Erasure Codes Made So Simple, You'll Really Like Them

W. David Schwaderer
August 7, 2014
schwaderer_01@comcast.net

Agenda

- Errors Versus Erasures
- HDD Bit Error Rate Implications
- RAID 4, 5, and 6 Review
- Objects and Dispersed Storage
- The Math
- Summary
- Questions

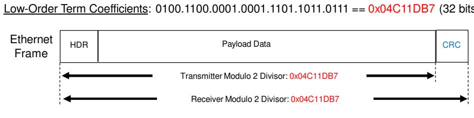
Errors Versus Erasures

Data Communications *Error Detection*

- Asynchronous (Start/Stop) Communication – Parity (e.g. Even/Odd)
- TCP/IP Checksums
- Ethernet CCITT-32 CRC

(x+1) * *Special* Polynomial: $(x^{32}+x^{26}+x^{23}+x^{22}+x^{16}+x^{12}+x^{11}+x^{10}+x^{8}+x^{7}+x^{5}+x^{4}+x^{2}+x^{1}+1)$

Low-Order Term Coefficients: 0100.1100.0001.0001.1101.1011.0111 == 0x04C11DB7 (32 bits)



Erasure: Error in known position

Key Take Away:

Special Polynomials; Magic Numbers; Bit-Wise, Modulo 2 Arithmetic

HDD Facts of Life

Nominal raw BER values: 10^{-5} to 10^{-6} => Improve reliability using ECC

ECC-corrected BER ratings:
Desktop hard disks ~ 1 in 10^{14}
Enterprise hard disks ~ 1 in 10^{15} (ten times better).

Desktop hard disks have ~8X the capacity of an enterprise disks.

=> Large capacity disks experience inevitable uncorrectable bit errors

Disk arrays use RAID to improve reliability.

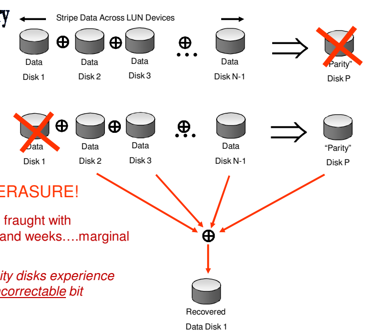
(Spreads data across multiple disks)

RAID ~ Redundant Array of *Inexpensive* Disks

~ Redundant Array of *Independent* Disks

W. David Schwaderer - FMS 2014 - August 7, 2014 4

RAID 4 & 5 Parity



RAID 5 *Distributes* Parity

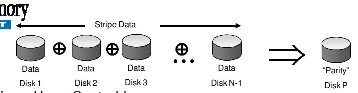
Rebuilds are fraught with peril….days and weeks….marginal reliability…

Large capacity disks experience inevitable uncorrectable bit errors…

"N" Total Disks, "N-1" Data Disks, "*N-1 of N*"
Can Lose One Disk, May Have Rebuild Problems

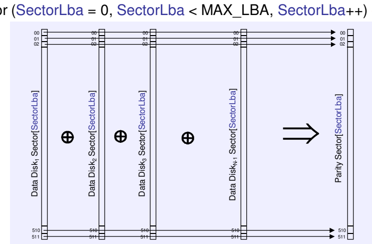
W. David Schwaderer - FMS 2014 - August 7, 2014 5

RAID 4 & 5 Parity



RAID 5 Distributes Parity

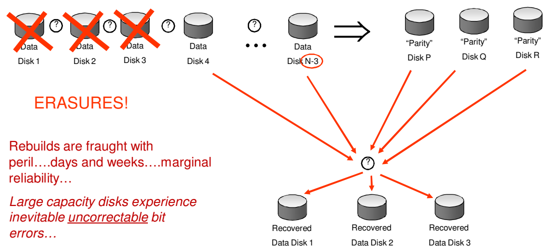
```
unsigned long SectorLba;
for (SectorLba = 0, SectorLba < MAX_LBA, SectorLba++) {
```




```
}
```

W. David Schwaderer - FMS 2014 - August 7, 2014 6

RAID 6 Parity



ERASURES!

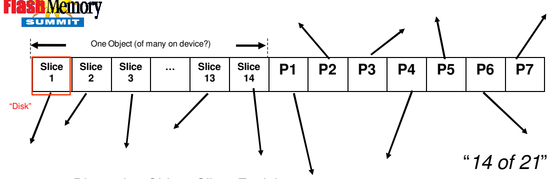
Rebuilds are fraught with peril….days and weeks….marginal reliability…

Large capacity disks experience inevitable uncorrectable bit errors…

"N" Total Disks, "N-3" Data Disks, "*N-3 of N*"
Can Lose Three Disks, Even Fewer Rebuild Problems

W. David Schwaderer - FMS 2014 - August 7, 2014 7

Objects & Information Dispersal



Dispersing Object *Slices* Exploits:

- Collective storage device pool isolation/reliability
- Distributed scale-out infrastructure design strengths
  1. Distributed rebuilds for lost object slices
  2. Network bandwidth load balancing

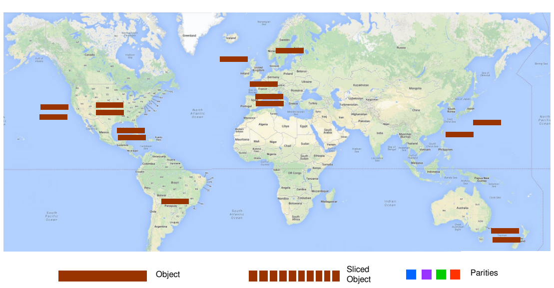
Eliminates 7X data redundancy and RAID rebuild weaknesses.

However, extreme data availability brings no free lunch:

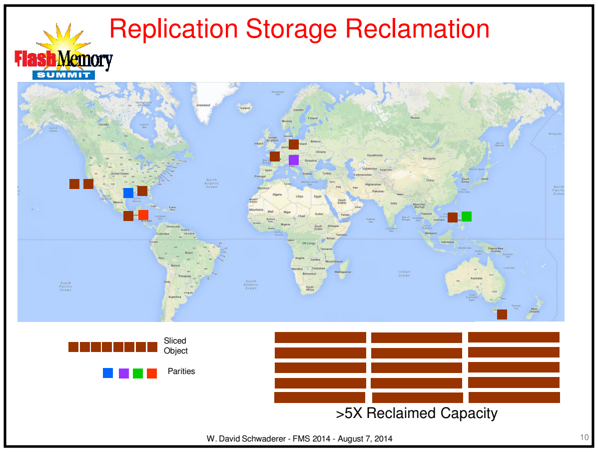
- Computations
- Read/Update/Write Workload Amplification
- *Scale at Tail* Laggard Compensations

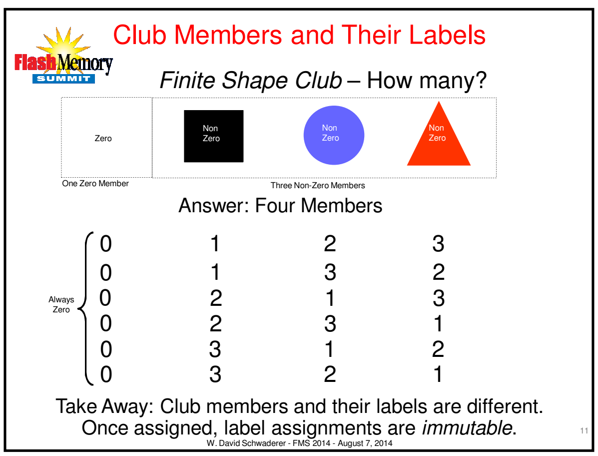
W. David Schwaderer - FMS 2014 - August 7, 2014 8

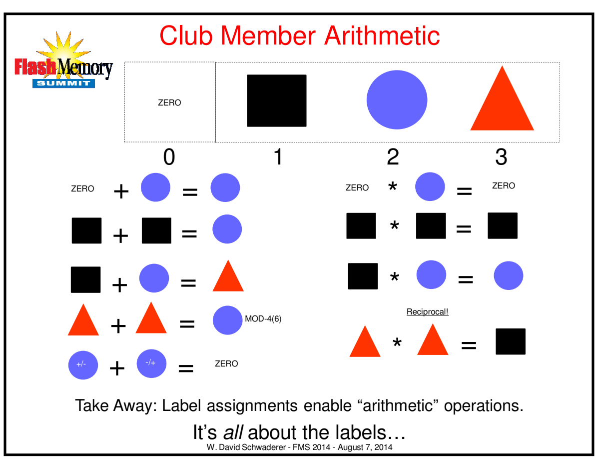
Replication Storage Reclamation



W. David Schwaderer - FMS 2014 - August 7, 2014 9


Replication Storage Reclamation
Sliced Object
Parities
>5X Reclaimed Capacity
W. David Schwaderer - FMS 2014 - August 7, 2014
10


Club Members and Their Labels
Finite Shape Club – How many?
Zero
Non Zero
Non Zero
Non Zero
One Zero Member
Three Non-Zero Members
Answer: Four Members
Always Zero
0 1 2 3
0 1 3 2
0 2 1 3
0 2 3 1
0 3 1 2
0 3 2 1
Take Away: Club members and their labels are different.
Once assigned, label assignments are *immutable*.
W. David Schwaderer - FMS 2014 - August 7, 2014
11


Club Member Arithmetic
ZERO
0 1 2 3
ZERO + =
+ =
+ =
+ = MOD-4(6)
+/- + -/+ = ZERO
ZERO * = ZERO
* =
* =
Reciprocal!
* =
Take Away: Label assignments enable "arithmetic" operations.
It's *all* about the labels…
W. David Schwaderer - FMS 2014 - August 7, 2014

Polynomial Shorthand

Consider Degree 8 Polynomials

$aX^8 + bX^7 + cX^6 + dX^5 + eX^4 + fX^3 + gX^2 + hX^1 + kX^0$

Example Degree 8 Polynomial

$4X^8 + 0X^7 + 2X^6 + 16X^5 + 10X^4 + 8X^3 + (-22)X^2 + 19X^1 + 36X^0$

Example Degree 8 Polynomial Shorthand

4.0.2.16.10.8.-22.19.36 Label!

W. David Schwaderer - FMS 2014 - August 7, 2014 13

$X^8 + X^4 + X^3 + X^2 + 1$ (*Special* Polynomial)

$X^8 + 0X^7 + 0X^6 + 0X^5 + 1X^4 + 1X^3 + 1X^2 + 0X^1 + 1X^0$ = 1.0001.1101
= 0x1D Polynomial *Label!*

If α is a root, by definition:

$\alpha^8 + \alpha^4 + \alpha^3 + \alpha^2 + 1 = 0$

So,

$\alpha^8 + \alpha^4 + \alpha^3 + \alpha^2 + 1 = 0$
$\oplus(\alpha^4 + \alpha^3 + \alpha^2 + 1) = \oplus(\alpha^4 + \alpha^3 + \alpha^2 + 1)$

or: $\alpha^8 = \alpha^4 + \alpha^3 + \alpha^2 + 1$

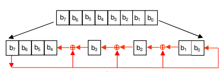
$\alpha^0 = 1 = 0\alpha^7 + 0\alpha^6 + 0\alpha^5 + 0\alpha^4 + 0\alpha^3 + 0\alpha^2 + 0\alpha + 1$ ~ 0x01
$\alpha^1 = \alpha = 0\alpha^7 + 0\alpha^6 + 0\alpha^5 + 0\alpha^4 + 0\alpha^3 + 0\alpha^2 + 1\alpha + 0$ ~ 0x02
$\alpha^2 = \alpha^2 = 0\alpha^7 + 0\alpha^6 + 0\alpha^5 + 0\alpha^4 + 0\alpha^3 + 1\alpha^2 + 0\alpha + 0$ ~ 0x04
$\alpha^3 = \alpha^3 = 0\alpha^7 + 0\alpha^6 + 0\alpha^5 + 0\alpha^4 + 1\alpha^3 + 0\alpha^2 + 0\alpha + 0$ ~ 0x08
$\alpha^4 = \alpha^4 = 0\alpha^7 + 0\alpha^6 + 0\alpha^5 + 1\alpha^4 + 0\alpha^3 + 0\alpha^2 + 0\alpha + 0$ ~ 0x10
$\alpha^5 = \alpha^5 = 0\alpha^7 + 0\alpha^6 + 1\alpha^5 + 0\alpha^4 + 0\alpha^3 + 0\alpha^2 + 0\alpha + 0$ ~ 0x20
$\alpha^6 = \alpha^6 = 0\alpha^7 + 1\alpha^6 + 0\alpha^5 + 0\alpha^4 + 0\alpha^3 + 0\alpha^2 + 0\alpha + 0$ ~ 0x40
$\alpha^7 = \alpha^7 = 1\alpha^7 + 0\alpha^6 + 0\alpha^5 + 0\alpha^4 + 0\alpha^3 + 0\alpha^2 + 0\alpha + 0$ ~ 0x80

$\alpha^8 = \alpha^4 + \alpha^3 + \alpha^2 + 1$
$= 0\alpha^7 + 0\alpha^6 + 0\alpha^5 + 1\alpha^4 + 1\alpha^3 + 1\alpha^2 + 0\alpha + 1$ ~ 0x1D

$\alpha^9 = \alpha(\alpha^8) = \alpha(\alpha^4 + \alpha^3 + \alpha^2 + 1) = \alpha^5 + \alpha^4 + \alpha^3 + \alpha$ ~ 0x3A
$\alpha^{10} = \alpha^2(\alpha^8) = \alpha^2(\alpha^4 + \alpha^3 + \alpha^2 + 1) = \alpha^6 + \alpha^5 + \alpha^4 + \alpha^2$ ~ 0x74
$\alpha^{11} = \alpha^3(\alpha^8) = \alpha^3(\alpha^4 + \alpha^3 + \alpha^2 + 1) = \alpha^7 + \alpha^6 + \alpha^5 + \alpha^3$ ~ 0xE8

$\alpha^{12} = \alpha^4(\alpha^8) = \alpha^4(\alpha^4 + \alpha^3 + \alpha^2 + 1) = \alpha^8 + \alpha^7 + \alpha^6 + \alpha^4$
$= (\alpha^4 + \alpha^3 + \alpha^2 + 1) + \alpha^7 + \alpha^6 + \alpha^4$
$= \alpha^7 + \alpha^6 + \alpha^3 + \alpha^2 + 1$ ~ 0xCD

...



Linear Feedback Shift Register

$\alpha^{254} = \alpha^7 + \alpha^3 + \alpha^2 + \alpha^1$ ~ 0x8E

$\alpha^{255} = \alpha(\alpha^{254}) = \alpha(\alpha^7 + \alpha^3 + \alpha^2 + \alpha^1) = \alpha^8 + \alpha^4 + \alpha^3 + \alpha^2$
$= (\alpha^4 + \alpha^3 + \alpha^2 + 1) + \alpha^4 + \alpha^3 + \alpha^2$
$= 1$
$= \alpha^0$

W. David Schwaderer - FMS 2014 - August 7, 2014 14

Linear Algebra Review

X + Y + 4 + 2 = 9
3X - Y + 6 - 5 = 2

X + Y = 3
\+ 3X - Y = 1

4X = 4

=> X == 1
X + Y = 3 => Y == 2

15

W. David Schwaderer - FMS 2014 - August 7, 2014

Galois Field Overview

A <u>finite</u> set of elements (e.g. 2^8, 2^{16}, etc.) that have:

- "Arithmetic" operator $\oplus \sim +$
- "Multiplicative" operator $\otimes \sim *$
- A *zero* member
- Negatives and Reciprocals

Field operations always produce another element in the set. (Closure)

In this discussion, our Galois Field has:

- 256 elements (*thingees*) <u>labeled</u> 0x01 to 0xFF, and 0x00 - [*GF(8)*]
- An "Arithmetic" operator (addition and subtraction) ~ XOR $\oplus$
- A "Multiplicative" operator (* and ÷) ~ exponent operations $\otimes$

W. David Schwaderer - FMS 2014 - August 7, 2014 16

Galois "Multiplicative" Operator

$$(X)\otimes(Y) = (\alpha^{\log_\alpha X})\otimes(\alpha^{\log_\alpha Y}) = \alpha^{[(\log_\alpha X + \log_\alpha Y) \text{ modulus } 255]}$$

255 non-zero members

Noting that $\alpha^K \otimes \alpha^{-K} = \alpha^{K-K} = \alpha^0 = 1$

...it follows that $1/(\alpha^K) = \alpha^{-K}$

So...

$$(X)/(Y) = (\alpha^{\log_\alpha X})/(\alpha^{\log_\alpha Y})$$
$$= (\alpha^{\log_\alpha X})\otimes(1/(\alpha^{\log_\alpha Y}))$$
$$= (\alpha^{\log_\alpha X})\otimes(\alpha^{-\log_\alpha Y})$$
$$= \alpha^{[(\log_\alpha X - \log_\alpha Y) \text{ modulus } 255]}$$

Note: Exponent additions/subtractions operations are base 16.

W. David Schwaderer - FMS 2014 - August 7, 2014 17

Example Calculations

$0x2B \otimes 0x6F = 0x8F$

<u>From Log Table (Table 2)</u>

$0x2B = \alpha^{0xDA_{16}} = \alpha^{218_{10}}$ (Base 10 for Understanding Ease)

$0x6F = \alpha^{0x3D_{16}} = \alpha^{61_{10}}$

<u>So</u>

$$0x2B \otimes 0x6F = \alpha^{218_{10}} \otimes \alpha^{61_{10}}$$
$$= \alpha^{(218_{10} + 61_{10})\bmod 255_{10}}$$
$$= \alpha^{(279_{10})\bmod 255_{10}}$$
$$= \alpha^{24_{10}}$$
$$= \alpha^{0x18_{16}}$$

<u>From Powers Table (Table 1)</u>

$\alpha^{0x18} = 0x8F$

W. David Schwaderer - FMS 2014 - August 7, 2014 18

Example Calculations (cont.)

Arithmetic (Addition, Subtraction):

$0x2D \oplus 0x28 == 0010.1101_b$ XOR

$0010.1000_b = 0000.0101_b = 0x05$

Multiplication: $0x3F \otimes 0x12$

$= \alpha^{0xA6} \otimes \alpha^{0xE0}$ (From Table 2)

$= \alpha^{0xA6 + 0xE0}$ (Base 16 Regular Addition)

$= \alpha^{(0x0186) \bmod 255}$

$= \alpha^{0x87}$

$= 0xA9$ (From Table 1)

Division: $0x3F \otimes (1/(0x12))$

$= \alpha^{0xA6} \otimes \alpha^{-0xE0}$ (From Table 2)

$= \alpha^{0xA6 - 0xE0}$ (Base 16 Regular Subtraction)

$= \alpha^{0x01A5 - 0xE0}$ (0xA6 = 0xA6 + 0xFF = ox01A5)

$= \alpha^{0xC5} = 0x8D$ (From Table 1)

W. David Schwaderer - FMS 2014 - August 7, 2014 19

Oddities

Negative Values

$\oplus \approx$ XOR

Thus, for any element V, V = -V since (V XOR V) = 0

Reciprocals

Let $\log_\alpha V == 0xNN$ ($V \neq 0$)

Then $1/V = 1/(\alpha^{0xNN}) = \alpha^{-0xNN}$

Proof: $v/v = 1 = (\alpha^{0xNN}) / (\alpha^{0xNN}) = (\alpha^{0xNN}) \otimes [1/(\alpha^{0xNN})]$

$= (\alpha^{0xNN}) \otimes (\alpha^{-0xNN})$ (?)

$= (\alpha^{0xNN + -0xNN})$

$= \alpha^0$

$= 1$

W. David Schwaderer - FMS 2014 - August 7, 2014 20

Double Data Disk Failure

Remembering

Parity Disk P Generation: $D_1 \oplus D_2 \oplus D_3 \oplus \ldots \oplus D_{N-2} = P$ (Normal RAID 5 Parity Generation)

Parity Disk Q Generation: $(K_1 \otimes D_1) \oplus (K_2 \otimes D_2) \oplus (K_3 \otimes D_3) \oplus \ldots \oplus (K_{N-2} \otimes D_{N-2}) = Q$

Losing Data Disk 1 and 2 gives

$D_1 \oplus D_2 = P \oplus D_3 \oplus \ldots \oplus D_{N-2} = V_1$ (1)

$(K_1 \otimes D_1) \oplus (K_2 \otimes D_2) = Q \oplus (K_3 \otimes D_3) \oplus \ldots (K_{N-2} \otimes D_{N-2}) = V_2$ (2)

So … $(K_1 \otimes D_1) \oplus (K_1 \otimes D_2) = (K_1 \otimes V_1)$

$\oplus$ $(K_1 \otimes D_1) \oplus (K_2 \otimes D_2) = V_2$

Gives … $D_2 \otimes (K_1 \oplus K_2) = (K_1 \otimes V_1) \oplus V_2$

Or … $D_2 = ((K_1 \otimes V_1) \oplus V_2) / (K_1 \oplus K_2)$

Recover D_1 using RAID 5 Logic with P values…

W. David Schwaderer - FMS 2014 - August 7, 2014 21

Summary

Easy to Solve Independent Linear Equations

$$X + Y = 3$$

$$3*X - Y = 1$$

Similarly, Losing Two Data Values Results in:

$$X \oplus Y = V_1$$

$$0x3 \otimes X \oplus Y = V_2$$

W. David Schwaderer - FMS 2014 - August 7, 2014

22

The End

Thank you!

schwaderer_01@comcast.net

W. David Schwaderer - FMS 2014 - August 7, 2014

23