Extending Flash Memory Lifetime Using Coding Techniques

Ryan Gabrys

Lara Dolecek

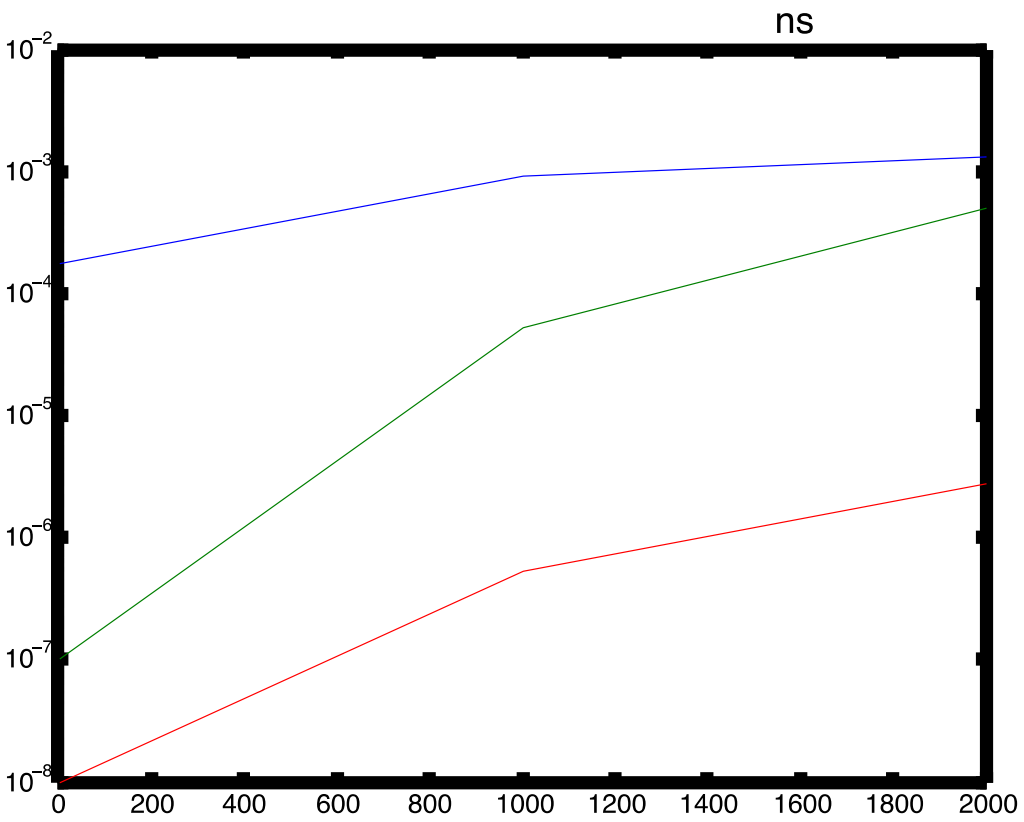
Outline

- Difficulties with flash
- Write once memory codes
- Error correction codes
- Conclusion

Difficulties with flash

- Erases can be more detrimental than programs in flash memory since erases are performed at the block level
- Multilevel flash memory cells are exponentially more error prone than single level cells

Flash error rates



[1] A. Jagmohan et al., “Adaptive endurance coding for NAND Flash,” IEEE Globecomm 2010.

Write once memory codes

Memory Model

The memory state is modeled as a vector $\mathbf{y}^j$ of length n where j is the current write or generation, each position can have q values $\{0, 1, 2, \ldots, q-1\}$. At generation j, encoder writes message M_i to memory by updating y^{j-1} to y^j satisfying WOM-constraint where $\mathbf{y}^j \geq \mathbf{y}^{j-1}$. Decoder sees y^j.

Definition

If M_j codewords can be represented at generation j, then generation j has rate $\frac{1}{n}\log(M_j)$. Sum rate is the sum of rates across generations.

Fixed-rate and unrestricted-rate codes

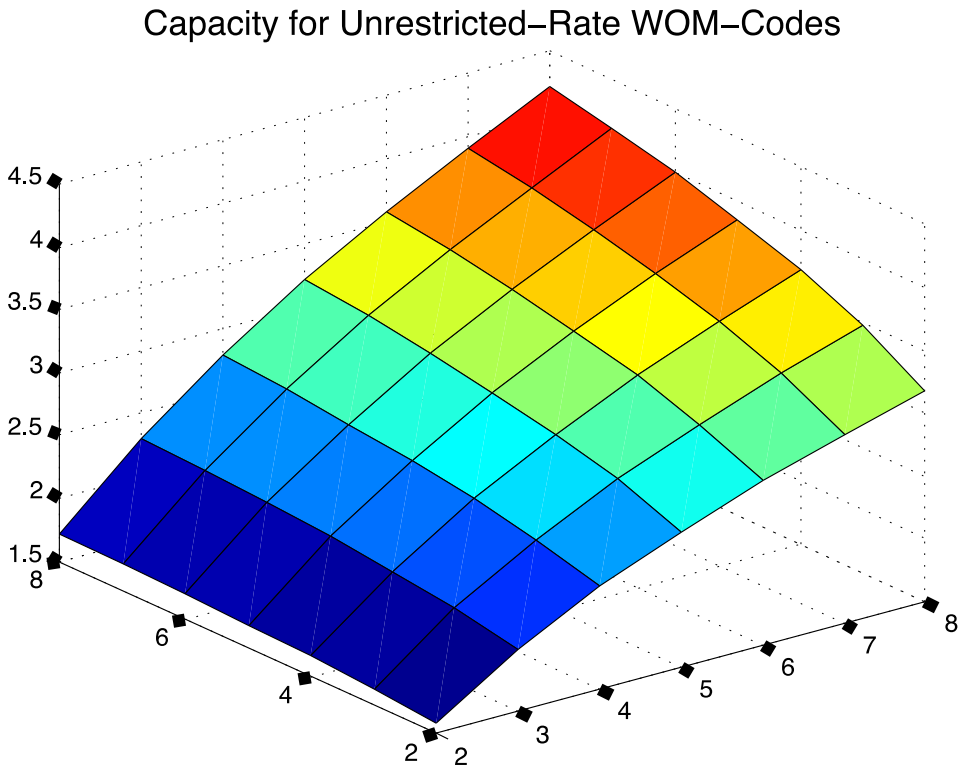
- If we write the same amount of information at each generation, the WOM-code is said to be a fixed rate WOM-code.
- Otherwise, the code is an unrestricted rate WOM-code.

Rivest and Shamir write twice code

Information	First Generation	Second Generation
00	000	111
01	001	110
10	010	101
11	100	011

[1] R.L. Rivest, A. Shamir “How to reuse a write-once memory” IEEE Communications Letters, 2003. Vol. 55, No. 1-3, pp. 1-19 Dec. 1982.

Capacity results for unrestricted rate



[1] F. Fu and A.J. Han Vinck, “On the Capacity of Generalized Write-Once Memory with State transitions Described by an Arbitrary Directed Acyclic Graph,” IEEE Trans. Inform. Theory, vol. 45, no. 1, pp. 308-313, Sept. 1999.

Rates achieved for two writes

q	Achieved Sum-rate	Capacity	Achieved/Capacity
4	2.9856	3.3219	0.8988
8	4.4784	5.1699	0.8662
16	6.3083	7.0875	0.8901
32	8.3083	9.0444	0.9186
64	10.3083	11.0224	0.9352
128	12.3083	13.0112	0.9460

[1] R. Gabrys et al., “Non-binary WOM-Codes for Multilevel Flash Memories,” Information Theory Workshop 2011.

Rates achieved for q=4

t	Achieved Sum-rate	Capacity	Achieved/Capacity
2	2.9856	3.3219	0.8988
3	3.2200	4.3219	0.7450
4	3.7128	5.1293	0.7238
5	3.9328	5.8074	0.6772
6	4.2594	6.3923	0.6663
7	4.3394	6.9069	0.6283

[1] R. Gabrys et al., “Non-binary WOM-Codes for Multilevel Flash Memories,” Information Theory Workshop 2011.

Rates achieved for q=8

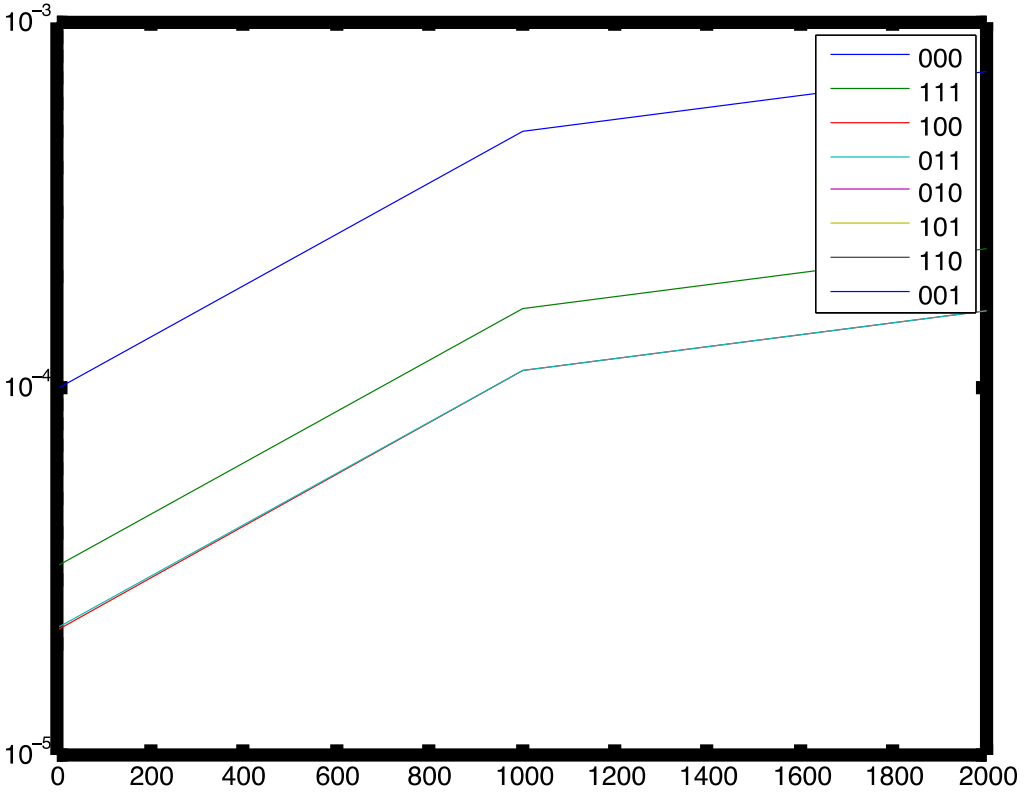
t	Achieved Sum-rate	Capacity	Achieved/Capacity
2	4.4784	5.1699	0.8662
3	4.8300	6.9069	0.6993
4	5.5692	8.3663	0.6657
5	5.8992	9.6294	0.6126
6	6.3891	10.7448	0.5946
7	6.5091	11.7448	0.5542

[1] R. Gabrys et al., “Non-binary WOM-Codes for Multilevel Flash Memories,” Information Theory Workshop 2011.

Observed error patterns in TLC flash memory

- Data was measured from 16 blocks of a TLC chip split evenly across the 2 planes.
  - The chip was repeatedly rewritten with random data and error measurements were taken every 100 cycles.
- Patterns
  - Approximately 69% were in same direction.
  - Errors were limited in magnitude.
  - Errors not bursty or persistent.

Data dependent BER



Only 4 states exhibit a BER more than 10^-5

Errors between state transitions

Programmed state	Errored state	Percentage of errors
000	010	0.2467
001	001	0.2444
111	101	0.0820
111	110	0.0807
000	100	0.0669
011	001	0.0556
100	110	0.0550
011	010	0.0547
100	101	0.0540
111	011	0.0217

Idea

- Instead of providing ECC that guards against ALL types of error patterns, target errors that flip a prescribed number of bit flips.

- Reduces the redundancy of the code resulting in more efficient error correction codes

Example

- Suppose we have q=32 levels and we want to correct all errors that flip at MOST 2 bits.
- Naïve Scheme: map the symbols to an ECC whose alphabet has cardinality 32.
- Better idea: map the symbols to a code with symbol cardinality 16.

Evaluation

- For the TLC data collected, we were able to correct 99.17% of the errors over the lifetime of the chip.
  - Used a mapping into an alphabet of size 4.
- More generally, suppose the probability that given 2^{n1} symbols a bit flip occurs is p1. Suppose we have n symbols and n1 and n get large. We want to correct t such symbol errors.

Evaluation

- The ratio of the naïve encoding to the redundancy of this scheme when optimal block codes are used is: $2^{tn1(1-H(p1))}$, where H is the binary entropy function.

- For fixed $p1 \neq 1/2$ as n1 continues to increase this scheme will yield more efficient codes.

Conclusion

- Bit error rates are continuing to increase with each generation of flash memory.
- Techniques for controlling this trend include write once memory codes and error correcting codes.
- Efficient error correction code should take advantage of the unique patterns observed.