

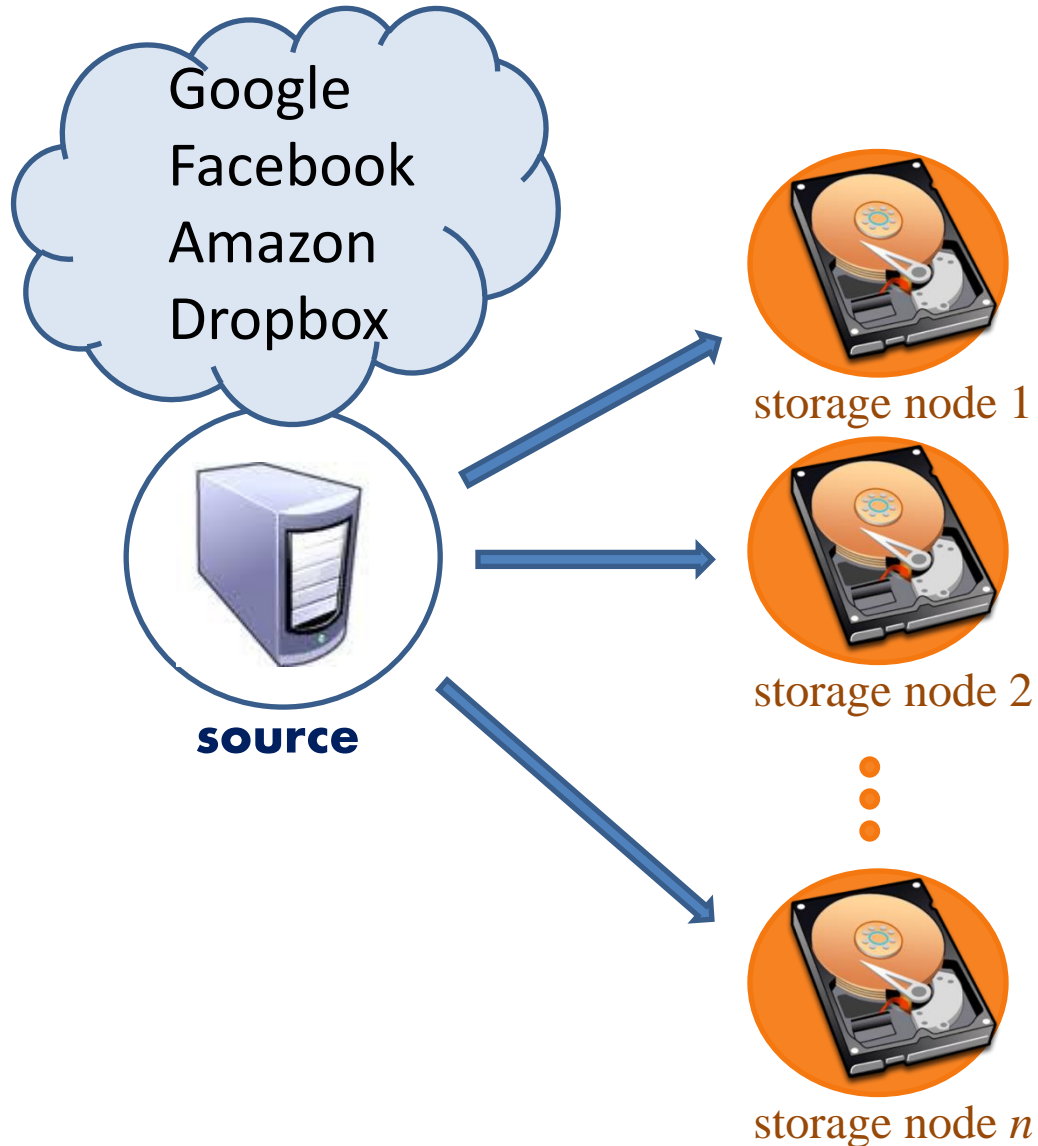
Optimal Fractional Repetition Codes for Distributed Storage Systems

Natalia Silberstein

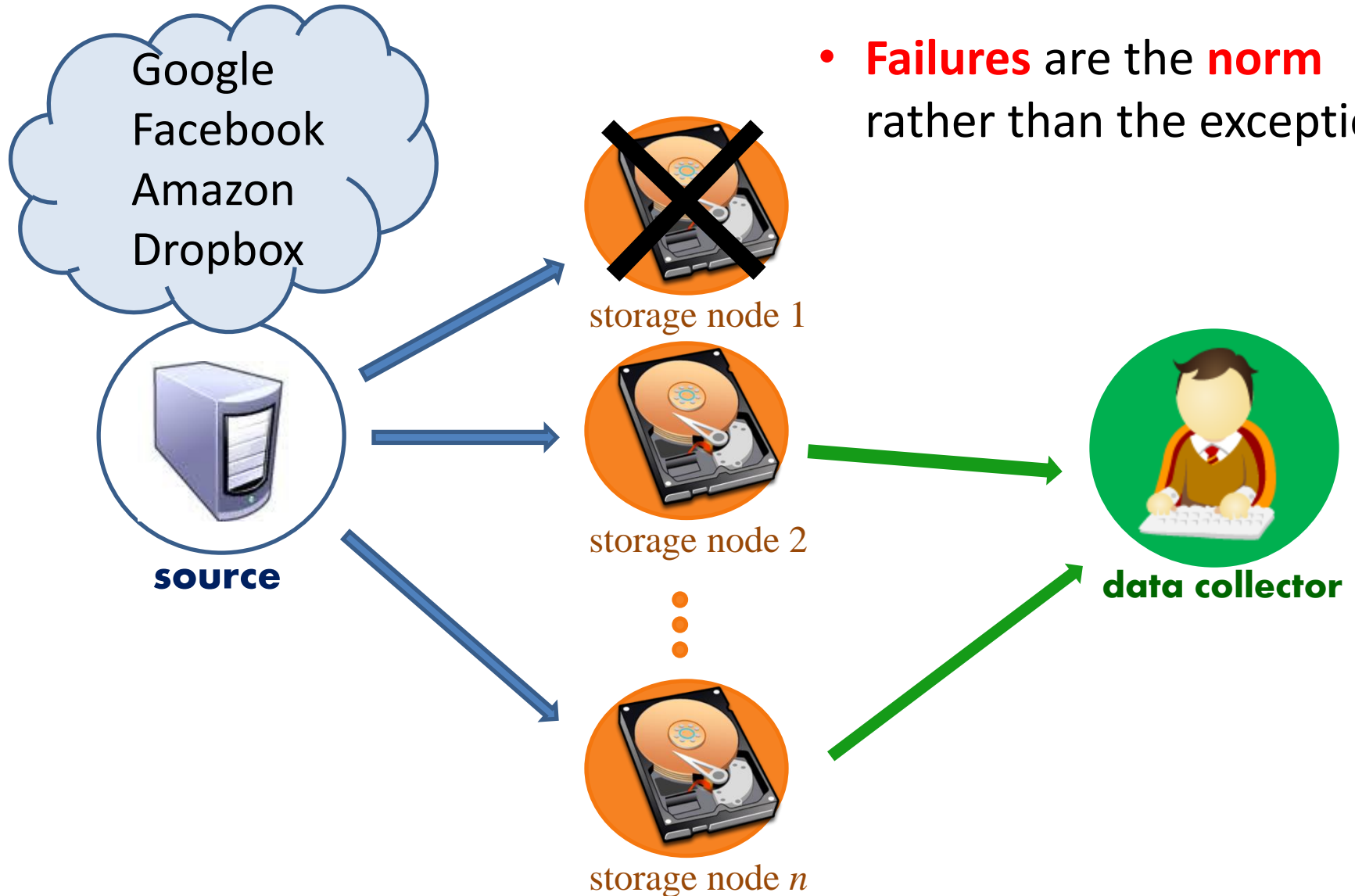
Joint work with Tuvi Etzion



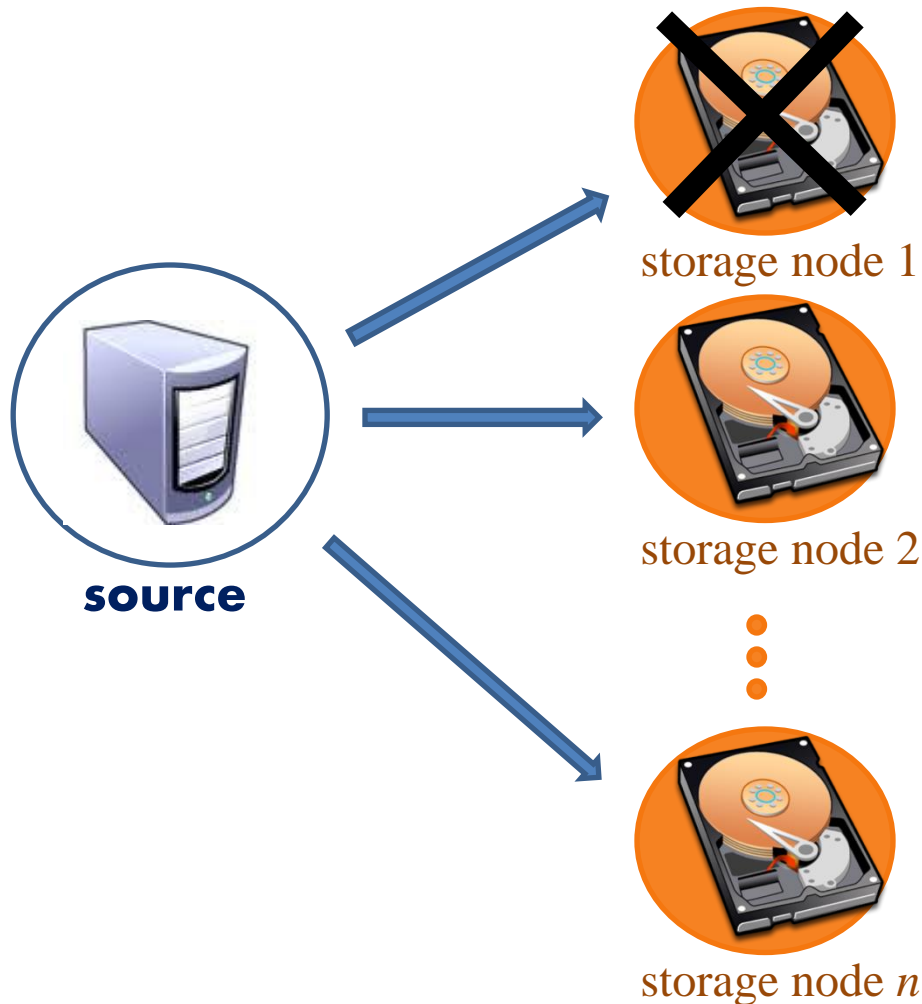
Distributed storage system (DSS)



Distributed storage system (DSS)



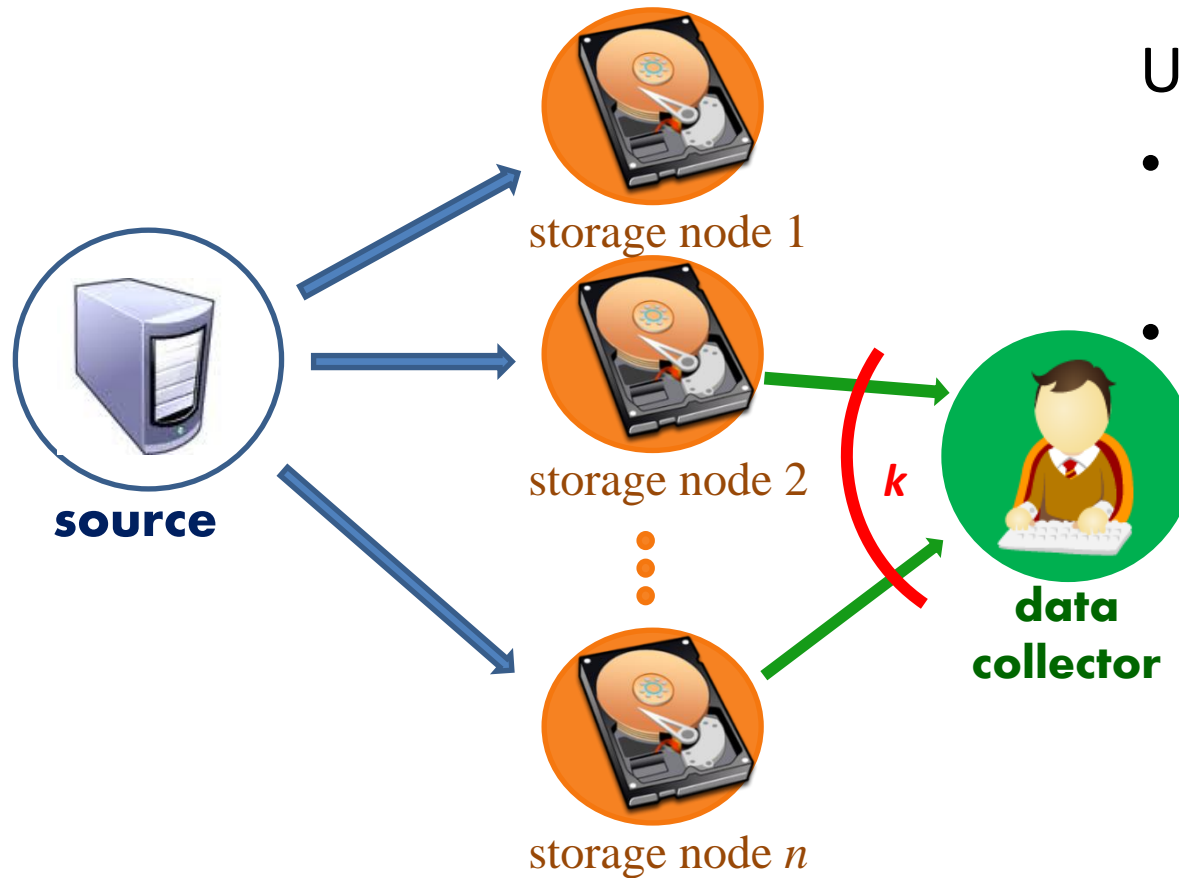
Distributed storage system (DSS)



- **Failures** are the **norm** rather than the exception

- **Redundancy** for reliability
 - **Replication**
 - **Erasure coding**

Coding for distributed storage system



- Erasure codes:

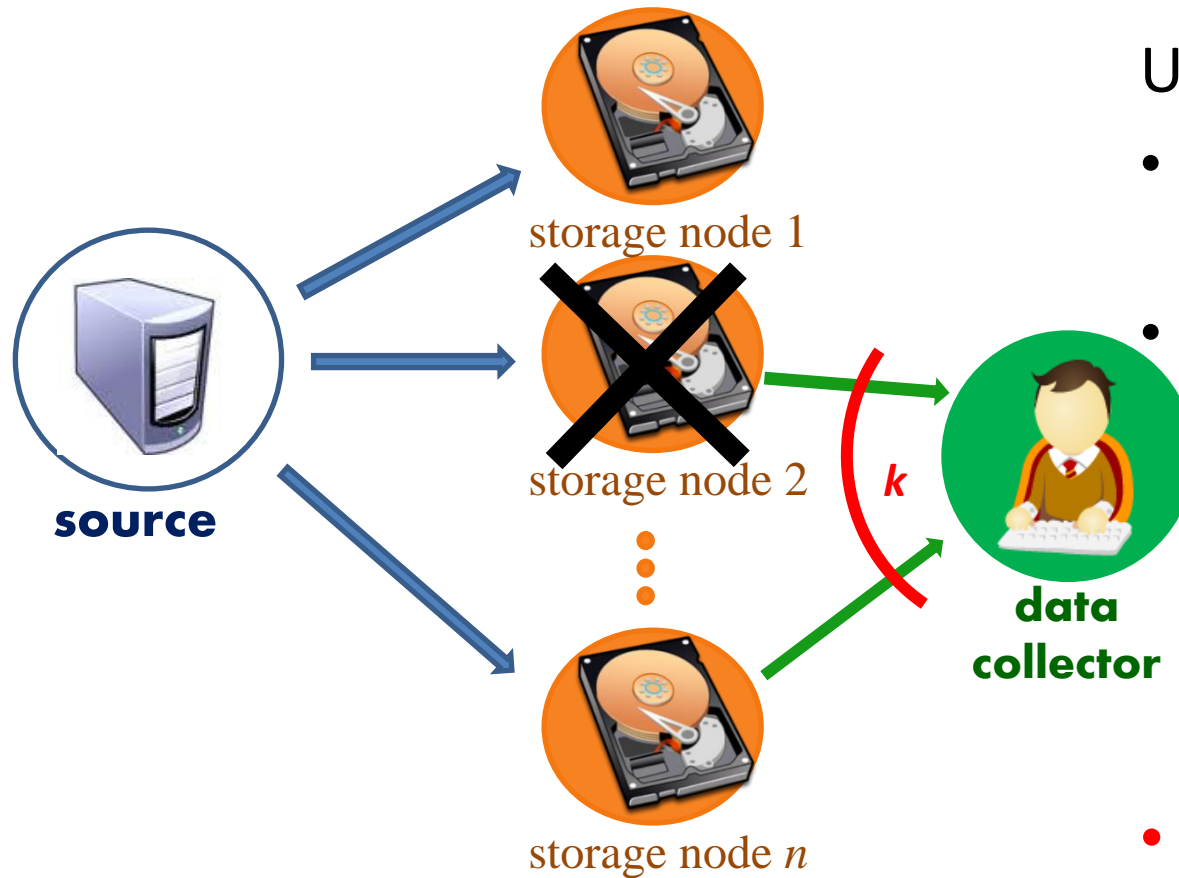
Using (n, k) erasure code:

- Partition the data into k blocks
- Generate n blocks, store each block in a new node

(n, k) MDS property:

Reconstruct the data from any k nodes

Coding for distributed storage system



- Erasure codes:

Using (n, k) erasure code:

- Partition the data into k blocks
- Generate n blocks, store each block in a new node

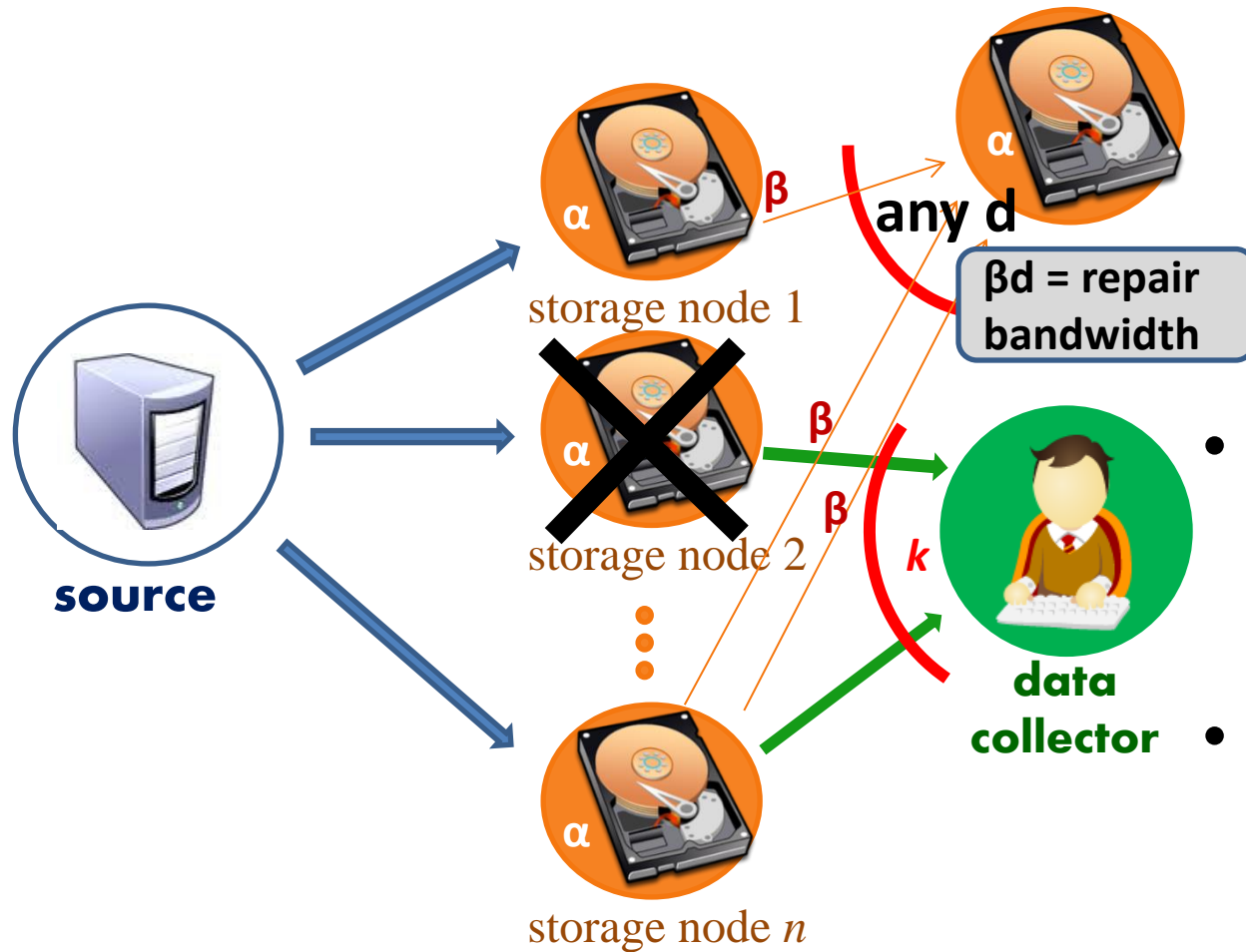
- (n, k) MDS property:

Reconstruct the data from any k nodes

- Regenerating codes

efficient node repairs

Regenerating codes (*)



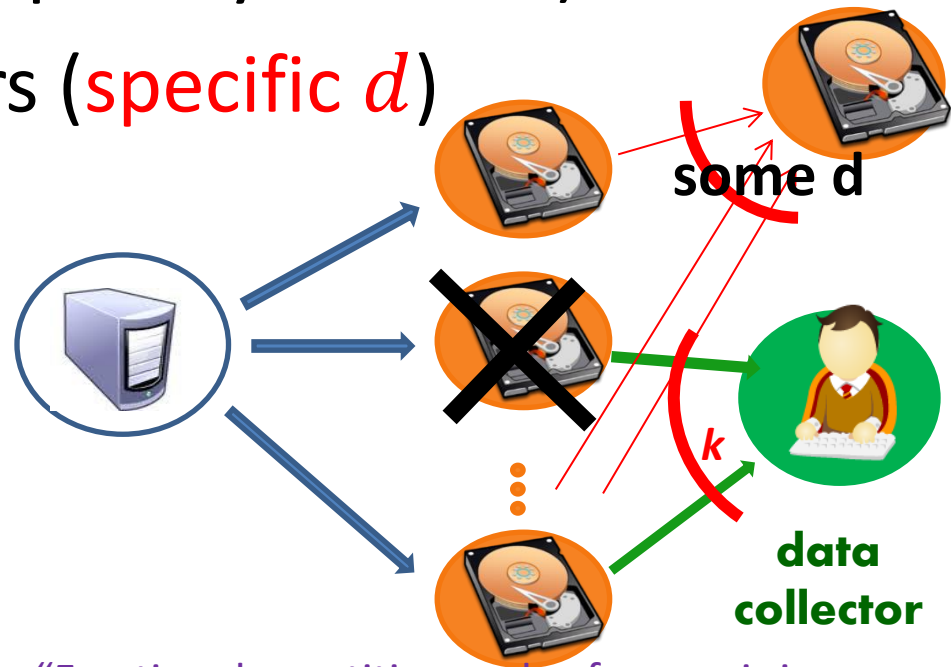
- Minimum **storage** regenerating (**MSR**) codes
- Minimum **bandwidth** regenerating (**MBR**) codes

(*) A. G. Dimakis, P. B. Godfrey, M. J. Wainwright and K. Ramchandran, "Network coding for distributed storage systems, 2007"

DRESS codes^(*)

(Distributed Replication based Exact Simple Storage)

- Minimum repair bandwidth (like MBR)
- **Uncoded** repair (repair by transfer)
- Table based repairs (*specific d*)



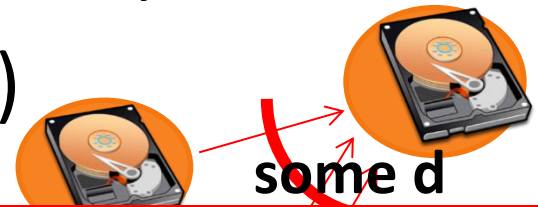
(*) S. El Rouayheb and K. Ramchandran, "Fractional repetition codes for repair in distributed storage systems", 2010

S. Pawar, N. Noorshams, S. El Rouayheb, and K. Ramchandran, "DRESS codes for the storage cloud: Simple randomized constructions", 2013.

DRESS codes^(*)

(Distributed Replication based Exact Simple Storage)

- Minimum repair bandwidth (like MBR)
- **Uncoded** repair (repair by transfer)
- Table based repairs (*specific d*)



Allow to store **more** data
than **MBR** codes!



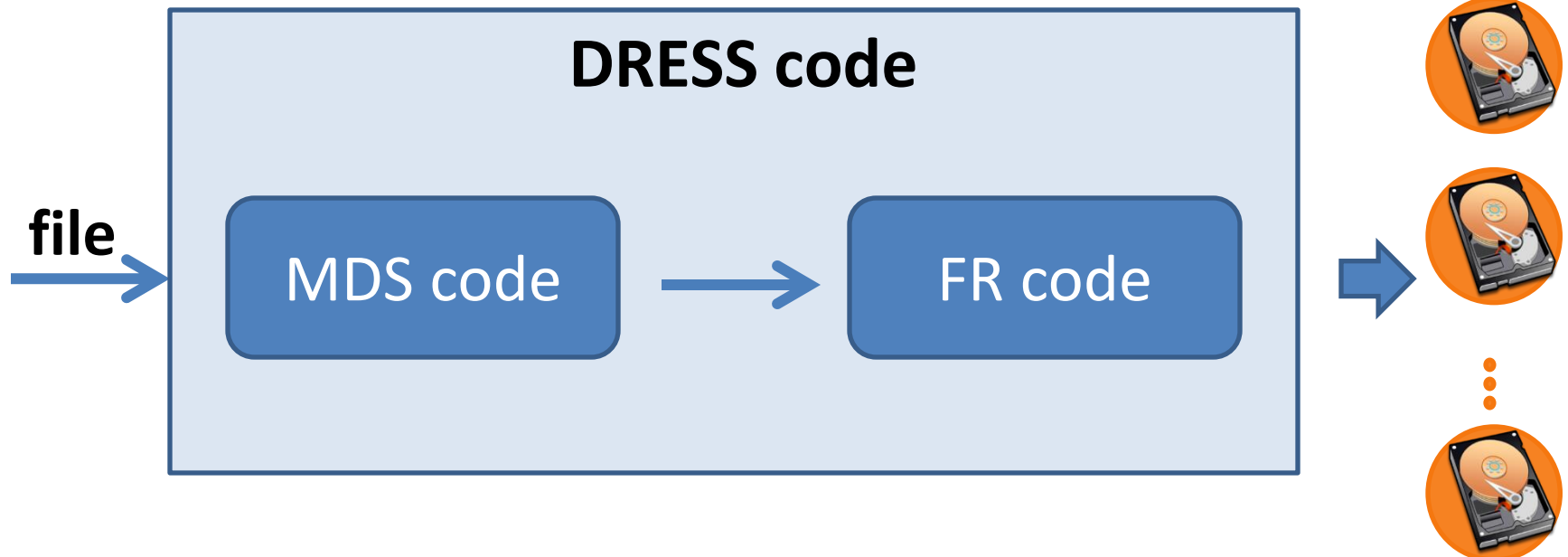
(*) S. El Rouayheb and K. Ramchandran, “Fractional repetition codes for repair in distributed storage systems”, 2010

S. Pawar, N. Noorshams, S. El Rouayheb, and K. Ramchandran, “DRESS codes for the storage cloud: Simple randomized constructions “ , 2013.

DRESS codes^(*)

(Distributed Replication based Exact Simple Storage)

- consist of the concatenation of
 - **Outer** maximum distance separable (**MDS**) code
 - And **inner** fractional repetition (**FR**) code.



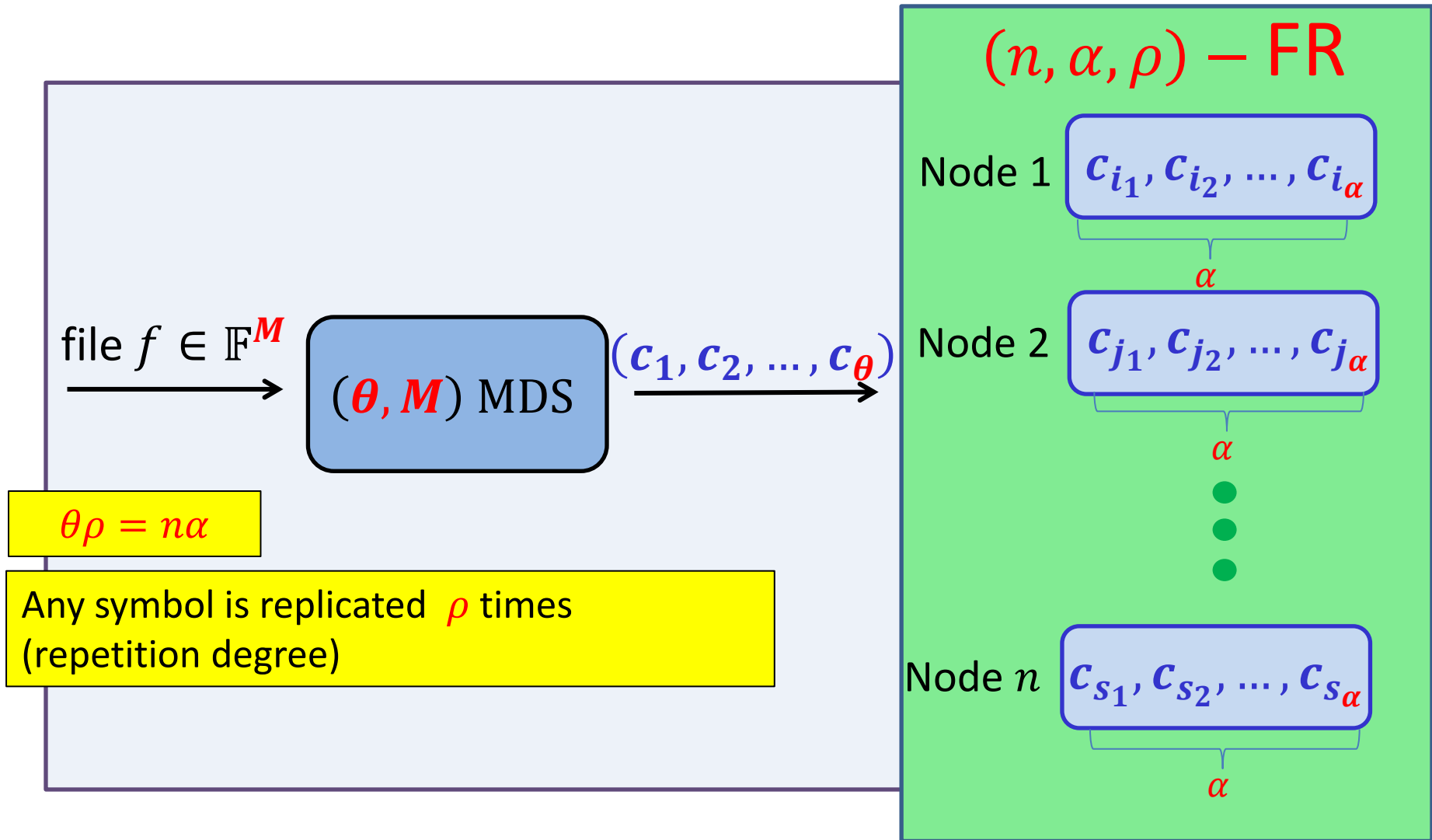
FR code: definition

- An (n, α, ρ) FR code C is a collection of n subsets N_1, \dots, N_n of $[\theta]$, for $n\alpha = \rho\theta$, such that
 - $|N_i| = \alpha$, for $1 \leq i \leq n$;
 - each element of $[\theta]$ belongs to exactly ρ subsets in C .

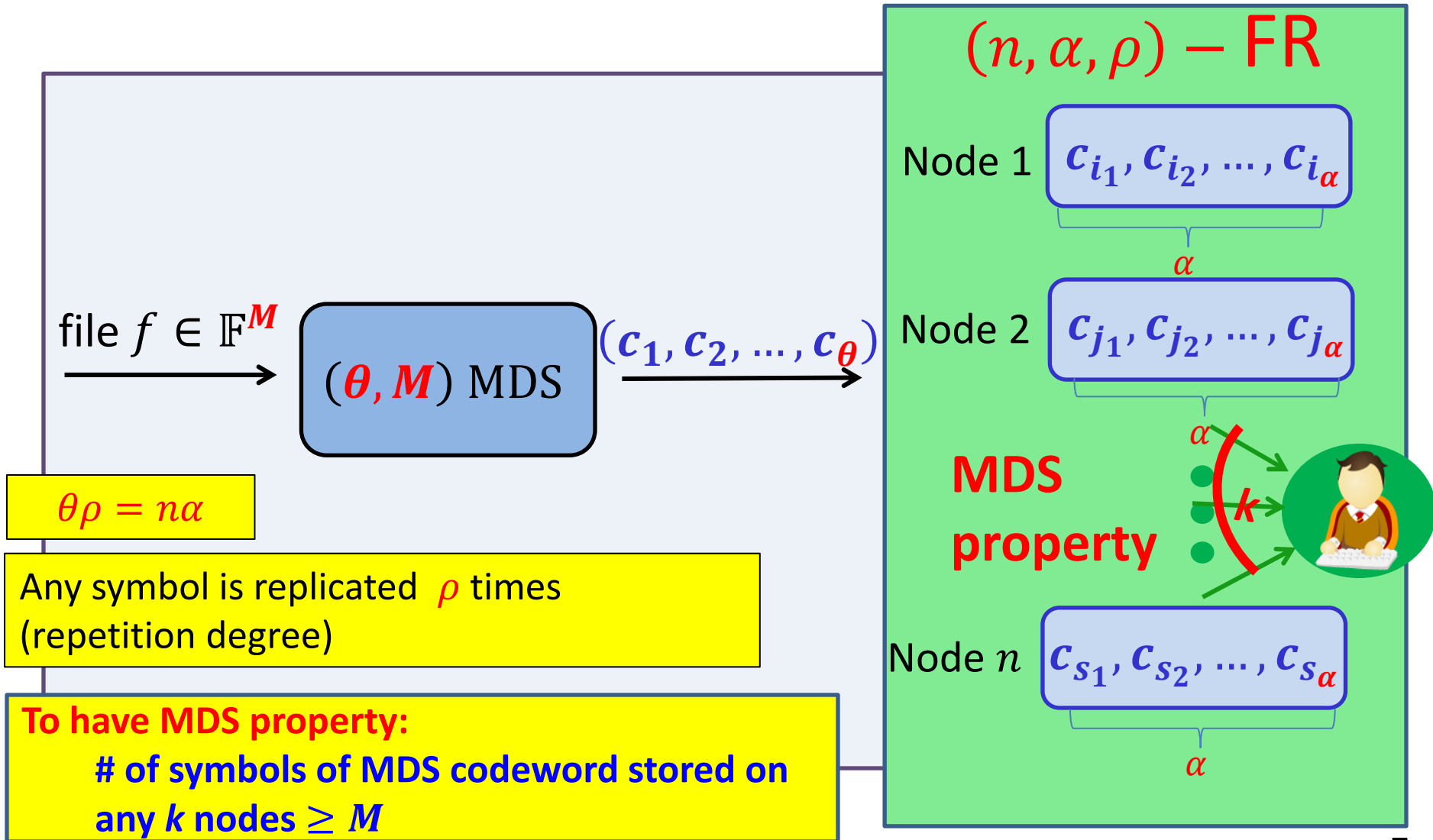
FR code: definition

- An (n, α, ρ) FR code C is a collection of n subsets N_1, \dots, N_n of $[\theta]$, for $n\alpha = \rho\theta$, such that
 - $|N_i| = \alpha$, for $1 \leq i \leq n$;
 - each element of $[\theta]$ belongs to exactly ρ subsets in C .
- Node i stores the symbols of MDS codeword indexed by N_i

$[(\theta, M), k, (n, \alpha, \rho)]$ - DRESS code



$[(\theta, M), k, (n, \alpha, \rho)]$ - DRESS code



FR code: MDS property

- **To have MDS property:**

of symbols of MDS codeword stored on any k nodes $\geq M$

$$\Leftrightarrow \left| \bigcup_{i \in I, |I|=k} N_i \right| \geq M$$

FR code: MDS property

- **To have MDS property:**

of symbols of MDS codeword stored on any k nodes $\geq M$

$$\Leftrightarrow \left| \bigcup_{i \in I, |I|=k} N_i \right| \geq M$$

- The same FR code can be used in many DRESS codes, with different k 's

FR code: MDS property

- **To have MDS property:**

of symbols of MDS codeword stored on any k nodes $\geq M$

$$\Leftrightarrow \left| \bigcup_{i \in I, |I|=k} N_i \right| \geq M$$

- The same FR code can be used in many DRESS codes, with different k 's
- For a given k , denote the file size $M(k)$

FR code: MDS property

- **To have MDS property:**

of symbols of MDS codeword stored on any k nodes $\geq M$

$$\Leftrightarrow \left| \bigcup_{i \in I, |I|=k} N_i \right| \geq M$$

- The same FR code can be used in many DRESS codes, with different k 's
- For a given k , denote the file size $M(k)$
- Want to store more than MBR code:

$$M(k) > k\alpha - \binom{k}{2}$$

MBR

DRESS code: maximum file size

- Denote by $A(n, k, \alpha, \rho)$ the upper bound on $M(k)$ for an $[(\theta, M(k)), k, (n, \alpha, \rho)]$ - DRESS code

$A(n, k, \alpha, \rho) \leq \varphi(k)$, where

$$\varphi(1) = \alpha, \varphi(k + 1) = \varphi(k) + \alpha - \left\lceil \frac{\rho\varphi(k) - k\alpha}{n - k} \right\rceil^{(*)}$$

(*) Salim El Rouayheb and Kannan Ramchandran, “Fractional repetition codes for repair in distributed storage systems”, 2010

DRESS code: maximum file size

- Denote by $A(n, k, \alpha, \rho)$ the upper bound on $M(k)$ for an $[(\theta, M(k)), k, (n, \alpha, \rho)]$ - DRESS code

$A(n, k, \alpha, \rho) \leq \varphi(k)$, where

$$\varphi(1) = \alpha, \varphi(k + 1) = \varphi(k) + \alpha - \left\lceil \frac{\rho\varphi(k) - k\alpha}{n - k} \right\rceil^{(*)}$$



Is it possible to achieve this bound?

(*) Salim El Rouayheb and Kannan Ramchandran, “Fractional repetition codes for repair in distributed storage systems”, 2010

DRESS code: maximum file size

- Denote by $A(n, k, \alpha, \rho)$ the upper bound on $M(k)$ for an $[(\theta, M(k)), k, (n, \alpha, \rho)]$ - DRESS code

$A(n, k, \alpha, \rho) \leq \varphi(k)$, where

$$\varphi(1) = \alpha, \varphi(k + 1) = \varphi(k) + \alpha - \left\lceil \frac{\rho\varphi(k) - k\alpha}{n-k} \right\rceil^{(*)}$$



Is it possible to achieve this bound?

- $A(n, k, \alpha, \rho)$ is determined by the parameters of the inner FR code, for a given k .
- **Optimal FR** code stores maximum possible file

(*) Salim El Rouayheb and Kannan Ramchandran, “Fractional repetition codes for repair in distributed storage systems”, 2010

Known constructions of FR codes

- S. El Rouayheb and K. Ramchandran, “*Fractional repetition codes for repair in distributed storage systems*,” Allerton 2010.
- J. C. Koo and J. T. Gill. III, “*Scalable constructions of fractional repetition codes in distributed storage systems*,” Allerton 2011.
- O. Olmez and A. Ramamoorthy, “*Repairable replication-based storage systems using resolvable designs*,” Allerton 2012.
- S. Anil, M. K. Gupta, and T. A. Gulliver, “*Enumerating some fractional repetition codes*,” arXiv 2013.
- S. Pawar, N. Noorshams, S. El Rouayheb, and K. Ramchandran, “*Dress codes for the storage cloud: Simple randomized constructions*,” ISIT 2013.
- B. Zhu, K. W. Shum, H. Li, and H. Hou, “*General fractional repetition codes for distributed storage systems*” IEEE Communications Letters, Apr. 2014.

Our results: optimal FR codes

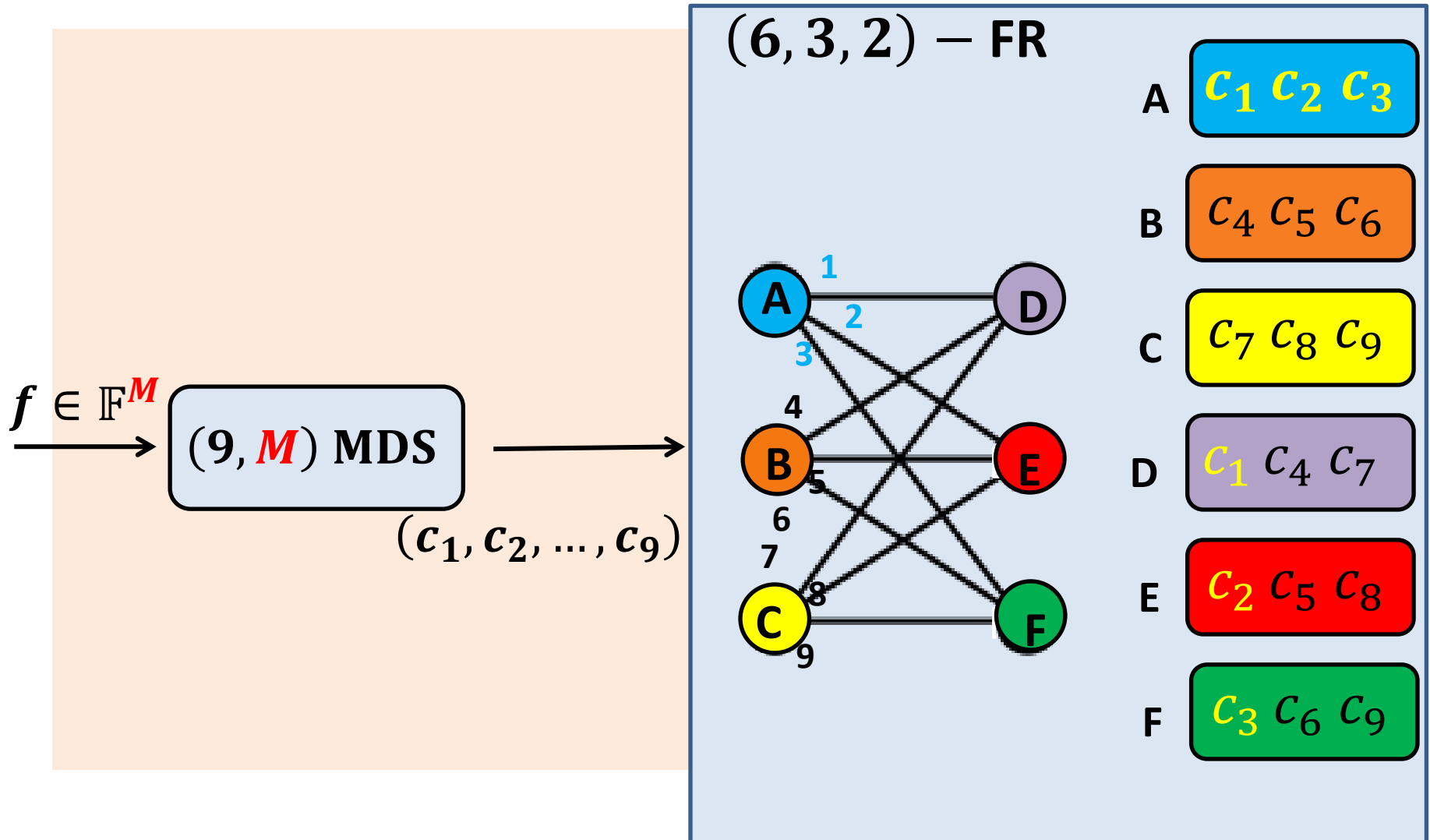
- $\rho = 2$:
 - Based on complete r -partite graphs (Turán graphs)
 - Based on regular graphs with a given girth (e.g. cage graphs)
- $\rho > 2$:
 - Based on transversal designs
 - Based on biregular bipartite graphs with a given girth (e.g. generalized polygons)

Our results: optimal FR codes

- $\rho = 2$:
 - Based on complete r -partite graphs (Turán graphs)
 - Based on regular graphs with a given girth (e.g. cage graphs)
- $\rho > 2$:
 - Based on transversal designs
 - Based on biregular bipartite graphs with a given girth (e.g. generalized polygons)

Optimal FR codes with $\rho = 2$:

Example: Bipartite graph based FR code



Optimal FR codes with $\rho = 2$:

Example: Bipartite graph based FR code

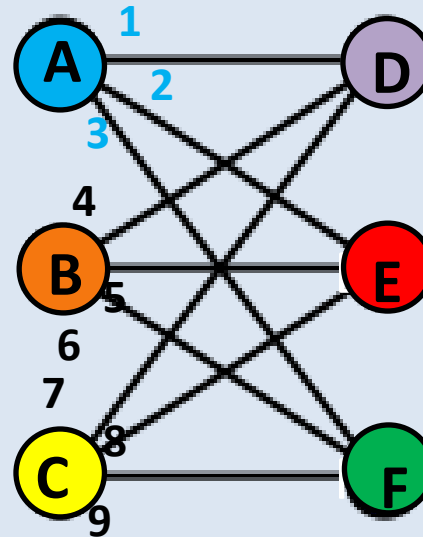
k	$M(k)$	MBR
1	3	3
2	5	5
3	7	6

$f \in \mathbb{F}^M$

(9, M) MDS

(c_1, c_2, \dots, c_9)

(6, 3, 2) – FR



A

$c_1 c_2 c_3$

B

$c_4 c_5 c_6$

C

$c_7 c_8 c_9$

D

$c_1 c_4 c_7$

E

$c_2 c_5 c_8$

F

$c_3 c_6 c_9$

Optimal FR codes with $\rho = 2$:

- α -regular graph $G = (V, E) \Leftrightarrow$ FR code C_G with $\rho = 2$

Vertex $v \in V \Leftrightarrow$ node N_i

Edge $e \in E \Leftrightarrow$ symbol of an MDS code

$|V| = n =$ number of nodes

$|E| = \theta =$ length of MDS code

Degree $\alpha =$ storage α in a node

Optimal FR codes with $\rho = 2$:

Lemma 1

The file size $M(k)$ of C_G based on $G = (V, E)$ is given by

$$M(k) = k\alpha - \max_{G'=(V',E') \in G_k} |E'|,$$

where G_k is the family of induced subgraphs of G with k vertices.

Optimal FR codes with $\rho = 2$:

Lemma 1

The file size $M(k)$ of C_G based on $G = (V, E)$ is given by

$$M(k) = k\alpha - \max_{G'=(V',E') \in G_k} |E'|,$$

where G_k is the family of induced subgraphs of G with k vertices.

Corollary 1

A graph G contains a k -clique *iff* $M(k) = k\alpha - \binom{k}{2}$ for C_G .

MBR

Optimal FR codes with $\rho = 2$:

Lemma 1

The file size $M(k)$ of C_G based on $G = (V, E)$ is given by

$$M(k) = k\alpha - \max_{G'=(V',E') \in G_k} |E'|,$$

where G_k is the family of induced subgraphs of G with k vertices.

Corollary 1

A graph G contains a k -clique *iff* $M(k) = k\alpha - \binom{k}{2}$ for C_G .

Corollary 2

FR code C_G stores **more** than MBR code *iff* G does not contain a k -clique.

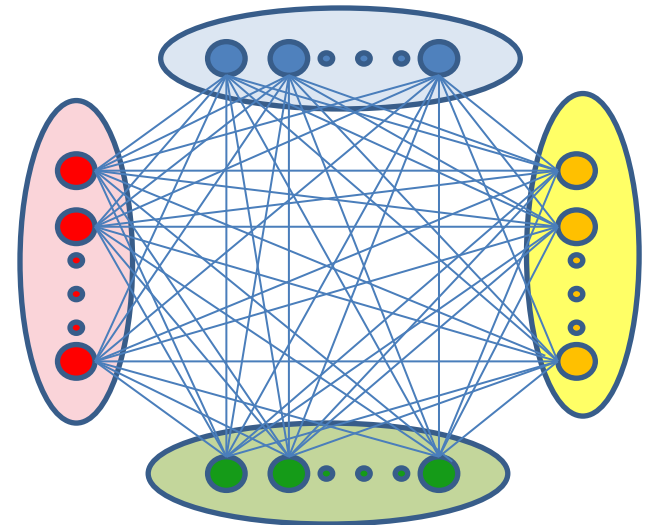
MBR

Turàn graph based FR code

- (n, r) -Turàn graph T is a complete r -partite graph, $r|n$:

- Does not contain a $(r+1)$ -clique
- Regular graph of degree

$$\alpha = (r - 1) \frac{n}{r}$$



Turán graph based FR code

- (n, r) -Turán graph T is a complete r -partite graph, $r|n$:

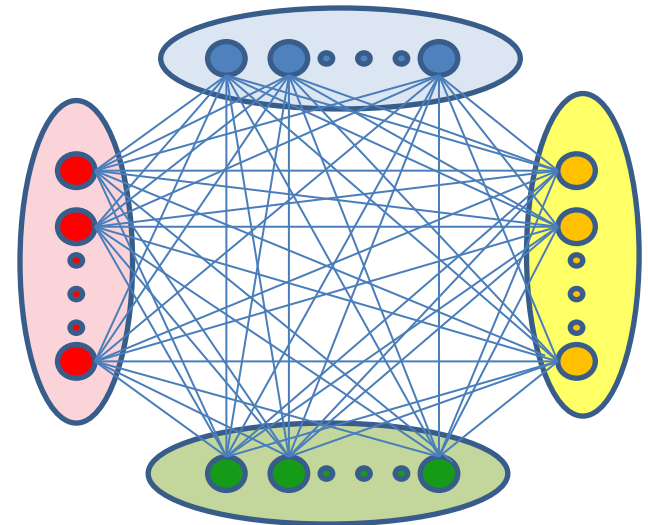
- Does not contain a $(r+1)$ -clique
- Regular graph of degree

$$\alpha = (r - 1) \frac{n}{r}$$

- The file size of the FR code C_T :

$$M_{C_T}(k) = k\alpha - \binom{k}{2} + r \binom{b}{2} + bt$$

where $b = \left\lfloor \frac{k}{r} \right\rfloor$, $t \equiv k \pmod{r}$



by Lemma 1: $M(k) = k\alpha - \max_{G'=(V',E') \in G_k} |E'|$

Turàn graph based FR code

- (n, r) -Turàn graph T is a complete r -partite graph, $r|n$:

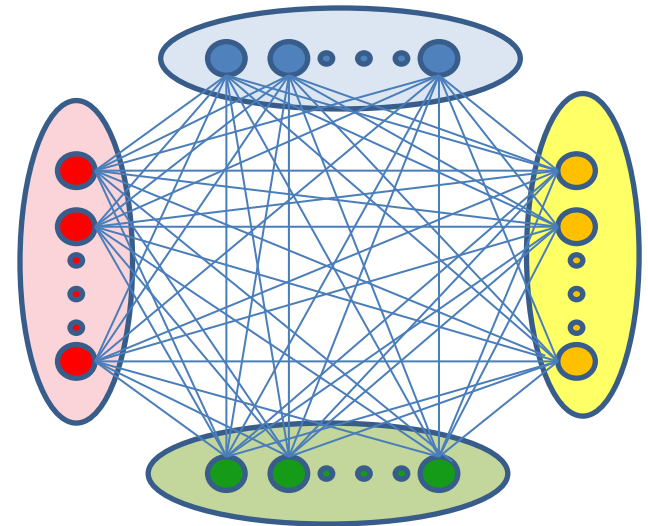
- Does not contain a $(r+1)$ -clique
- Regular graph of degree

$$\alpha = (r - 1) \frac{n}{r}$$

- The file size of the FR code C_T :

$$M_{C_T}(k) = k\alpha - \binom{k}{2} + r \binom{b}{2} + bt$$

where $b = \left\lfloor \frac{k}{r} \right\rfloor$, $t \equiv k \pmod{r}$



MBR

Turàn graph based FR code

- (n, r) -Turàn graph T is a complete r -partite graph, $r|n$:

- Does not contain a $(r+1)$ -clique
- Regular graph of degree

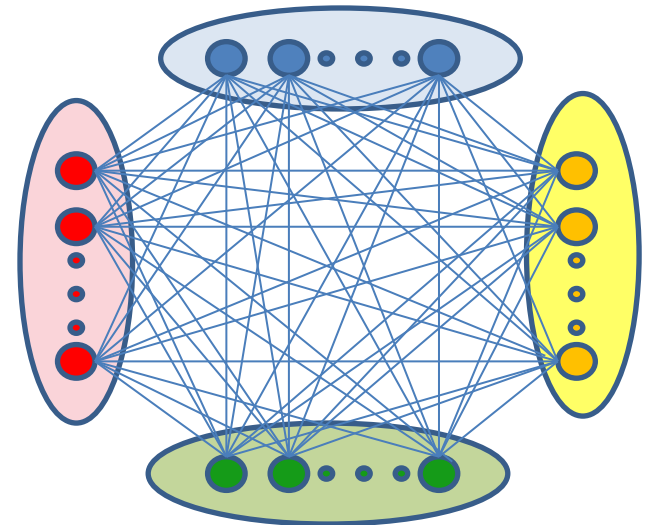
$$\alpha = (r - 1) \frac{n}{r}$$

- The file size of the FR code C_T :

$$M_{C_T}(k) = k\alpha - \binom{k}{2} + r \binom{b}{2} + bt$$

where $b = \left\lfloor \frac{k}{r} \right\rfloor$, $t \equiv k \pmod{r}$

MBR



Attains the upper bound on the file size for all $1 \leq k \leq \alpha$

Optimal FR codes with $\rho = 2$: FR code based on a graph with large girth

Lemma 2

The file size $M(k)$ of C for any $1 \leq k \leq \alpha$ satisfies

$$M(k) \leq k\alpha - (k - 1)$$

$$\text{Lemma 1: } M(k) = k\alpha - \max_{G'=(V',E') \in G_k} |E'|$$

Optimal FR codes with $\rho = 2$: FR code based on a graph with large girth

Lemma 2

The file size $M(k)$ of C for any $1 \leq k \leq \alpha$ satisfies

$$M(k) \leq k\alpha - (k - 1)$$

$$\text{Lemma 1: } M(k) = k\alpha - \max_{G'=(V',E') \in G_k} |E'|$$

- **Girth** g of a graph G is the length of the shortest cycle
- The file size of C_G is $M(k) = k\alpha - (k - 1)$ iff the girth of G is at least $k + 1$.

Optimal FR codes with $\rho = 2$:

FR code based on a graph with large girth

Lemma 2

The file size $M(k)$ of C for any $1 \leq k \leq \alpha$ satisfies

$$M(k) \leq k\alpha - (k - 1)$$

$$\text{Lemma 1: } M(k) = k\alpha - \max_{G'=(V',E') \in G_k} |E'|$$

- **Girth** g of a graph G is the length of the shortest cycle
- The file size of C_G is $M(k) = k\alpha - (k - 1)$ iff the girth of G is at least $k + 1$.
- The FR code C_G based on an α -regular graph G with girth g is optimal for each $k \leq g - 1$.

FR codes with $\rho = 2$

- To obtain a specific value of $M(k)$,

$$k\alpha - \binom{k}{2} \leq M(k) \leq k\alpha - k + 1$$

we need to exclude certain subgraphs from G :

$$M(k) \geq k\alpha - \binom{k}{2} + 1 \text{ if } G \text{ does not contain a } k\text{-clique } K_k$$

$$M(k) \geq k\alpha - \binom{k}{2} + 2 \text{ if } G \text{ does not contain a } K_k - e$$

FR codes with $\rho = 2$

- To obtain a specific value of $M(k)$,

$$k\alpha - \binom{k}{2} \leq M(k) \leq k\alpha - k + 1$$

we need to exclude certain subgraphs from G :

$$M(k) \geq k\alpha - \binom{k}{2} + 1 \text{ if } G \text{ does not contain a } k\text{-clique } K_k$$

$$M(k) \geq k\alpha - \binom{k}{2} + 2 \text{ if } G \text{ does not contain a } K_k - e$$

- Turán type problem: find the minimum number of vertices in a graph which does not contain a specific subgraph.
 - **Turán graphs** are the graphs which do not contain a clique and have minimum number of vertices
 - **Cages** are the graphs with the given degree and girth have minimum number of vertices

Bound on reconstruction degree k

- Given M, θ, n , and α , find the smallest reconstruction degree k to provide the maximum failure resilience.
- [Lemma](#). Let C be an (n, α, ρ) FR code which stores a file of a given size M . The reconstruction degree k should satisfy

$$k \geq \left\lceil \frac{n \binom{M-1}{\alpha}}{\binom{\theta}{\alpha}} \right\rceil + 1$$

Bound on reconstruction degree k

- Given M, θ, n , and α , find the smallest reconstruction degree k to provide the maximum failure resilience.
- Lemma. Let C be an (n, α, ρ) FR code which stores a file of a given size M . The reconstruction degree k should satisfy

$$k \geq \left\lceil \frac{n \binom{M-1}{\alpha}}{\binom{\theta}{\alpha}} \right\rceil + 1$$

- FR code based on K_n attains this bound for $k = n - 2$
- FR code based on K_n without one perfect matching attains this bound for $k = n - 2$

Optimal FR codes with $\rho > 2$
Transversal Designs based codes

Optimal FR codes with $\rho > 2$

Transversal Designs based codes

- FR codes based on Transversal Designs
 - Nodes = points of the design
 - MDS symbols = blocks of the design

Optimal FR codes with $\rho > 2$

Transversal Designs based codes

- FR codes based on Transversal Designs
 - Nodes = points of the design
 - MDS symbols = blocks of the design
- (n, α, ρ) FR code:
 - There are n points in the design
 - Each block contains ρ points
 - Each point is contained in α blocks

Optimal FR codes with $\rho > 2$ Transversal Designs based codes

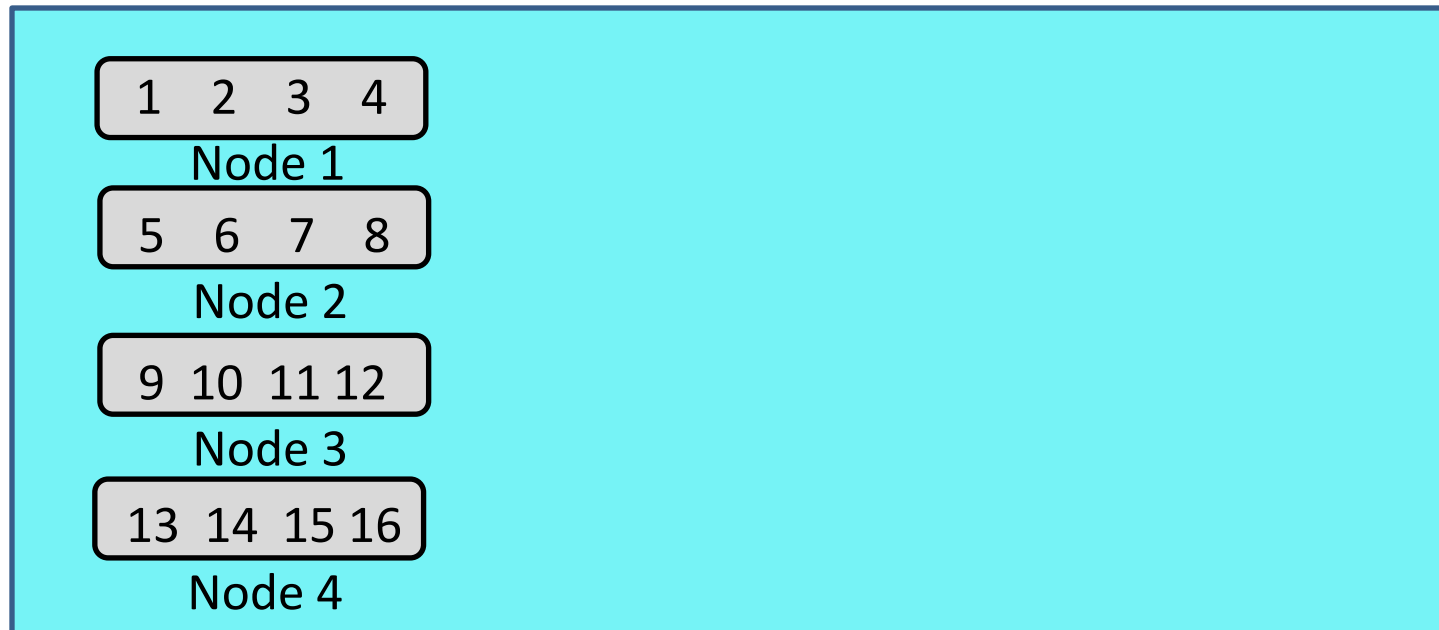
- FR codes based on Transversal Designs
 - Nodes = points of the design
 - MDS symbols = blocks of the design
- The file size of the FR code C_{TD} based on TD is strictly larger than for MBR code
- For the parameter α large enough it is an optimal FR code

Allowing parallel reads using our FR codes

- **Parallel** reads of **a subset** of **any** data symbols:
 - Multiple concurrent data collectors
 - Reading **at most one** element from each node

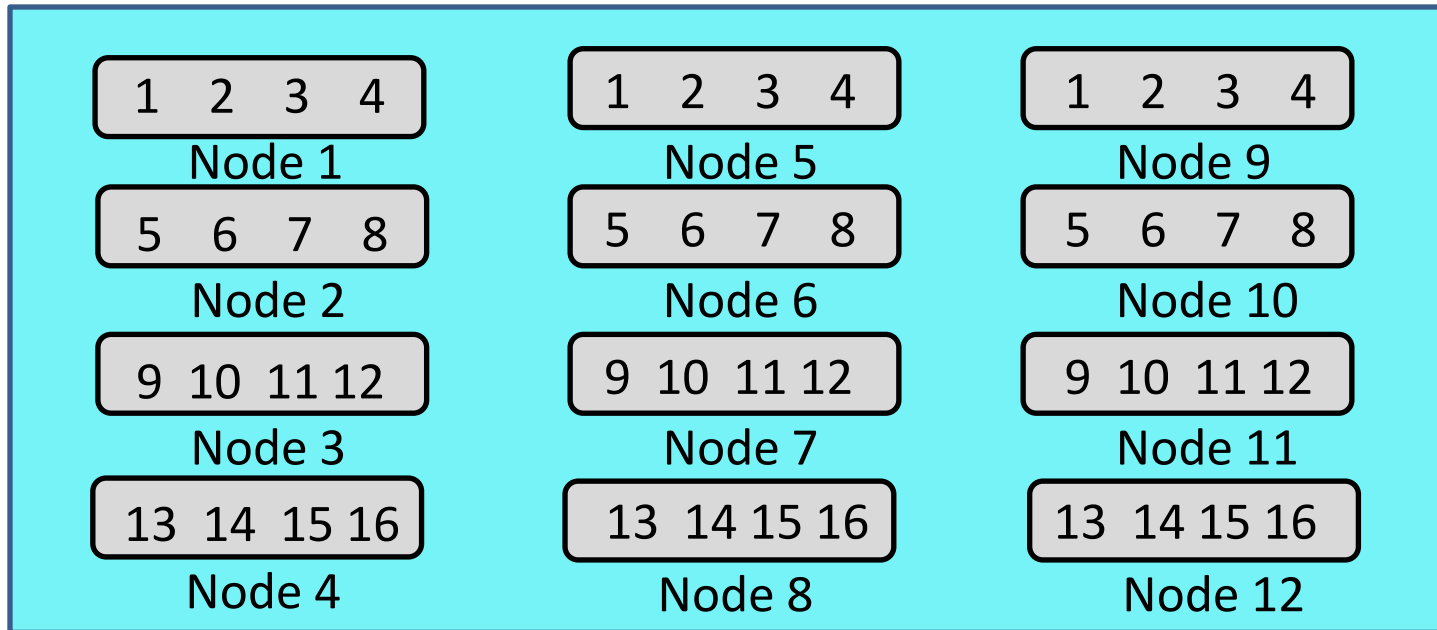
Allowing parallel reads using our FR codes

- **Parallel** reads of **a subset** of **any** data symbols:
 - Multiple concurrent data collectors
 - Reading **at most one** element from each node



Allowing parallel reads using our FR codes

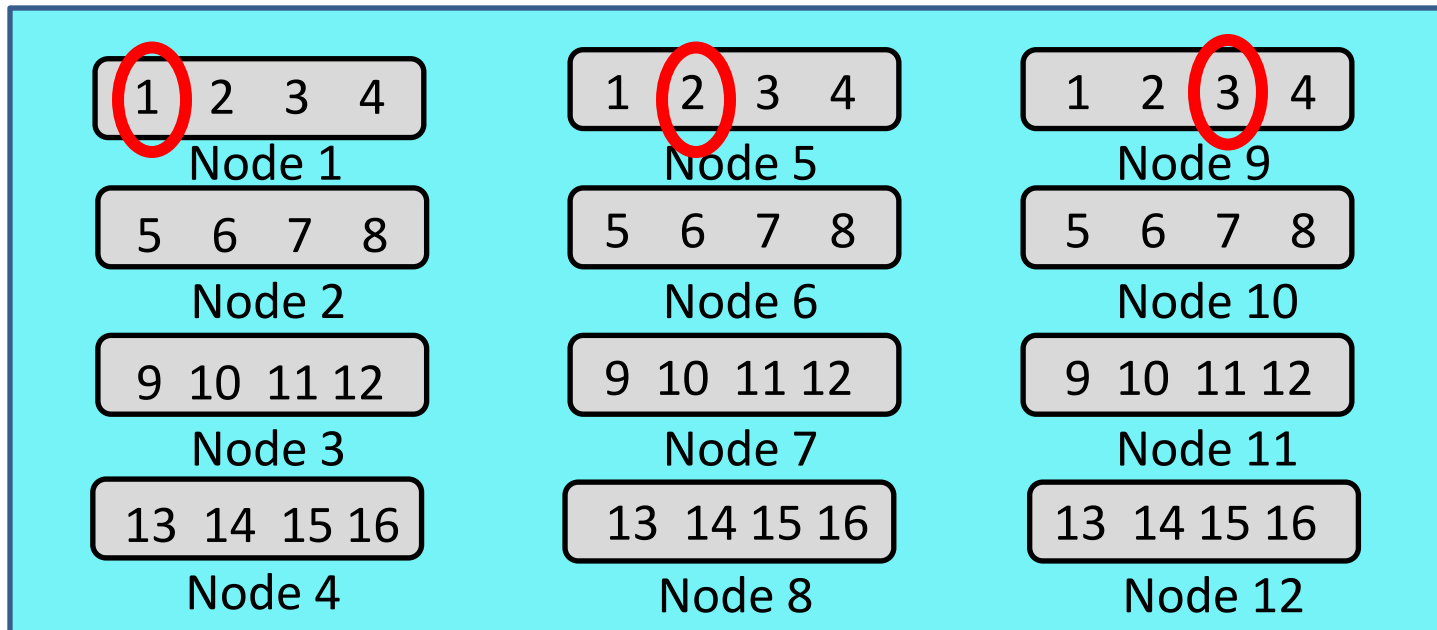
- **Parallel** reads of **a subset** of **any** data symbols:
 - Multiple concurrent data collectors
 - Reading **at most one** element from each node



{1,2,3,4} can not be read in parallel!

Allowing parallel reads using our FR codes

- **Parallel** reads of **a subset** of **any** data symbols:
 - Multiple concurrent data collectors
 - Reading **at most one** element from each node



{1,2,3,4} can not be read in parallel!

Allowing parallel reads using our FR codes

- Parallel reads of a subset of any data symbols:
 - Multiple concurrent data collectors
 - Reading **at most one** element from each node
- **Batch codes**^(*) are designed to have this property

(*) Y. Ishai, E. Kushilevitz, R. Ostrovsky, and A. Sahai, “Batch codes and their applications”, 2004

Allowing parallel reads using our FR codes

- Parallel reads of a subset of any data symbols:
 - Multiple concurrent data collectors
 - Reading **at most one** element from each node
- **Batch codes** are designed to have this property
- FR codes based on **transversal designs** are good batch codes^(*):
 - FR code based on **TD($q - 1, q$)**, for a prime power q , allows **$n - 1 = q^2 - q - 1$** parallel reads

Allowing parallel reads using our FR codes

- Parallel reads of a subset of any data symbols:
 - Multiple concurrent data collectors
 - Reading **at most one** element from each node
- **Batch codes** are designed to have this property
- FR codes based on **transversal designs** are good batch codes:
 - FR code based on $TD(q - 1, q)$, for a prime power q , allows $n - 1 = q^2 - q - 1$ parallel reads

Define **FR Batch** codes that have the properties
of **FR** and **batch** codes **simultaneously**

FR batch code based on TD(3,4)

$$\alpha = 4, \rho = 3, n = 12, k = 4, M = 11$$

f

1
2
3
4
5
6
7
8
9
10
11

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16



1 2 3 4
Node 1

1 5 9 13
Node 5

1 6 12 15
Node 9

5 6 7 8
Node 2

2 6 10 14
Node 6

2 5 11 16
Node 10

9 10 11 12
Node 3

3 7 11 15
Node 7

3 8 10 13
Node 11

13 14 15 16
Node 4

4 8 12 16
Node 8

4 7 9 14
Node 12

Any 11 symbols can be read in parallel!

Conclusion

- New constructions of optimal FR codes based on
 - Turàn regular graphs
 - Graphs with large girth
 - Transversal designs
- FR batch codes
 - Multiple parallel reads

Thank you!

Optimal FR codes with $\rho > 2$

Transversal Designs based codes

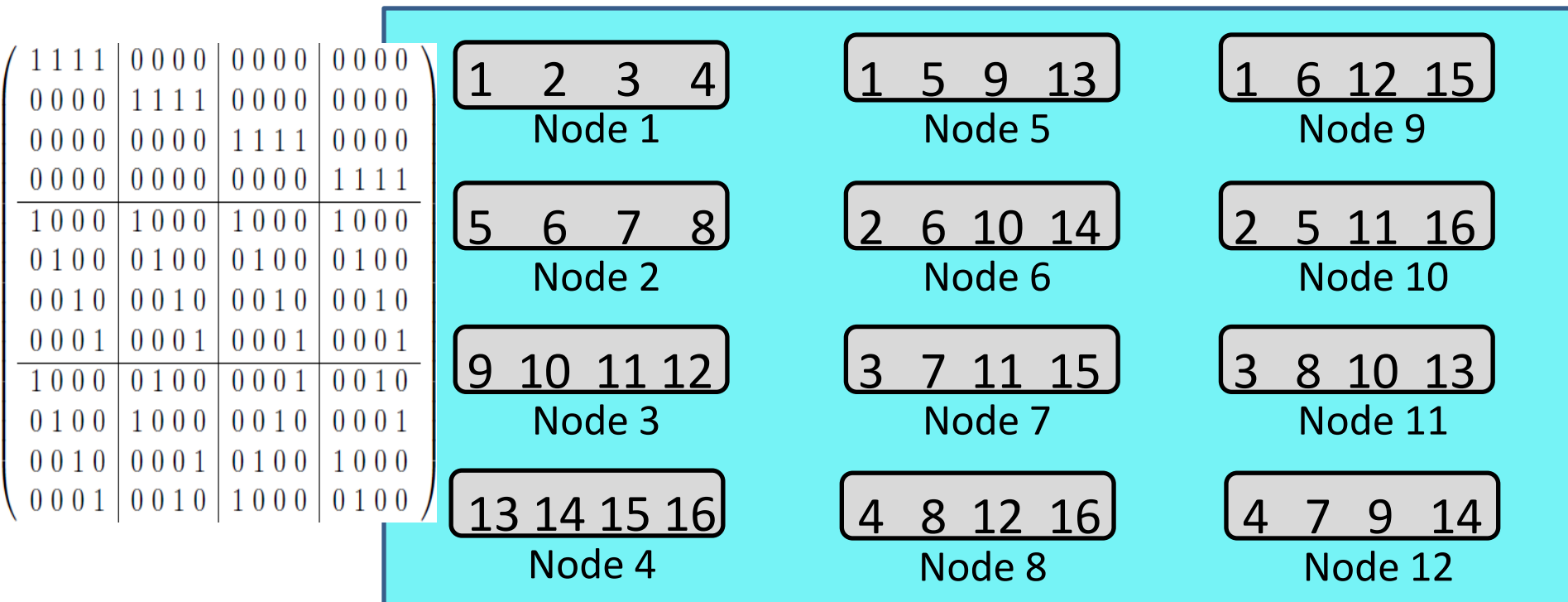
A transversal design $TD(\rho, \alpha)$ of block size ρ and group size α is a triple $(\mathcal{P}, \mathcal{G}, \mathcal{B})$ where

- \mathcal{P} is a set of $\rho\alpha$ points;
 - \mathcal{G} is a partition of \mathcal{P} into ρ sets (groups) of size α each;
 - \mathcal{B} is a collection of ρ -subsets of \mathcal{P} (blocks);
 - each block meets each group in exactly one point;
 - any pair of points from different groups is contained in exactly one block.
-
- FR codes based on Transversal Designs
 - Nodes = points of the design ($\rho\alpha$)
 - MDS symbols = blocks of the design (α^2)

Optimal FR codes with $\rho > 2$

Transversal Designs based codes

- TD(3,4): $n=12$ points, $\theta=16$ blocks of size $\rho=3$,
each point in $\alpha=4$ blocks

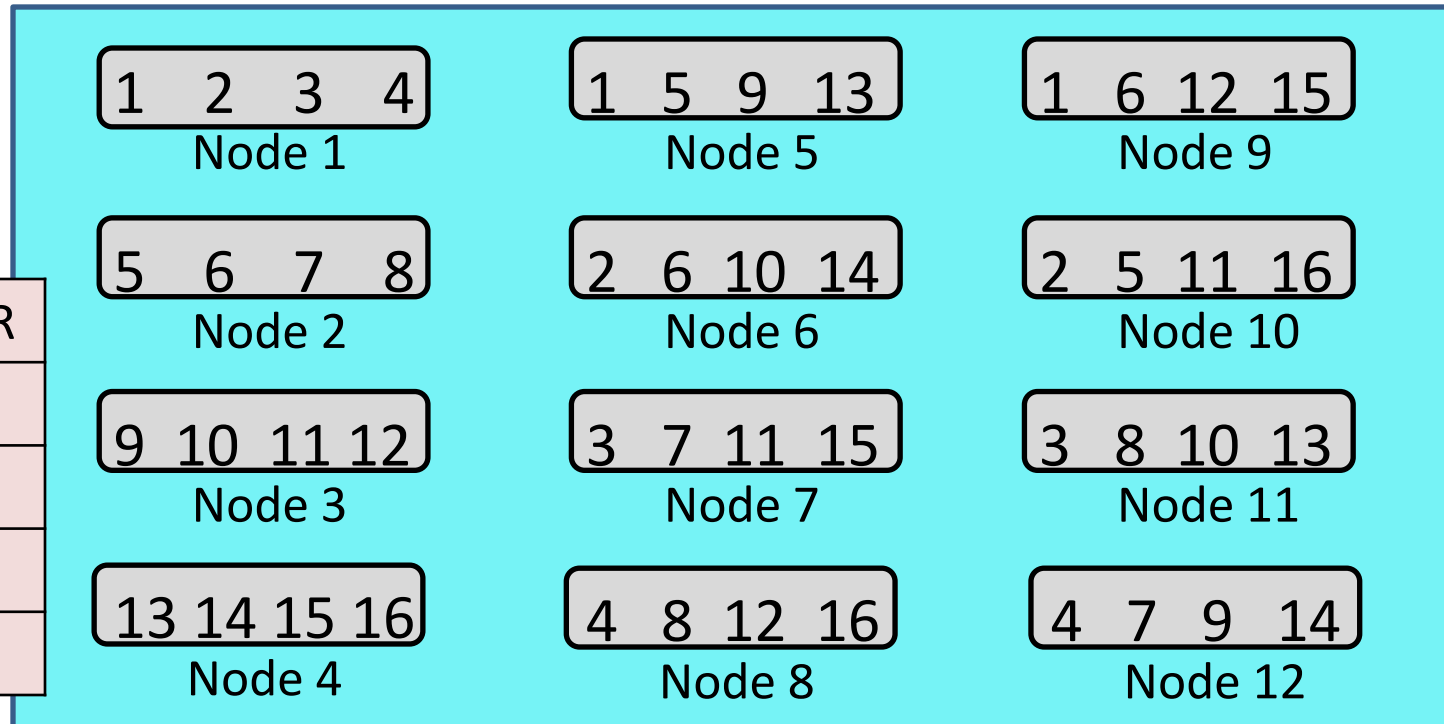


Optimal FR codes with $\rho > 2$

Transversal Designs based codes

- TD(3,4): $n=12$ points, $\theta=16$ blocks of size $\rho=3$,
each point in $\alpha=4$ blocks

k	R_C	MBR
1	4	4
2	7	7
3	9	9
4	11	10



Optimal FR codes with $\rho > 2$

Transversal Designs based codes

- Theorem: Let C_{TD} be FR code based on $TD(\rho, \alpha)$

$$R_{C_{TD}}(k) \geq k\alpha - \binom{k}{2} + \rho \binom{b}{2} + bt$$

where $k = b\rho + t, t \leq r - 1$

Proof: similar to Turán graphs

Optimal FR codes with $\rho > 2$

Transversal Designs based codes

- Theorem: Let C_{TD} be FR code based on $TD(\rho, \alpha)$

$$R_{C_{TD}}(k) \geq k\alpha - \binom{k}{2} + \rho \binom{b}{2} + bt$$

where $k = b\rho + t, t \leq r - 1$

Proof: similar to Turán graphs

For $\alpha \geq \alpha_0(k, \rho)$:

- $R_{C_{TD}}(k) = k\alpha - \binom{k}{2} + \rho \binom{b}{2} + bt$
- Attains the upper bound on FR capacity