An Explicit Construction of a High-Rate Minimum Storage Regenerating Code with Low Sub-Packetization and Selectable Repair Degree

Birenjith Sasidharan, Myna Vajha and P. Vijay Kumar†

Department of Electrical Communication Engineering,
Indian Institute of Science, Bangalore

Mathematical Methods for Cryptography
Organized by Selmer Center, Department of Informatics, University of Bergen,
Dedicated to celebrate Prof. Tor Helleseth's 70th birthday

Thon Hotel Lofoten, Svolvær
Archipelago Lofoten, Norway
September 4-8, 2017

Organization

- Distributed Storage
- Regenerating Codes
- High-Rate MSR Codes

Acknowledgement

- Thanks to Lilya, Mathew and Chunlei for the invitation and the superb organization,

Acknowledgement

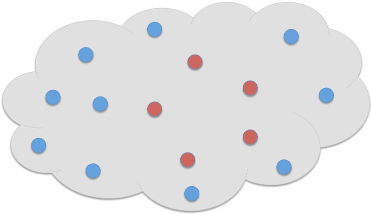
- Thanks to Lilya, Mathew and Chunlei for the invitation and the superb organization,
- It has been my privilege to have known and worked with Tor
- I have learned a great deal from him over the years...

Acknowledgement

- Thanks to Lilya, Mathew and Chunlei for the invitation and the superb organization,
- It has been my privilege to have known and worked with Tor
- I have learned a great deal from him over the years...
- and look forward to the continued interaction..

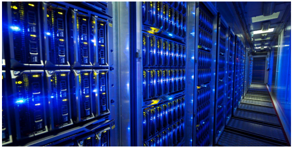
Distributed Storage Setting

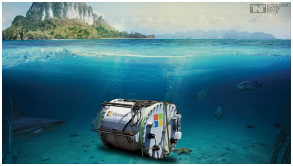
Distributed Storage Setting



- data pertaining to a single file is distributed across storage nodes
- nodes are inexpensive storage devices
  (a) prone to failure,
  (b) down for maintenance,
  (c) unavailable, busy serving other demands..

Data Center Setting: Size Matters!

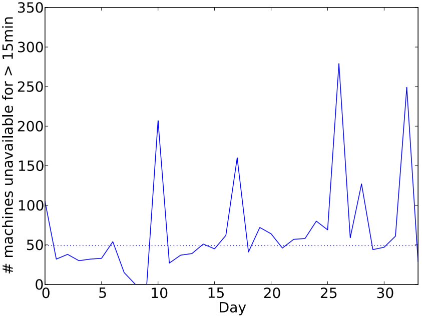




- NSA data centre (right) estimated to store $3-12$ Exabytes ($10^6\,GB$)

- `https://c.slashgear.com/wp-content/uploads/2012/10/google-datacenter-tech-13.jpg`
- `http://www.techworm.net/wp-content/uploads/2016/02/microsoft-dives-underwater-to-build-a-cool-data-center.jpg`
- `https://nsa.gov1.info/utah-data-center/`

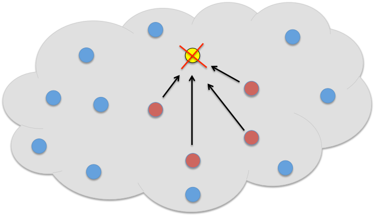
Node Failures in Facebook Warehouse Cluster



- thousands of storage units each storing 24-36 TB
- a total of several hundred petabytes of data
- number of failures per day over 30-day period (2013)

K. V. Rashmi et al., "A Solution to the Network Challenges of Data Recovery in Erasure-coded Distributed Storage Systems: A Study on the Facebook Warehouse Cluster", USENIX HotStorage, June 2013.

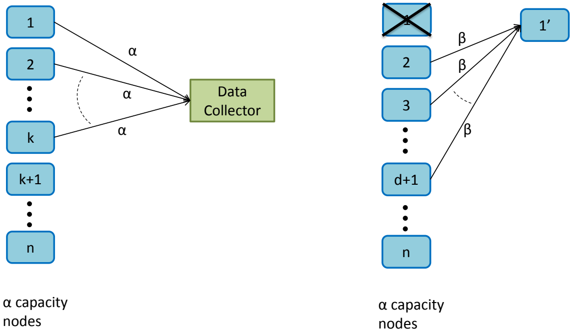
Distributed Storage Setting



- Thus there is node for efficient repair of a failed node
- Focus on
  (a) repair bandwidth - amount of data download
  (b) repair degree - number of helper nodes contacted

Regenerating Codes

Parameters: $(\ (n, k, d), \ (\alpha, \beta), \ B, \ \mathbb{F}_q \ )$

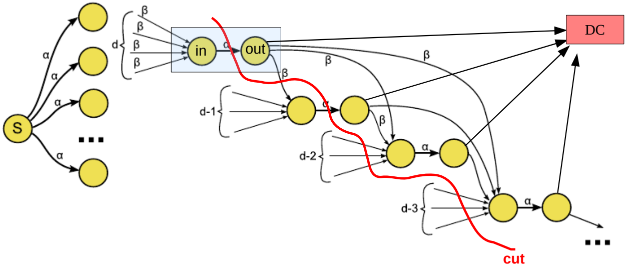


- Data can be recovered by connecting to any k of n nodes
- A failed node can be repaired by connecting to any d nodes, downloading β symbols from each node; ($d\beta <<$ file size B)
- We restrict to Minimum-Storage-Regenerating (MSR) codes – repair-optimal MDS codes.

Cut-Set Bound from Network Coding

Given code parameters $\{[n, k, d], (\alpha, \beta)\}$:

$$B \;\leq\; \sum_{i=1}^{k} \min\{\alpha, (d-i+1)\beta\}.$$



(can be shown to be achievable under functional repair)

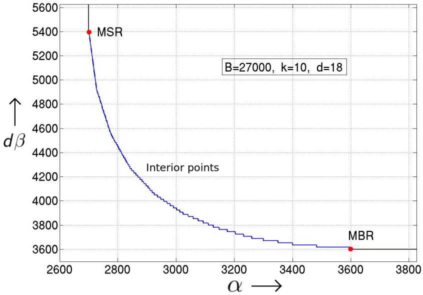
Dimakis, Godfrey, Wu,Wainwright, Ramchandran, T-IT, Sep. 2010
Wu, IEEE JSAC, Feb. 2010.

The Storage-Repair Bandwidth Tradeoff

The upper bound on file size (Dimakis et al.):

$$B \leq \sum_{i=1}^{k} \min\{\alpha, (d-i+1)\beta\} \quad \text{(multiple } (\alpha,\beta) \text{ pairs can achieve bound)}$$

- Tradeoff curve drawn for fixed (k, d), B.
- Extreme points: MSR & MBR
  - MSR=Minimum Storage Regenerating Point
  - MSR=Minimum Bandwidth Regenerating Point
  - MSR codes are MDS codes over the vector alphabet $\mathbb{F}_{q^\alpha}$ and have minimum possible repair bandwidth..



Regenerating Code Constructions

- General Construction for MBR Codes - available
- Construction for MSR Codes - Rate $R \leq \frac{1}{2}$ - available
- Bounds and Constructions for Interior Points
  - much progress
  - improved bounds
  - some constructions
  - but still open!
- Focus here on constructions for high-rate MSR codes

Explicit Constructions of High-Rate MSR Codes

Code	Explicit Construction	Rate	Sub-Packetization Level	Optimal Access	Repair Degree
Ye-Barg 1a	Yes	High	High	No	selectable
Ye-Barg 1b	Yes	High	High	Yes	$(n-1)$
Ye-Barg 2	Yes	High	Low	Yes	$(n-1)$
Li, Tang, Tian	Yes	High	Low	Yes	$(n-1)$
Present Coupled-Layer Construction	Yes	High	Low	No	selectable

Optimal Access: means that helper nodes do not need to do any computation..

(first explicit construction with low sub-packetization
and selectable repair degree, $d \in \{k, k+1, \cdots, (n-1)\}$)

Some Relevant References

1. M. Ye and A. Barg, "Explicit constructions of high-rate MDS array codes with optimal repair bandwidth," CoRR, vol. 1604.00454, April 2016.
2. Jie Li, Xiaohu Tang, Chao Tian, "Enabling All-Node-Repair in Minimum Storage Regenerating Codes, " arXiv:1604.07671, April 2106.
3. M. Ye and A. Barg, "Explicit constructions of optimal-access MDS codes with nearly optimal sub-packetization,"CoRR, vol. abs/1605.08630, May 2016.
4. B. Sasidharan, M. Vajha, and P. V. Kumar, "An explicit, coupled-layer construction of a high-rate MSR code with low sub-packetization level, small field size and all-node repair,"CoRR, vol. abs/1607.07335, July 2016.
5. Birenjith Sasidharan, Myna Vajha, P. Vijay Kumar, "An Explicit, Coupled-Layer Construction of a High-Rate MSR Code with Low Sub-Packetization Level, Small Field Size and $d < (n-1)$," arXiv:1701.07447, Jan. 2017.
6. Jie Li , Jie, Xiaohu Tang, Chao Tian, "A Generic Transformation for Optimal Repair Bandwidth and Rebuilding Access in MDS Codes," *Proc. of the 2017 IEEE Internl. Symp. Inform. Th.*, Aachen, Germany, June 2017.

Talk with Overlapping Results

This talk has some overlap with talk by

- Xiaohu Tang, “MDS codes for distributed storage system,” MMC-17, Solvær, Sep. 6, 2017. Thursday, 4 pm.

Talk with Overlapping Results

This talk has some overlap with talk by

- Xiaohu Tang, "MDS codes for distributed storage system," MMC-17, Solvær, Sep. 6, 2017. Thursday, 4 pm.

Not surprising! This has now become a very competitive field to work in!

Results to be presented are drawn from the January 2017 preprint (later ISIT 2017):

(1) Sasidharan, Vajha, Kumar, "An Explicit, Coupled-Layer Construction of a High-Rate MSR Code with Low Sub-Packetization Level, Small Field Size and $d < (n-1)$," arXiv:1701.07447, Jan. 2017.

Will begin with the case $d = (n-1)$ as this was our first step:

(2) Sasidharan, Vajha, and Kumar, "An explicit, coupled-layer construction of a high-rate MSR code with low sub-packetization level, small field size and all-node repair," CoRR, vol. abs/1607.07335, July 2016.

This first step was our rediscovery in July 2016, of an earlier result by Ye and Barg in May 2016:

(3) Ye and Barg, "Explicit constructions of optimal-access MDS codes with nearly optimal sub-packetization," CoRR, vol. abs/1605.08630, May 2016.

We will present this using a slightly different, coupled-layer perspective from that in (3).

As will be seen, both (2) and (3) overlap with the work by Li-Tang-Tian in :

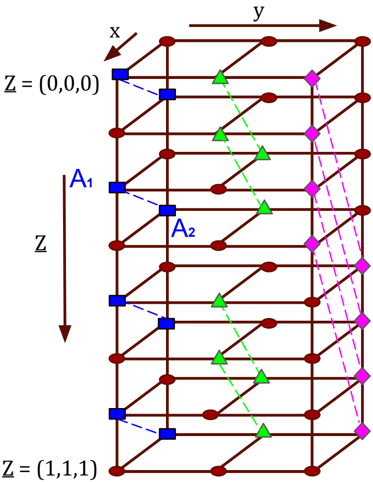
(4) Li, Tang and Tian, "Enabling All-Node-Repair in Minimum Storage Regenerating Codes, " arXiv:1604.07671, April 2106.

(5) Li , Tang and Tian, "A Generic Transformation for Optimal Repair Bandwidth and Rebuilding Access in MDS Codes," *Proc. of the 2017 IEEE Internl. Symp. Inform. Th.*, Aachen, Germany, June 2017.

---

(6) Vajha, Kini, Puranik, Ramkumar, Lobo, Sasidharan, Kumar, Ye, Barg, Hussain, Narayanamurthy, and Nandi, "Pairing up for Regeneration: The Mantra for Fast and Efficient Node Repair," poster presentation at *USENIX ATC 2017*. (Emulation)

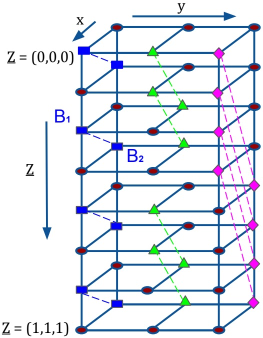
Coupled-Layer Construction: Codeword as a Data Cube



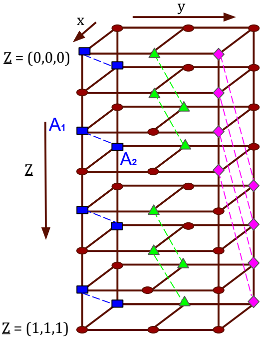
- Each vertical column corresponds to:
  - a storage node
  - a vector code symbol
- the (x, y) indexing of vector code symbols is simply for convenience..
- Thus each codeword is of the form:

$$\{ \quad \underbrace{A(x, y; \underline{z})}_{\text{vector code symbol}} \quad | \, (x, y)\}$$

Coupled-Layer Construction: Virtual and Real Data Cubes



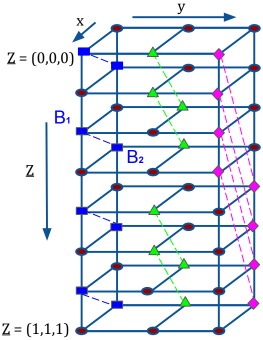
Virtual data cube B.

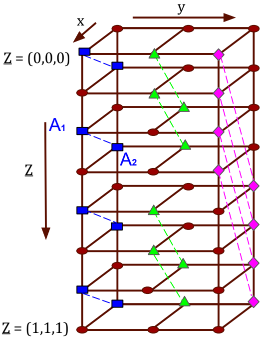


Real data cube A.

As we shall see, the virtual data cube $B(x, y; \underline{z})$ will assist in constructing the real data cube $A(x, y; \underline{z})$.

Structure of the Real and Virtual Data Cubes

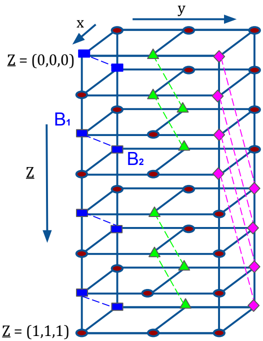




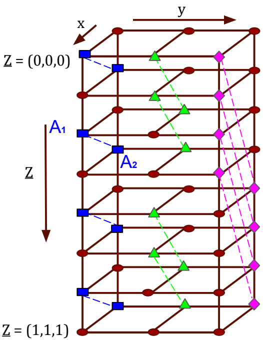
Step 1: Each layer of the B data cube on left is an independent MDS code.

Step 2: The A data cube on the right is obtained by pairwise coupling across layers..

Coupled-Layer Construction: Pairwise-Coupling Transform



Step 1: Independent Layers



Step 2: Coupled Layers

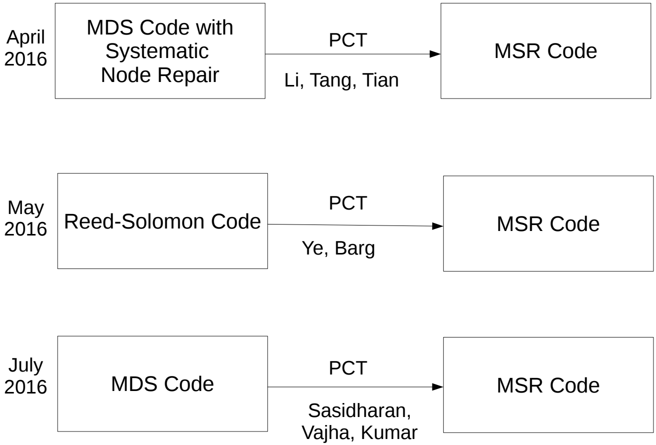
Pairwise Coupling Transformation (PCT):

$$\begin{bmatrix} A_1 \\ A_2 \end{bmatrix} = \begin{bmatrix} 1 & \gamma \\ \gamma & 1 \end{bmatrix} \begin{bmatrix} B_1 \\ B_2 \end{bmatrix},$$

replace (B_1, B_2) by (A_1, A_2) etc. to get the data cube on the right.

Pairwise Coupling Transformation (PCT)

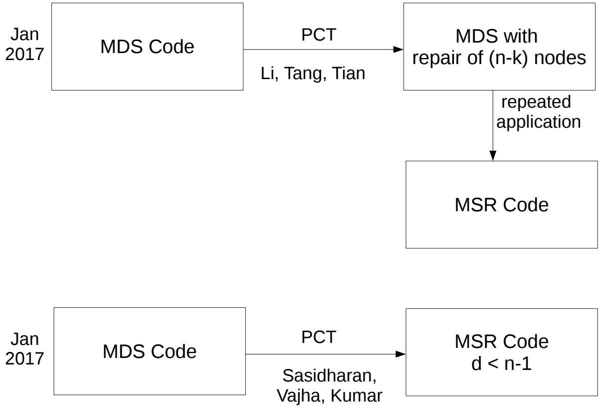
Independent discoveries of same transformation (in retrospect)



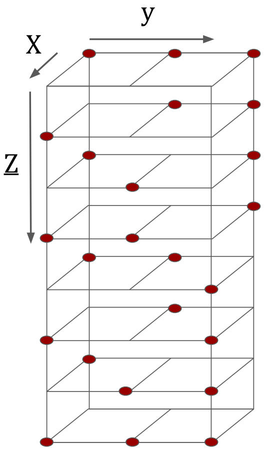
* Coupled-layer perspective of the Ye-Barg construction was introduced in July 2016 work (arXiv).

Pairwise Coupling Transformation (PCT)

Most recently, in ISIT 2017:



Coupled-Layer Construction: Parameters of an Example Construction



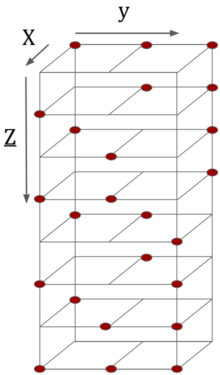
General Parameters:
$\{(n, k, d),\ (\alpha, \beta),\ B,\ \mathbb{F}_q\ \}$

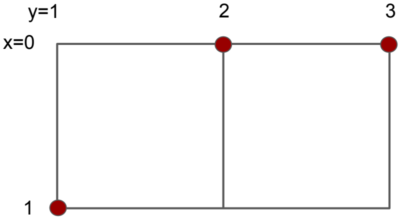
Example Parameters:
($n = 6$ nodes, $k = 4$, $d = 5$),

($\alpha = 8$ sub-packetization level)
($\beta = 4$ symbols downloaded from each helper node)

(file size $B = 32$)
(field size $\mathbb{F} = 7 \Rightarrow \mathbb{F}_7$)

Notation to Identify the 8 Layers





(layer $\underline{z} = (1, 0, 0)$ identified through the placement of red dots in the appropriate coordinates)

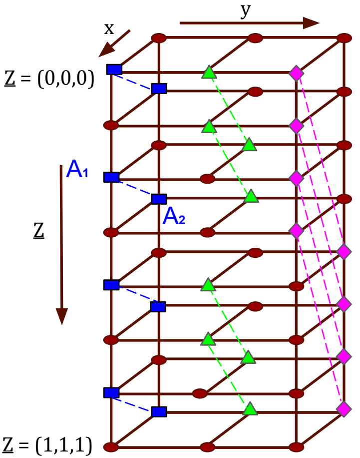
Mathematical Notation for the Companion Code Symbol

Obtaining the companion :

$$A^c(x,y;\underline{z}) \quad = \quad \underbrace{A(x,y;\underline{z})}_{\text{swap } x \text{ and } z_y}$$

$$({\color{red}1},1;\ {\color{red}0}00) \quad \Leftrightarrow \quad ({\color{red}0},1;\ {\color{red}1}00)$$

Graphically $\Rightarrow$:



Similarly,

$$B^c(x,y;\underline{z}) \quad = \quad \underbrace{B(x,y;\underline{z})}_{\text{swap } x \text{ and } z_y}\ .$$

Pairwise Coupling Transformation

$$\begin{bmatrix} A(x,y;\underline{z}) \\ A^c(x,y;\underline{z}) \end{bmatrix} = \begin{bmatrix} 1 & \gamma \\ \gamma & 1 \end{bmatrix}^{-1} \begin{bmatrix} B(x,y;\underline{z}) \\ B^c(x,y;\underline{z}) \end{bmatrix}$$

$$\begin{bmatrix} B(x,y;\underline{z}) \\ B^c(x,y;\underline{z}) \end{bmatrix} = \begin{bmatrix} 1 & \gamma \\ \gamma & 1 \end{bmatrix} \begin{bmatrix} A(x,y;\underline{z}) \\ A^c(x,y;\underline{z}) \end{bmatrix}.$$

Can verify that any two of

$$\{A(x,y;\underline{z})\ A^c(x,y;\underline{z})\ B(x,y;\underline{z})\ B(x,y;\underline{z})\ \},$$

suffice to recover the other two.

Parity-Check Constraints Satisfied by the Code

Parity-Check Constraint: Each layer of the B code must satisfy the $(n-k)$ constraints of some scalar or vector MDS code...

$$\sum_{y \in [t]} \sum_{x \in \mathbb{Z}_u} h_\lambda(x,y) B(x,y;\underline{z}) \;\; = \;\; 0.$$

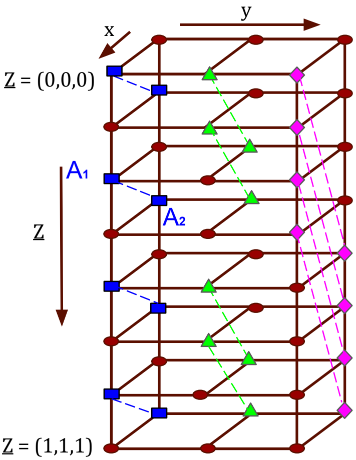
The constraints on the A code are thus, expressed in terms of the B code.

Encoding and Node repair

We provide an animation-based overview of systematic encoding and of node repair.

Data Collection: We Adopt a Layer-by-Layer Approach

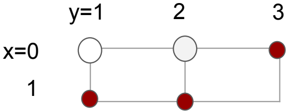
Recall: Data Collection Means Recovery from $(n-k)$Erasures



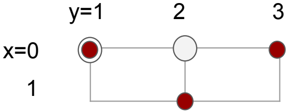
- Proceeds Layer-by-Layer
- in order of increasing intersection score $\Downarrow$

Data Collection: Intersection Score of Layer

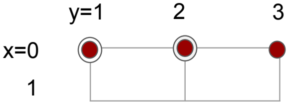
1. Assume $(n-k)=2$ nodes are erased.
2. Erasures are indicated by unfilled circle.
3. Layers viewed from above



Intersection score $=0$

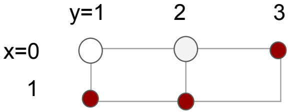


Intersection score $=1$



Intersection score $=2$

Data Collection: Case of Intersection Score Zero



Intersection Score $= 0$

A(0,1) erased	A(0,2) erased	A(0,3)
A(1,1)	A(1,2)	A(1,3)

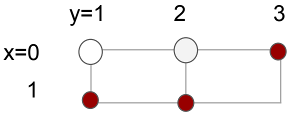
$$+$$

$A^c(0,1)$	$A^c(0,2)$	$A^c(0,3)$
$A^c(1,1)$	$A^c(1,2)$	$A^c(1,3)$

$$\Downarrow$$

$B(0,1) = ?$	$B(0,2) = ?$	B(0,3)
B(1,1)	B(1,2)	B(1,3)

Data Collection: Case of Intersection Score Zero



Intersection Score $= 0$

Solve for A(0,1)	**Sole for A(0,2)**	A(0,3)
A(1,1)	A(1,2)	A(1,3)

$\Uparrow$

$A^c(0,1)$	$A^c(0,2)$	$A^c(0,3)$
$A^c(1,1)$	$A^c(1,2)$	$A^c(1,3)$

$+$

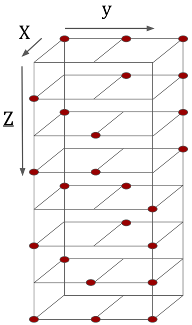
Solve for B(0,1)	**Solve for B(0,2)**	B(0,3)
B(1,1)	B(1,2)	B(1,3)

Solve using MDS code

Data Collection: Case of Intersection Score > 0

- Can be similarly computed
- Given that layers of lesser intersection score have already been decoded

The Coupled-Layer Construction for $d < (n-1)$



- Can be viewed as a subcode of the coupled-layer $d = (n-1)$ construction.
- obtained by adding parity-check constraints on contents of each node
- data in each node is now constrained to be a codeword in an

$$[(n-k)\beta,\ (d-k+1)\beta]$$

MDS code.

- Parameters before nodal constraints

$$\{(n, k, d_0 = (n-1)), (\alpha_0 = (d_0 - k + 1)\beta, \beta), B_0 = \alpha_0 k, \mathbb{F}_q\}$$

- Parameters after nodal constraints (only α, B change)

$$\{(n, k, d_1 < (n-1)), (\alpha_1 = (d_1 - k + 1)\beta, \beta), B_1 = \alpha_1 k, \mathbb{F}_q\}$$

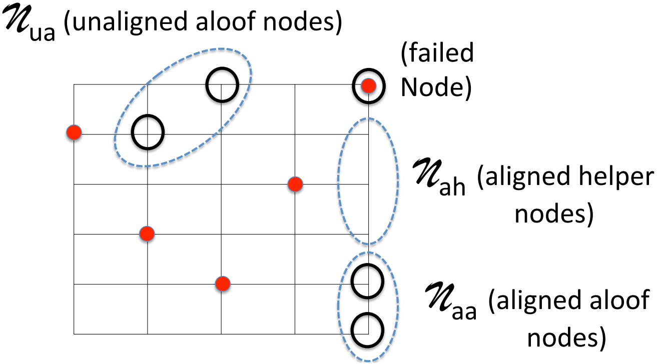
Example Parameters

$$\left(\ n = uv,\ \ k = u(v-1),\ \ d = n-1-a\ \right),$$
$$\left(\alpha = (u-a)\cdot u^{v-1}, \beta = u^{v-1}\right) \text{ and } \quad q \leq n\ .$$

We note that MSR codes having any (n, k, d) can be obtained through shortening. This is realized by first constructing the MSR code given by parameters $u = n-k$, $v = \lceil \frac{n}{u} \rceil$ and $a = n-1-d$ and then shortening it by $\Delta = uv - n$ symbols.

n	k	$d < (n-1)$	α
12	8	9	32
11	8	9	54
10	6	7	32
16	12	13	128
24	16	20	192

Node Repair



$\mathcal{N}_{\text{ua}}$ (unaligned aloof nodes)
(failed Node)
$\mathcal{N}_{\text{ah}}$ (aligned helper nodes)
$\mathcal{N}_{\text{aa}}$ (aligned aloof nodes)

Node Repair and Data Collection

1. Presence of nodal parity does not impact data collection.
2. The two sets of parity are designed so as to work together and permit some nodes to remain aloof.

Thanks!