conference on File and Storage Technologies**, San Francisco, December, 2005, pp. 211-224.

- C. Huang, M. Chen and J. Li, "*Pyramid Codes: Flexible Schemes to Trade Space for Access Efficiency in Reliable Data Storage Systems,*" **ACM Transactions on Storage**, 9(1), February, 2013. Comment: Maximally Recoverable (MR) codes defined. Generalized Pyramid Codes (GPC) are MR codes.

- C. Huang, H. Simitci, Y. Xu, A. Ogus, B. Calder, P. Gopalan, J. Li and S. Yekhanin, "*Erasure Coding in Windows Azure Storage,*" **USENIX Annual Technical Conference**, Boston, June, 2012. Comment: Local Reconstruction Codes (LRC) are MR codes.

- P. Gopalan, C. Huang, H. Simitci and S. Yekhanin, "*On the Locality of Codeword Symbols,*" **IEEE Transactions on Information Theory**, 58(11), November, 2012. Comment: This paper might be too theoretical for most system readers. But, it is good to be aware of two fundamental trade-offs discovered in the paper: 1) in general, one has to sacrifice storage efficiency to gain repair efficiency; 2) in general, data blocks and parity blocks are in conflict in terms of repair efficiency.

- M. Blaum, J. L. Hafner and S. Hetzler, "*Partial-MDS Codes and Their Application to RAID Type of Architectures,*" IBM Research Report RJ10498 (ALM1202-001), February, 2012. (A version of this paper will appear in **IEEE Transactions on Information Theory**). Comment: Partial-MDS Codes (PMDS) are MR codes.

- J. S. Plank, M. Blaum and J. L. Hafner, "*SD Codes: Erasure Codes Designed for How Storage Systems*

Really Fail," **FAST-2013: 11th Usenix Conference on File and StorageTechnologies**, San Jose, February, 2013.

- D. S. Papailiopoulos, J. Luo, A. G. Dimakis, C. Huang and J. Li,"*Simple Regenerating Codes: Network Coding for Cloud Storage,"* **Proceedings of IEEE INFOCOM Mini-Conference**, 2012.

Efficient Repair of Existing Codes

The first paper below is about efficient repair of RDP codes (a class of RAID6 codes). The next two papers are about efficient repair of arbitrary binary codes.

- L. Xiang, Y. Xu, J. C. S. Lui and Q. Chang, "*Optimal Recovery of Single Disk Failure in RDP Code Storage Systems,"* **ACM SIGMETRICS**, June, 2010.
- O. Khan, R. Burns, J. S. Plank and C. Huang, "*In Search of I/O-Optimal Recovery from Disk Failures,"* **HotStorage '11: 3rd Workshop on Hot Topics in Storage and File Systems**, Portland, USENIX, June, 2011.
- O. Khan, R. Burns, J. S. Plank, W. Pierce and C. Huang, "*Rethinking Erasure Codes for Cloud File Systems: Minimizing I/O for Recovery and Degraded Reads,"* **FAST-2012: 10th Usenix Conference on File and Storage Technologies**, San Jose, February, 2012.

MDS Codes with Efficient Repair

- A. G. Dimakis, P. B. Godfrey, Y. Wu, M. Wainwright and K. Ramchandran, "*Network Coding for Distributed Storage Systems,"* **IEEE Transactions on Information Theory**, 2010. Comment: Regenerating Codes and theoretical bound on repair efficiency.
- A. G. Dimakis, K. Ramchandran, Y. Wu and C. Suh, "*A Survey on Network Codes for Distributed Storage,"* **Proceedings of the IEEE**, 99(3), March, 2011. Comment: Very good survey paper on regenerating codes covering the state-of-the-art until 2011.
- C. Suh and K. Ramchandran, "*Exact-repair MDS code construction using interference alignment,"* **IEEE Transactions on Information Theory**, 57(3), March, 2011. Comment: A class of practical regenerating codes with storage overhead at 2x (and above). Optimal disk I/O (and network) for data block repair. Optimal network (not disk I/O) for parity block repair.
- K. V. Rashmi, N. B. Shah and P. V. Kumar, "*Optimal Exact-Regenerating Codes for Distributed Storage at the MSR and MBR Points via a Product-Matrix Construction,"* **IEEE Transactions on Information Theory**, August, 2011. Comment: A class of practical regenerating codes with storage overhead at 2x (and above). Optimal network (not disk I/O) for both data and parity block repair.
- V. R. Cadambe, C. Huang, J. Li and S. Mehrotra, "*Polynomial Length MDS Codes With Optimal Repair in Distributed Storage,"* **Proceedings of Asilomar Conference on Signals, Systems and Computers**, 2011. Comment: A class of regenerating codes with storage overhead at 1.5x (above 1.5x can be easily achieved by setting data blocks to zero -- a standard technique called "shortening"). Optimal disk I/O (and network) for data block repair. No optimality for parity block repair. Practical when k is not too big ($(k/2)^2$ is reasonably small).

- I. Tamo, Z. Wang and J. Bruck, "*Zigzag codes: MDS array codes with optimal rebuilding,*" **submitted to IEEE Transactions on Information Theory**, October, 2011. Comment: Arbitrary storage overhead. Optimal disk I/O (and network) for data block repair. No optimality for parity block repair. Practical when $n - k = 2$.
- O. Khan, R. Burns, J. S. Plank, W. Pierce and C. Huang, "*Rethinking Erasure Codes for Cloud File Systems: Minimizing I/O for Recovery and Degraded Reads,*" **FAST-2012: 10th Usenix Conference on File and Storage Technologies**, San Jose, February, 2012. Comment: Rotated Reed-Solomon codes. Practical with arbitrary storage overhead. No optimality for either disk I/O or network.