Western Digital.

A Flexible and Low-Complexity Locally Erasure Recovery Scheme

Xinmiao Zhang
Senior Technologist

Steven Sprouse
Engineering Fellow

August 8, 2017

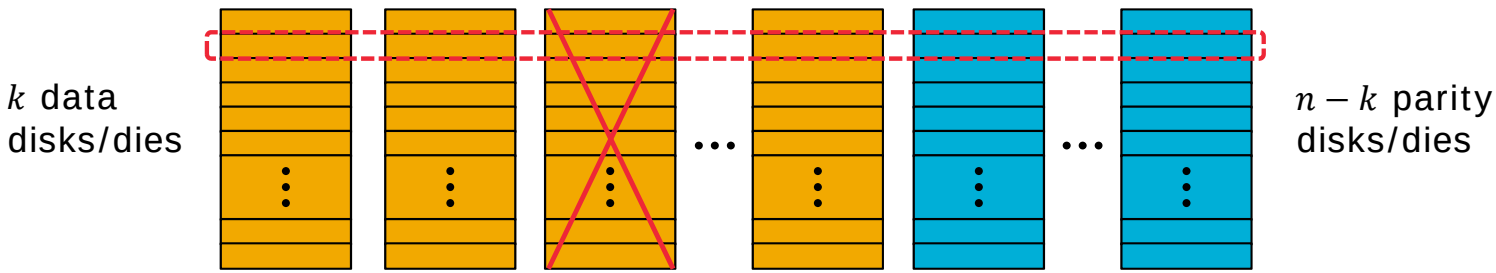
Flash Memory Summit

SAFE HARBOR | DISCLAIMERS
Forward-Looking Statements

This presentation contains forward-looking statements that involve risks and uncertainties, including, but not limited to, statements regarding our addressable market, our product and technology positioning and compute platforms, the anticipated benefits of our new technologies, executing on our integrated strategic plans, realizing our strategic imperatives, including our solid-state drives and storage technologies. Forward-looking statements should not be read as a guarantee of future performance or results, and will not necessarily be accurate indications of the times at, or by, which such performance or results will be achieved, if at all. Forward-looking statements are subject to risks and uncertainties that could cause actual performance or results to differ materially from those expressed in or suggested by the forward-looking statements.

Additional key risks and uncertainties include the impact of continued uncertainty and volatility in global economic conditions; actions by competitors; difficulties associated with go-to-market capabilities; business conditions; growth in our markets; and pricing trends and fluctuations in average selling prices. More information about the other risks and uncertainties that could affect our business are listed in our filings with the Securities and Exchange Commission (the “SEC”) and available on the SEC’s website at www.sec.gov, including our most recently filed periodic report, to which your attention is directed. We do not undertake any obligation to publicly update or revise any forward-looking statement, whether as a result of new information, future developments or otherwise, except as otherwise required by law.

Traditional Erasure Codes for Distributed Storage



- Large-scale storage system needs to be able to recover or tolerate multiple failed or unavailable disk/SSD/NAND dies
- Goal of traditional erasure coding schemes: minimize redundancy
  - An (n,k) maximum distance separable (MDS) code has $t = d_{min} - 1 = n - k$
- Traditional erasure codes

  t: erasure-correcting capability
  - XOR-single parity $(t = 1)$
  - RAID-6 (EVENODD, Row-Diagonal Parity (RDP), extended RS…) $(t = 2)$
  - Reed-Solomon (RS) codes $(t \geq 2)$
- Need to read k symbols in order to recover any erasure

Erasure Codes with Local Recovery

- Distributed storage needs erasure codes accessing much less than k symbols for recovery (possible if actual erasure number $< t$)
  - Lower network traffic
  - Reduced recovery latency
  - Better data availability
- Recent regenerating and local erasure recovery codes
  - Modified EVENODD and RDP (have constraints on n and k; $t = 2$)
  - Rotated RS ($t = 2, 3$)
  - Piggybacking (locality depends on failure indices)
  - Codes based on partitions of finite fields (high en/decoder complexity)
  - Sparse parity columns in the generator matrix, each parity covers overlapping data symbols (worse locality for data symbol recovery)
- Our new scheme for local recovery
  - Not much constraint on n, k, t
  - Easy tradeoff on locality and redundancy
  - Allows unequal protection
  - Low-complexity en/decoder

MDS Codes With Systematic Parity Check Matrices

- For a linear block code, x is a code word iff $Hx^T = 0$
- $H = [A|I]$ $\longrightarrow$ x consists of data symbols followed by parities

$$H = \begin{bmatrix} 1 & 1 & \cdots & 1 & 1 & 0 & \cdots & 0 \\ \alpha^0 & \alpha^1 & \cdots & \alpha^{k-1} & 0 & 1 & \cdots & 0 \\ \alpha^0 & \alpha^2 & \cdots & \alpha^{2(k-1)} & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \alpha^0 & \alpha^{t-1} & \cdots & \alpha^{(t-1)(k-1)} & 0 & 0 & \cdots & 1 \end{bmatrix}$$

Vandermonde matrix

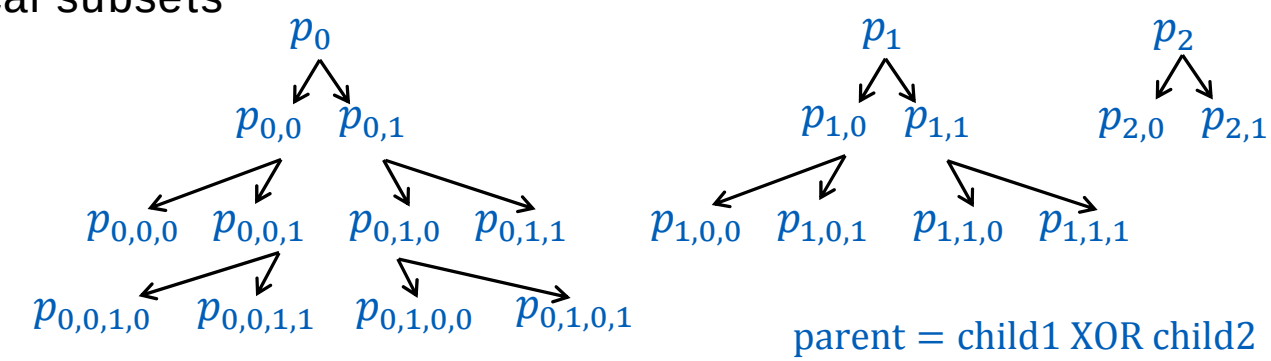
α: primitive element of $GF(2^r)$

- Using a Vandermonde matrix leads to simple encoder and decoder
- Not many constraints on code parameters in order to be MDS for small t
  - $t = 2$ or 3 requires $k \leq 2^r - 1$
  - $t = 4$ requires $k \leq 27$ over $GF(2^8)$; $k \leq 67$ over $GF(2^{10})$; $\cdots$

Parity Splitting for Local Recovery

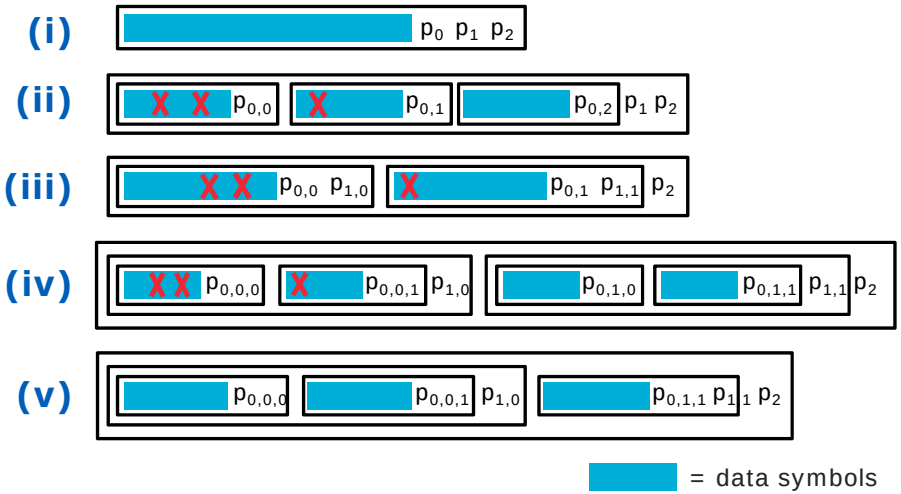
$$H = [A|I] = \begin{array}{c} \begin{matrix} d_0 & d_1 & d_2 & d_3 & d_4 & d_5 & d_6 & d_7 & d_8 & d_9 & d_{10} & d_{11} & p_0 & p_1 & p_2 & p_3 \end{matrix} \\ \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ \alpha^0 & \alpha^1 & \alpha^2 & \alpha^3 & \alpha^4 & \alpha^5 & \alpha^6 & \alpha^7 & \alpha^8 & \alpha^9 & \alpha^{10} & \alpha^{11} & 0 & 1 & 0 & 0 \\ \alpha^0 & \alpha^2 & \alpha^4 & \alpha^6 & \alpha^8 & \alpha^{10} & \alpha^{12} & \alpha^{14} & \alpha^{16} & \alpha^{18} & \alpha^{20} & \alpha^{22} & 0 & 0 & 1 & 0 \\ \alpha^0 & \alpha^3 & \alpha^6 & \alpha^9 & \alpha^{12} & \alpha^{15} & \alpha^{18} & \alpha^{21} & \alpha^{24} & \alpha^{27} & \alpha^{30} & \alpha^{33} & 0 & 0 & 0 & 1 \end{bmatrix} \end{array}$$

- MDS $\longleftrightarrow$ every square submatrix of A is nonsingular
- Any columns in t' rows of A and $I_{t' \times t'}$ form the H matrix of a t'-erasure-correcting MDS code
- Divide the data into subsets and split the parities; $t' < t$ erasures are correctable by using local subsets



parent = child1 XOR child2

Examples



	# of parities	locality for recovering		
		1 erasure	2 erasures	3 erasures
(i)	3	k	k	k
(ii)	5	$k/3$	$2k/3$ (1,1) k (2,0)	k
(iii)	5	$k/2$	$k/2$ (2,0) k (1,1)	k
(iv)	7	$k/4$	$k/2$	$k/2$ (2,1,0,0) $3k/4$ (1,1,1,0) (2,0,1,0) k (3,0,0,0)
(v)	6	$k/3$	$k/3$ (0,0,2) $2k/3$ (1,1,0) (1,0,1) (2,0,0)	$2k/3$ (2,1,0) (1,0,2) k (1,1,1) (2,0,1) (3,0,0) (0,0,3)

- Locality is defined as the number of symbols to access for recovery
- Optimal subset division and parity splitting are dependent on erasure patterns & probabilities

*$(a,b,\cdots)$ denote that the numbers of erasures in the last-level subsets are $a,b,\cdots$

Encoder Hardware Architecture

With Vandermonde matrix

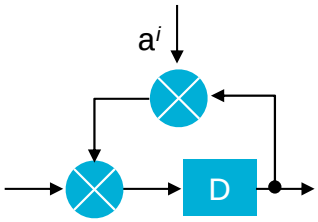
$$p_i = \sum_{j=0}^{k-1} \alpha^{ij} d_j = (\cdots((d_{k-1}\alpha^i + d_{k-2})\alpha^i + d_{k-3})\cdots)\alpha^i + d_0$$

Parity splitting

$$p_{i0} = \sum_{j=0}^{k/2-1} \alpha^{ij} d_j$$
$$p_{i1} = \sum_{j=k/2}^{k-1} \alpha^{ij} d_j$$

modified split parity

$$p'_{i1} = \sum_{j=0}^{k/2-1} \alpha^{ij} d_{j+k/2}$$
$$= p_{i1}/(\alpha^{ik/2})$$



Multiply-Add (MA) Unit

- Split parities are derived by resetting the register when the data symbols of the next subset arrives
- The constant scalar is multiplied back before the split parities are added up to derive the parent parity if needed in the decoder

Decoding Formulas for $t=4$

- erasure indices: w,x,y,z

$$d_w\alpha^{wi}+d_x\alpha^{xi}+d_y\alpha^{yi}+d_z\alpha^{iz} = p_i+\sum_{j\neq w,x,y,z}\alpha^{ij}d_j = q_i$$

Solve linear equations

$$d_z = \frac{\alpha^{w+x+y}q_0+(\alpha^{w+x}+\alpha^{w+y}+\alpha^{x+y})q_1+(\alpha^{w}+\alpha^{x}+\alpha^{y})q_2+q_3}{(\alpha^{w}+\alpha^{z})(\alpha^{x}+\alpha^{z})(\alpha^{y}+\alpha^{z})}$$

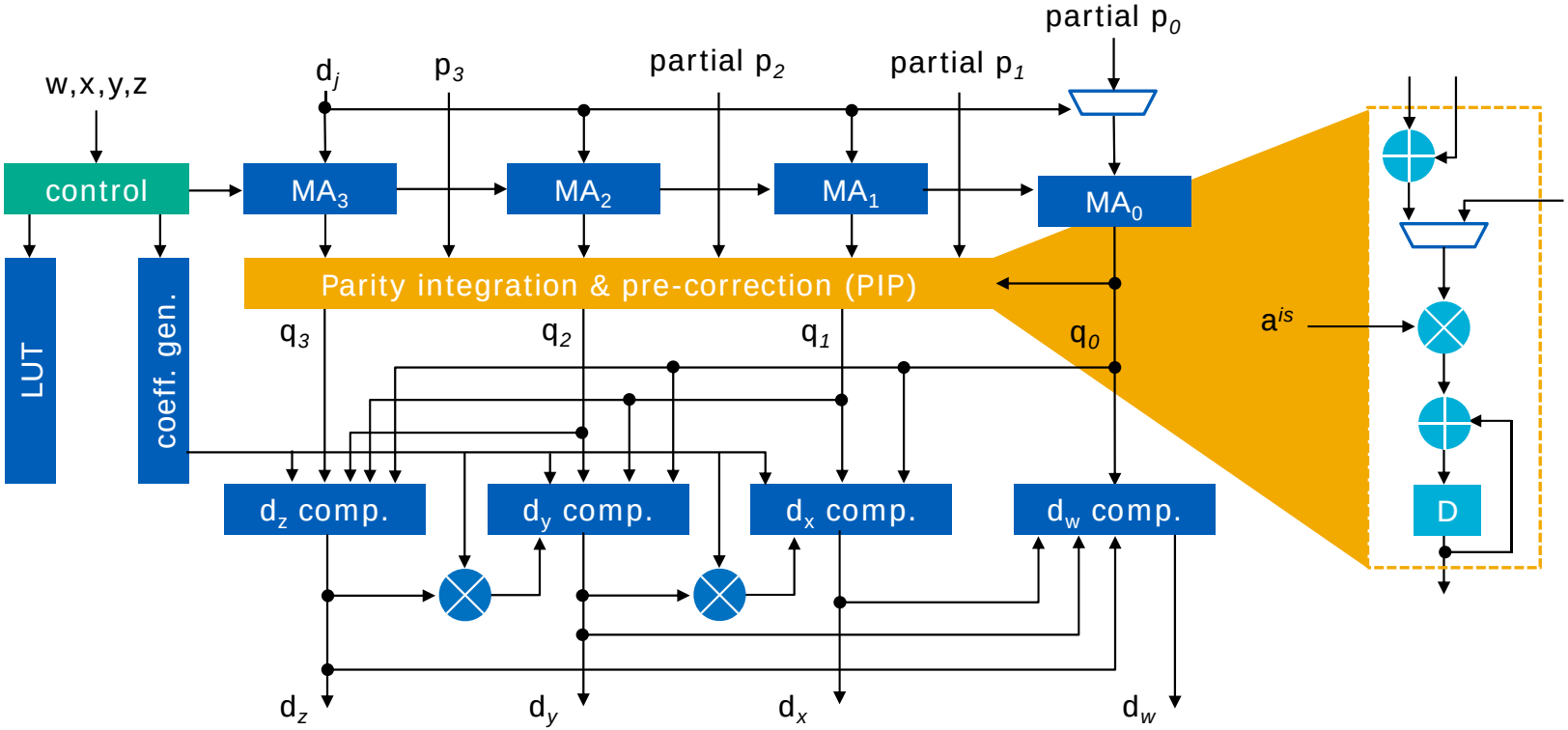
$$d_y = \frac{\alpha^{w+x}q_0+(\alpha^{w}+\alpha^{x})q_1+q_2+(\alpha^{w}+\alpha^{z})(\alpha^{x}+\alpha^{z})d_z}{(\alpha^{w}+\alpha^{y})(\alpha^{x}+\alpha^{y})}$$

$$d_x = \frac{\alpha^{w}q_0+q_1+(\alpha^{w}+\alpha^{y})d_y}{(\alpha^{w}+\alpha^{x})}$$

$$d_w = q_0 + d_x + d_y + d_z$$

- Formulas would be more complex if a non-Vandermonde matrix is used
- Complexity is lower than that of Berlekamp-Massey RS decoder

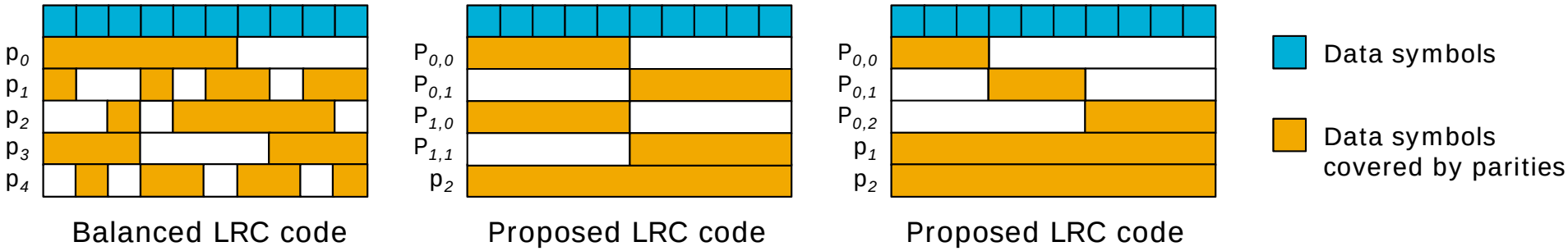
Decoder Hardware Architecture for $t=4$



- Single-erasure pre-correction is done first
- Overhead compared to non-local decoder is less than 15%

Comparisons with Existing Codes with Parity Manipulation

- Microsoft Azure locally recoverable code is a special case that splits only p_0
- Pyramid codes split parities and divide data sets in a similar way
  - Use existing MDS codes, whose systematic H generally has irregular entries
  - Has higher hardware complexity
- Codes with sparse parity columns in systematic generator matrix covering overlapping message symbols
  - Inferior locality for recovering data symbols



Western Digital®