

An Explicit Construction of Systematic MDS Codes with Small Sub-packetization for All-Node Repair

Katina Kralevska and Danilo Gligoroski

Dep. of Information Security and Communication Technologies, NTNU, Norwegian University of Science and Technology

Email: {katinak, danilog}@ntnu.no

Abstract—An explicit construction of systematic MDS codes, called HashTag+ codes, with arbitrary sub-packetization level for all-node repair is proposed. It is shown that even for small sub-packetization levels, HashTag+ codes achieve the optimal MSR point for repair of any parity node, while the repair bandwidth for a single systematic node depends on the sub-packetization level. Compared to other codes in the literature, HashTag+ codes provide from 20% to 40% savings in the average amount of data accessed and transferred during repair.

Index Terms: Explicit, systematic, MDS, MSR, small sub-packetization, all-node repair, access-optimal.

I. INTRODUCTION

Redundancy is essential to ensure reliability in distributed storage systems. Maximum Distance Separable (MDS) codes are optimal erasure codes in terms of the redundancy-reliability tradeoff. In particular, a (n, k) MDS code tolerates the maximum number of failures, up to $r = n - k$ failed nodes, for the added redundancy of r nodes. A *systematic* (n, k) MDS code is applied in such a way that the original data is equally divided into k parts without encoding and stored into k nodes, called systematic nodes, and r linear combinations of the k parts are stored into r nodes, called parity nodes. In addition to their redundancy-reliability optimality, systematic MDS codes are preferred in practical systems because data access from the systematic nodes can be done instantly without decoding.

Conventional MDS codes do not perform well in terms of the *repair bandwidth* defined as the amount of data that is transferred during a node repair. Dimakis et al. [1] proved that the lower bound of the repair bandwidth γ for a single node with a (n, k) MDS code is:

$$\gamma_{MSR}^{min} \geq \frac{M}{k} \frac{n-1}{n-k}, \quad (1)$$

where M is the file size. The equality is met when a fraction of $1/r$ -th of the stored data is transferred from all $n-1$ non-failed nodes. Minimum Storage Regenerating (MSR) codes satisfy the equality and they operate at the MSR point.

The exponential sub-packetization level is a fundamental limitation of any high-rate MSR code. The sub-packetization levels are $\alpha = r^{k/r}$ and $\alpha = r^{n/r}$ for optimal repair of systematic nodes and optimal repair of both systematic and parity nodes (all-node repair) [2], respectively. Large sub-packetization levels bring multiple practical challenges such as high I/O, high repair time, expensive computations, and

difficult management of meta-data. Thus constructing high-rate MDS codes with small sub-packetization levels has attracted a lot of attention in the recent years. Table I summarizes several high-rate MDS codes with small sub-packetization [3]–[7]. Three piggyback designs were presented in [4]. For the purpose of this paper, we compare with piggyback design 2 that optimizes all-node repair for $r \geq 3$ and sub-packetization of $(2r-3)m$ where $m \geq 1$. HashTag codes [4], [5] repair the systematic nodes with the lowest repair bandwidth in the literature for arbitrary sub-packetization $2 \leq \alpha \leq r^{\lceil k/r \rceil}$. Rawat et al. presented two approaches for all-node repair in [6]. The second approach, that is more flexible in terms of the sub-packetization, requires MSR codes and error correcting codes with specific parameters to obtain ϵ -MSR codes. However, codes with such specific parameters may not always be available. Additionally, there is a tradeoff between ϵ and the length of the code. Clay codes were recently presented in [7]. They are optimized for all-node repair. However, Clay codes require an exponential sub-packetization level, and for sub-packetization levels lower than the maximal exponential value, they are just MDS codes that do not achieve the optimal MSR point neither for the data nodes nor for the parity nodes. It is observed in [8] that 98.08% of the failures in Facebook's data-warehouse cluster that consists of thousands of nodes are single failures. Thus, we optimize the repair for single failures of any systematic or parity node.

In this paper we present a family of MDS codes called HashTag+ codes with the following properties: 1. They are systematic MDS codes; 2. They are exact-repairable codes; 3. They have a high-rate; 4. They have a flexible sub-packetization ($4 \leq \alpha \leq r^{\lceil n/r \rceil}$); 5. They achieve the MSR point for repair of single parity node for sub-packetization levels lower than or equal to the maximal exponential value of $r^{\lceil n/r \rceil}$; 6. They achieve the MSR point for repair of single systematic node for $\alpha = r^{\lceil n/r \rceil}$ and repair near-optimally for $\alpha < r^{\lceil n/r \rceil}$; 7. They are access-optimal (access and transfer the same amount of data). We combine the framework proposed by Li et al. [9] and the family of MDS codes called HashTag codes [5]. Compared to the work by Li et al. [9] where they focus on MSR codes with the maximal sub-packetization level $\alpha = r^{\lceil \frac{n}{r} \rceil}$, we construct explicit codes for the whole range of sub-packetization levels $4 \leq \alpha \leq r^{\lceil \frac{n}{r} \rceil}$ motivated by the practical importance of codes with small sub-packetization levels.

The rest of the paper is organized as follows. Section

TABLE I
COMPARISON OF HashTag+ CODES WITH EXISTING MDS CODES WITH SMALL SUB-PACKETIZATION FOR $n - 1$ HELPER NODES.

Code	Systematic	Explicit construction	Number of parities r	Sub-packetization α	All-node repair	Optimal parity repair for small α
Piggyback 2 [3]	Yes	Yes	$r \geq 3$	$(2r - 3)m, m \geq 1$	Yes	No
HashTag [5]	Yes	Yes	$r \geq 2$	$2 \leq \alpha \leq r^{\lceil k/r \rceil}$	No	No
Rawat et al. [6]	Yes	Yes	$r \geq 2$	$r^\tau, \tau \geq 1$	Yes	No
Clay codes [7]	Yes	Yes	$r \geq 2$	$\alpha \leq r^{n/r}$	Yes	No
HashTag+	Yes	Yes	$r \geq 2$	$4 \leq \alpha \leq r^{\lceil n/r \rceil}$	Yes	Yes

II presents HashTag+ code construction by first giving two examples and then presenting a general algorithm and performance comparison between HashTag+ and state-of-the-art codes. Section III concludes the paper.

Notations. For two integers $0 < i < j$, we denote the set $\{i, i + 1, \dots, j\}$ by $[i : j]$, while the set $\{0, 1, \dots, j - 1\}$ is denoted by $[j]$. Vectors and matrices are denoted with a bold font.

II. HashTag+ CODE CONSTRUCTION

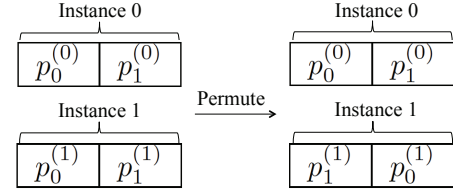
We now present two examples of HashTag+ codes with the maximal and a small sub-packetization, and we later give algorithms for general code construction and repair. An appealing feature of HashTag+ codes is that they support any values of code parameters $k, r \geq 2$, and sub-packetization $4 \leq \alpha \leq r^{\lceil \frac{n}{r} \rceil}$ including cases where r does not divide n .

Example 1: Consider a $(6, 4)$ HashTag MDS code with $\alpha = 2^{4/2} = 4$ as a base code. The code given in Fig. 1 is generated with Alg. 1 from [5] where the coefficients are from the finite field \mathbb{F}_{16} with irreducible polynomial $x^4 + x^3 + 1$. This code achieves the bound in Eq. (1) for repair of any single systematic node, i.e., 10 symbols are read and transferred for repair of 4 symbols of any systematic node. The repair of a single parity node is the same as Reed-Solomon codes, i.e., 16 symbols are read and transferred for repair of 4 parity symbols. The goal is to construct a code that provides optimal repair of the parity nodes as well (all-node repair).

In a first step, we generate $r - 1 = 1$ additional instances of the $(6, 4)$ HashTag MDS code with $\alpha = 4$ by using Alg. 1 from [5]. The data of the first systematic node d_0 stored in instance 0 is $a_{0,0}, \dots, a_{3,0}$ and in instance 2 is $a_{4,0}, \dots, a_{7,0}$ as it is shown in Fig. 2. In this way, we obtain a $(6, 4)$ code with sub-packetization level of $r \times \alpha = 2 \times 4 = 8$. A systematic node $d_j, j = 0, \dots, 3$, comprises the symbols $a_{i,j}$ from the two instances where $i = 0, \dots, 7$ and $j = 0, \dots, 3$, and a parity node $p_l, l = 0, 1$, comprises the symbols $p_{i,l}$ from the two instances where $i = 0, \dots, 7$ and $l = 0, 1$. Note that the base code from above has been renamed to instance 0.

In the second step, we permute the data in the two instances of the parity nodes p_0 and p_1 . First, this data is represented as $p_l^{(i)}$ where the superscript i denotes the instance and the subscript l denotes the parity node. Then, the permutation is

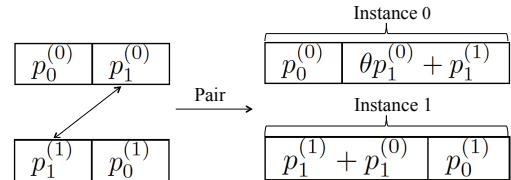
as follows: $p_l^{(i)} \rightarrow p_{l+i}^{(i)}$ where the index arithmetic is cyclic, i.e., modulo r (for example $l + i = 3 \rightarrow l + i = 1$).



In the third step, the data from the parity nodes is paired following this rule:

$$p_l^{(i)} = \begin{cases} p_l^{(i)}, & \text{if } i = l, \\ \theta_{l,i} p_l^{(i)} + p_i^{(l)}, & \text{otherwise} \end{cases} \quad (2)$$

where $\{\theta_{l,i}, \theta_{i,l}\} \subseteq \{1, \theta\}$ and $\theta \in \mathbb{F}_{16} \setminus \{0, 1\}$. The bidirectional arrows in the figure below shows which parity parts are paired together. This completes the code generation.



The final $(6, 4)$ HashTag+ code with $\alpha = 8$ that provides optimal all-node repair is given in Fig. 1.

We now illustrate that this code recovers optimally any systematic or parity node. Let us assume that node d_0 has failed. In order to recover $a_{0,0}, a_{1,0}$, we transfer 6 symbols $a_{0,1}, a_{1,1}, a_{0,2}, a_{1,2}, a_{0,3}, a_{1,3}$ from instance 0 of the non-failed systematic nodes and 2 non-paired symbols $p_{0,0}, p_{1,0}$ from the parity nodes. Next we recover $a_{4,0}, a_{5,0}$ by downloading 6 symbols $a_{4,1}, a_{5,1}, a_{4,2}, a_{5,2}, a_{4,3}, a_{5,3}$ from instance 1 of the systematic nodes and 2 non-paired symbols $p_{4,0}, p_{5,0}$ from the parity nodes. To recover the remaining symbols $a_{2,0}, a_{3,0}$ from instance 0, we transfer the paired symbols $\theta p_{0,1} + p_{4,1}, p_{4,1} + p_{0,1}$ and solve 2×2 system of linear equations. In a similar manner we recover the last two symbols $a_{6,0}, a_{7,0}$ by transferring the paired symbols $p_{5,1} + p_{1,1}, \theta p_{1,1} + p_{5,1}$. Thus, the repair of d_0 (or any other systematic node) requires 20 symbols in total, and it achieves the bound in Eq.(1).

Systematic nodes				Parity nodes			
$a_{0,0}$	$a_{0,1}$	$a_{0,2}$	$a_{0,3}$	$p_{0,0} = 6a_{0,0} + 13a_{0,1} + 15a_{0,2} + 7a_{0,3}$	$p_{0,1} = 8a_{0,0} + 12a_{0,1} + 8a_{0,2} + 4a_{0,3} + 12a_{2,0} + 8a_{1,2}$		
$a_{1,0}$	$a_{1,1}$	$a_{1,2}$	$a_{1,3}$	$p_{1,0} = 2a_{1,0} + 8a_{1,1} + 14a_{1,2} + 6a_{1,3}$	$p_{1,1} = 7a_{1,0} + 7a_{1,1} + 4a_{1,2} + 8a_{1,3} + 4a_{3,0} + 5a_{0,3}$		
$a_{2,0}$	$a_{2,1}$	$a_{2,2}$	$a_{2,3}$	$p_{2,0} = 14a_{2,0} + 12a_{2,1} + 7a_{2,2} + 14a_{2,3}$	$p_{2,1} = 2a_{2,0} + 7a_{2,1} + 15a_{2,2} + 5a_{2,3} + 6a_{0,1} + 15a_{3,2}$		
$a_{3,0}$	$a_{3,1}$	$a_{3,2}$	$a_{3,3}$	$p_{3,0} = 8a_{3,0} + 11a_{3,1} + 11a_{3,2} + 6a_{3,3}$	$p_{3,1} = 13a_{3,0} + 6a_{3,1} + 2a_{3,2} + 5a_{3,3} + 15a_{1,1} + 7a_{2,3}$		

Fig. 1. A systematic (6, 4) HashTag MDS code with $\alpha = 4$.

Instance 0											
d_0				p_0				p_1			
$a_{0,0}$	$a_{0,1}$	$a_{0,2}$	$a_{0,3}$	$p_{0,0} = 6a_{0,0} + 13a_{0,1} + 15a_{0,2} + 7a_{0,3}$				$p_{0,1} = 8a_{0,0} + 12a_{0,1} + 8a_{0,2} + 4a_{0,3} + 12a_{2,0} + 8a_{1,2}$			
$a_{1,0}$	$a_{1,1}$	$a_{1,2}$	$a_{1,3}$	$p_{1,0} = 2a_{1,0} + 8a_{1,1} + 14a_{1,2} + 6a_{1,3}$				$p_{1,1} = 7a_{1,0} + 7a_{1,1} + 4a_{1,2} + 8a_{1,3} + 4a_{3,0} + 5a_{0,3}$			
$a_{2,0}$	$a_{2,1}$	$a_{2,2}$	$a_{2,3}$	$p_{2,0} = 14a_{2,0} + 12a_{2,1} + 7a_{2,2} + 14a_{2,3}$				$p_{2,1} = 2a_{2,0} + 7a_{2,1} + 15a_{2,2} + 5a_{2,3} + 6a_{0,1} + 15a_{3,2}$			
$a_{3,0}$	$a_{3,1}$	$a_{3,2}$	$a_{3,3}$	$p_{3,0} = 8a_{3,0} + 11a_{3,1} + 11a_{3,2} + 6a_{3,3}$				$p_{3,1} = 13a_{3,0} + 6a_{3,1} + 2a_{3,2} + 5a_{3,3} + 15a_{1,1} + 7a_{2,3}$			
Instance 1											
$a_{4,0}$	$a_{4,1}$	$a_{4,2}$	$a_{4,3}$	$p_{4,0} = 6a_{4,0} + 13a_{4,1} + 15a_{4,2} + 7a_{4,3}$				$p_{4,1} = 8a_{4,0} + 12a_{4,1} + 8a_{4,2} + 4a_{4,3} + 12a_{6,0} + 8a_{5,2}$			
$a_{5,0}$	$a_{5,1}$	$a_{5,2}$	$a_{5,3}$	$p_{5,0} = 2a_{5,0} + 8a_{5,1} + 14a_{5,2} + 6a_{5,3}$				$p_{5,1} = 7a_{5,0} + 7a_{5,1} + 4a_{5,2} + 8a_{5,3} + 4a_{7,0} + 5a_{4,3}$			
$a_{6,0}$	$a_{6,1}$	$a_{6,2}$	$a_{6,3}$	$p_{6,0} = 14a_{6,0} + 12a_{6,1} + 7a_{6,2} + 14a_{6,3}$				$p_{6,1} = 2a_{6,0} + 7a_{6,1} + 15a_{6,2} + 5a_{6,3} + 6a_{4,1} + 15a_{7,2}$			
$a_{7,0}$	$a_{7,1}$	$a_{7,2}$	$a_{7,3}$	$p_{7,0} = 8a_{7,0} + 11a_{7,1} + 11a_{7,2} + 6a_{7,3}$				$p_{7,1} = 13a_{7,0} + 6a_{7,1} + 2a_{7,2} + 5a_{7,3} + 15a_{5,1} + 7a_{6,3}$			

Fig. 2. Two instances of a (6, 4) HashTag code with $\alpha = 8$.

The same amount of data is transferred when repairing the parity nodes p_0 or p_1 . We first repair the unpaired symbols $p_{0,0}, p_{1,0}, p_{2,0}, p_{3,0}$ from instance 0 by transferring all 16 symbols from instance 0 of the systematic nodes $a_{0,0}, a_{1,0}, a_{2,0}, a_{3,0}, a_{0,1}, a_{1,1}, a_{2,1}, a_{3,1}, a_{0,2}, a_{1,2}, a_{2,2}, a_{3,2}, a_{0,3}, a_{1,3}, a_{2,3}, a_{3,3}$. Next the paired symbols from p_0 are recovered by downloading the 4 symbols from instance 0 of p_1 . In total, 20 symbols are read and transferred for repair of 8 symbols from p_0 .

Example 2: We next give a (6, 4) HashTag+ code with $\alpha = 4$ in Fig. 4. The code is obtained by following the steps from the previous example where the base code is a (6, 4) HashTag code with $\alpha = 2$. Note that the sub-packetization level in this example is lower than the optimal one in Example 1. The goal is to illustrate that the code achieves the MSR point when repairing a single parity node although the sub-packetization is small.

Repairing any systematic node is near-optimal, i.e., 12 symbols for repair of 4 symbols. Let us assume that node d_0 has failed. In order to recover $a_{0,0}$, we transfer 3 symbols $a_{0,1}, a_{0,2}, a_{0,3}$ from instance 0 of the non-failed systematic nodes and 1 non-paired symbol $p_{0,0}$ from the parity node p_0 . Next we recover $a_{2,0}$ by downloading 3 symbols $a_{2,1}, a_{2,2}, a_{2,3}$ from instance 1 of the systematic nodes and 1 non-paired symbol $p_{2,0}$ from the parity node p_1 . To recover the remaining symbols $a_{1,0}, a_{3,0}$ from instance 0 and 1, we transfer the paired symbols $\theta p_{0,1} + p_{2,1}, p_{2,1} + p_{0,1}$ and $a_{1,2}, a_{3,2}$ (due to the small sub-packetization level) and solve 2×2 system of linear equations. Thus, the repair of d_0 (or any other systematic node) requires 12 symbols in total, and the repair bandwidth of the systematic nodes is the same as that of the base code (HashTag code).

However, the repair bandwidth for any parity node achieves

the lower bound in Eq.(1). In particular, all 4 symbols from p_0 are repaired by transferring all 8 symbols from instance 0 of the systematic nodes $a_{0,0}, a_{1,0}, a_{0,1}, a_{1,1}, a_{0,2}, a_{1,2}, a_{0,3}, a_{1,3}$ and 2 symbols $\theta p_{0,1} + p_{2,1}$ and $\theta p_{1,1} + p_{3,1}$ from instance 0 of p_1 . In total, 10 symbols are read and transferred for repair of 4 symbols from p_0 . Repair of p_1 requires the same amount of repair bandwidth.

A. General Code Construction

Consider a file of size $M = k\alpha$ symbols from a finite field \mathbb{F}_q stored in k systematic nodes d_j of capacity α symbols. We start the construction with a HashTag code [4], [5] as a base code that is defined as follows.

Definition 1: A $(n, k)_q$ HashTag linear code is a vector systematic code defined over an alphabet \mathbb{F}_q^α for some $2 \leq \alpha \leq r^{\lceil k/r \rceil}$. It encodes a vector $\mathbf{x} = (\mathbf{x}_0, \dots, \mathbf{x}_{k-1})$, where $\mathbf{x}_i = (x_{0,i}, x_{1,i}, \dots, x_{\alpha-1,i})^T \in \mathbb{F}_q^\alpha$ for $i \in [k]$, to a codeword $\mathcal{C}(\mathbf{x}) = \mathbf{c} = (\mathbf{c}_0, \mathbf{c}_1, \dots, \mathbf{c}_{n-1})$ where the systematic parts $\mathbf{c}_i = \mathbf{x}_i$ for $i \in [k]$ and the parity parts $\mathbf{c}_i = (c_{0,i}, c_{1,i}, \dots, c_{\alpha-1,i})^T$ for $i \in [k : n-1]$ are computed by the linear expressions that have a general form as follows:

$$c_{j,i} = \sum f_{\nu,j,i} x_{j_1,j_2}, \quad (3)$$

where $f_{\nu,j,i} \in \mathbb{F}_q$ and the index pair (j_1, j_2) is defined in the j -th row of the index array \mathbf{P}_{i-r-1} where $\nu \in [r]$. The r index arrays $\mathbf{P}_0, \dots, \mathbf{P}_{r-1}$ are defined as follows:

$$\mathbf{P}_0 = \begin{bmatrix} (0,0) & (0,1) & \dots & (0,k-1) \\ (1,0) & (1,1) & \dots & (1,k-1) \\ \vdots & \vdots & \ddots & \vdots \\ (\alpha-1,0) & (\alpha-1,1) & \dots & (\alpha-1,k-1) \end{bmatrix},$$

$a_{0,0}$	$a_{0,1}$	$a_{0,2}$	$a_{0,3}$	$p_{0,0} = 6a_{0,0} + 13a_{0,1} + 15a_{0,2} + 7a_{0,3}$	$\theta p_{0,1} + p_{4,1} = \theta(8a_{0,0} + 12a_{0,1} + 8a_{0,2} + 4a_{0,3} + 12a_{2,0} + 8a_{1,2}) + 8a_{4,0} + 12a_{4,1} + 8a_{4,2} + 4a_{4,3} + 12a_{6,0} + 8a_{5,2}$
$a_{1,0}$	$a_{1,1}$	$a_{1,2}$	$a_{1,3}$	$p_{1,0} = 2a_{1,0} + 8a_{1,1} + 14a_{1,2} + 6a_{1,3}$	$\theta p_{1,1} + p_{5,1} = \theta(7a_{1,0} + 7a_{1,1} + 4a_{1,2} + 8a_{1,3} + 4a_{3,0} + 5a_{3,3}) + 7a_{5,0} + 7a_{5,1} + 4a_{5,2} + 8a_{5,3} + 4a_{7,0} + 5a_{4,3}$
$a_{2,0}$	$a_{2,1}$	$a_{2,2}$	$a_{2,3}$	$p_{2,0} = 14a_{2,0} + 12a_{2,1} + 7a_{2,2} + 14a_{2,3}$	$\theta p_{2,1} + p_{6,1} = \theta(2a_{2,0} + 7a_{2,1} + 15a_{2,2} + 5a_{2,3} + 6a_{0,1} + 15a_{3,2}) + 2a_{6,0} + 7a_{6,1} + 15a_{6,2} + 5a_{6,3} + 6a_{4,1} + 15a_{7,2}$
$a_{3,0}$	$a_{3,1}$	$a_{3,2}$	$a_{3,3}$	$p_{3,0} = 8a_{3,0} + 11a_{3,1} + 11a_{3,2} + 6a_{3,3}$	$\theta p_{3,1} + p_{7,1} = \theta(13a_{3,0} + 6a_{3,1} + 2a_{3,2} + 5a_{3,3} + 15a_{1,1} + 7a_{2,3}) + 13a_{7,0} + 6a_{7,1} + 2a_{7,2} + 5a_{7,3} + 15a_{5,1} + 7a_{6,3}$
$a_{4,0}$	$a_{4,1}$	$a_{4,2}$	$a_{4,3}$	$p_{4,1} + p_{0,1} = 8a_{4,0} + 12a_{4,1} + 8a_{4,2} + 4a_{4,3} + 12a_{6,0} + 8a_{5,2} + 8a_{0,0} + 12a_{0,1} + 8a_{0,2} + 4a_{0,3} + 12a_{2,0} + 8a_{1,2}$	$p_{4,0} = 6a_{4,0} + 13a_{4,1} + 15a_{4,2} + 7a_{4,3}$
$a_{5,0}$	$a_{5,1}$	$a_{5,2}$	$a_{5,3}$	$p_{5,1} + p_{1,1} = 7a_{5,0} + 7a_{5,1} + 4a_{5,2} + 8a_{5,3} + 4a_{7,0} + 5a_{4,3} + 7a_{1,0} + 7a_{1,1} + 4a_{1,2} + 8a_{1,3} + 4a_{3,0} + 5a_{3,3}$	$p_{5,0} = 2a_{5,0} + 8a_{5,1} + 14a_{5,2} + 6a_{5,3}$
$a_{6,0}$	$a_{6,1}$	$a_{6,2}$	$a_{6,3}$	$p_{6,1} + p_{2,1} = 2a_{6,0} + 7a_{6,1} + 15a_{6,2} + 5a_{6,3} + 6a_{4,1} + 15a_{7,2} + 2a_{2,0} + 7a_{2,1} + 15a_{2,2} + 5a_{2,3} + 6a_{0,1} + 15a_{3,2}$	$p_{6,0} = 14a_{6,0} + 12a_{6,1} + 7a_{6,2} + 14a_{6,3}$
$a_{7,0}$	$a_{7,1}$	$a_{7,2}$	$a_{7,3}$	$p_{7,1} + p_{3,1} = 13a_{7,0} + 6a_{7,1} + 2a_{7,2} + 5a_{7,3} + 15a_{5,1} + 7a_{6,3} + 13a_{3,0} + 6a_{3,1} + 2a_{3,2} + 5a_{3,3} + 15a_{1,1} + 7a_{2,3}$	$p_{7,0} = 8a_{7,0} + 11a_{7,1} + 11a_{7,2} + 6a_{7,3}$

Fig. 3. Two instances of a (6, 4) HashTag+ code with $\alpha = 8$ where $\theta \in \mathbb{F}_{16} \setminus \{0, 1\}$.

$a_{0,0}$	$a_{0,1}$	$a_{0,2}$	$a_{0,3}$	$p_{0,0} = 15a_{0,0} + 12a_{0,1} + 2a_{0,2} + 15a_{0,3}$	$\theta p_{0,1} + p_{2,1} = \theta(3a_{0,0} + 7a_{0,1} + 13a_{0,2} + 3a_{0,3} + 6a_{1,0} + 14a_{1,2}) + 3a_{2,0} + 7a_{2,1} + 13a_{2,2} + 3a_{2,3} + 6a_{3,0} + 14a_{3,2}$
$a_{1,0}$	$a_{1,1}$	$a_{1,2}$	$a_{1,3}$	$p_{1,0} = 9a_{1,0} + 7a_{1,1} + 10a_{1,2} + 7a_{1,3}$	$\theta p_{1,1} + p_{3,1} = \theta(4a_{1,0} + 13a_{1,1} + 9a_{1,2} + 13a_{1,3} + 15a_{0,1} + 7a_{0,3}) + 4a_{3,0} + 13a_{3,1} + 9a_{3,2} + 13a_{3,3} + 15a_{2,1} + 7a_{2,3}$
$a_{2,0}$	$a_{2,1}$	$a_{2,2}$	$a_{2,3}$	$p_{2,1} + p_{0,1} = 3a_{2,0} + 7a_{2,1} + 13a_{2,2} + 3a_{2,3} + 6a_{3,0} + 14a_{3,2} + 3a_{0,0} + 7a_{0,1} + 13a_{0,2} + 3a_{0,3} + 6a_{1,0} + 14a_{1,2}$	$p_{2,0} = 15a_{2,0} + 12a_{2,1} + 2a_{2,2} + 15a_{2,3}$
$a_{3,0}$	$a_{3,1}$	$a_{3,2}$	$a_{3,3}$	$p_{3,1} + p_{1,1} = 4a_{3,0} + 13a_{3,1} + 9a_{3,2} + 13a_{3,3} + 15a_{2,1} + 7a_{2,3} + 4a_{1,0} + 13a_{1,1} + 9a_{1,2} + 13a_{1,3} + 15a_{0,1} + 7a_{0,3}$	$p_{3,0} = 9a_{3,0} + 7a_{3,1} + 10a_{3,2} + 7a_{3,3}$

Fig. 4. Two instances of a (6, 4) HashTag+ code with $\alpha = 4$ where $\theta \in \mathbb{F}_{16} \setminus \{0, 1\}$.

$$\mathbf{P}_i = \begin{bmatrix} (0,0) & \dots & (0,k-1) & \overbrace{(\cdot, \cdot)}^{\lceil \frac{k}{r} \rceil} & \dots & (\cdot, \cdot) \\ (1,0) & \dots & (1,k-1) & (\cdot, \cdot) & \dots & (\cdot, \cdot) \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ (\alpha-1,0) & \dots & (\alpha-1,k-1) & (\cdot, \cdot) & \dots & (\cdot, \cdot) \end{bmatrix}.$$

where the values of the indexes (\cdot, \cdot) are determined by a scheduling algorithm that guarantees the code is MDS, i.e. the entire information \mathbf{x} can be recovered from any k out of the n vectors \mathbf{c}_i . In addition, the algorithm ensures optimal or near-optimal repair by scheduling the indexes of the elements from \mathbf{x}_i into $\lceil \alpha/r \rceil$ rows in the $r-1$ index arrays \mathbf{P}_j where $j = 1, \dots, r-1$. ■

The scheduling algorithm for Def. 1 is presented in [4], [5]. Note that in the original presentation the indexing of the arrays is from 1 to r but in order to synchronize with the transformation of Li et al. [9] here we use the indexing of the arrays from 0 to $r-1$. The set of all symbols in d_j is partitioned in disjunctive subsets where at least one subset has $\lceil \alpha/r \rceil$ number of elements. The set of indexes $D = \{1, \dots, \alpha\}$, where the i -th index of $a_{i,j}$ from d_j is represented by i in D , is partitioned in r disjunctive subsets $D = \cup_{\rho=1}^r D_{\rho, d_j}$ where at least one subset has $\lceil \alpha/r \rceil$ elements. One subset D_{ρ, d_j} is assigned per disk. The indexes in D_{ρ, d_j} are the row positions where the pairs (i, j) with indexes $i \in D \setminus D_{\rho, d_j}$ are scheduled (the zero pairs are replaced with concrete (i, j) pairs). By using the code defined in Def. 1 as a base code, we next define HashTag+ code.

Definition 2: A $(n, k)_q$ HashTag+ linear code is a vector systematic code defined over an alphabet \mathbb{F}_q^α for some $4 \leq \alpha \leq r \lceil n/r \rceil$.

The algorithm for constructing a (n, k) HashTag+ code is given in Alg. 1.

The construction of HashTag+ codes given in Alg. 1 is sound and there always exists a finite field \mathbb{F}_q and a set of non-zero coefficients from the field such that the HashTag+ code is MDS due to the following Lemma:

Lemma 1: There exists a choice of non-zero coefficients $c_{l,i,j}$ where $l = 1, \dots, r$, $i = 1, \dots, \alpha$ and $j = 1, \dots, k$ from

Algorithm 1 HashTag+ code construction

Input: (n, k) HashTag code with sub-packetization α

Output: (n, k) HashTag+ code with sub-packetization $r \times \alpha$

- 1: Construct $r-1$ additional instances of a (n, k) HashTag code with sub-packetization α ;
- 2: Permute the data from the i -th instance in the l -th parity node as $p_l^{(i)} \rightarrow p_{l+i}^{(i)}$;
- 3: Compute the parity parts $p_l^{(i)}$ with the rule in Eq.(2).

\mathbb{F}_q such that the code is MDS if $q \geq \binom{n}{k} r \alpha$.

Proof: It is sufficient to combine Theorem 1 from [5] about the base HashTag codes and Theorem 2 and 3 from [9]. Namely, Theorem 1 from [5] guarantees that the size of the finite field for the base HashTag code is sufficient to be $q \geq \binom{n}{k} r \alpha$ in order to find a HashTag MDS code. Then, according to Theorem 2 and 3 from [9] the HashTag+ code has optimal repair bandwidth, has optimal rebuilding access and is a MDS code. ■

B. Repair of systematic nodes

Alg. 2 shows the repair of a systematic node where the systematic and the parity nodes are global variables. A set of $\lceil \alpha/r^2 \rceil$ symbols is accessed and transferred from all $n-1$ non-failed nodes from each instance.

Proposition 1: The repair bandwidth for a single systematic node γ_s is bounded between the following lower and upper bounds:

$$\frac{(n-1)}{\alpha} \lceil \frac{\alpha}{r} \rceil \leq \gamma_s \leq \frac{(n-1)}{\alpha} \lceil \frac{\alpha}{r} \rceil + \frac{(r-1)}{\alpha} \lceil \frac{\alpha}{r} \rceil \lceil \frac{k}{r} \rceil. \quad (4)$$

Proof: We read in total $k \lceil \frac{\alpha}{r} \rceil$ elements in the first for loop of Alg. 2. Additionally, $(r-1) \lceil \frac{\alpha}{r} \rceil$ elements are read in Step 7 of the second for loop. Assuming that we do not read more elements in Step 6, we determine the lower bound as $k \lceil \frac{\alpha}{r} \rceil + (r-1) \lceil \frac{\alpha}{r} \rceil = (n-1) \lceil \frac{\alpha}{r} \rceil$ elements, i.e., the lower bound is $\frac{(n-1)}{\alpha} \lceil \frac{\alpha}{r} \rceil$ (since every element has a size of $\frac{1}{\alpha}$). To derive the upper bound, we assume that we read all elements $a_{i,j}$ from the extra $\lceil \frac{k}{r} \rceil$ columns of the arrays $\mathbf{P}_0, \dots, \mathbf{P}_{r-1}$ in Step 6. Thus, the upper bound is $\frac{(n-1)}{\alpha} \lceil \frac{\alpha}{r} \rceil + \frac{(r-1)}{\alpha} \lceil \frac{\alpha}{r} \rceil \lceil \frac{k}{r} \rceil$. ■

Algorithm 2 Repair of systematic node d_j **Input:** j (where $j = 0, \dots, k-1$);**Output:** d_j ;**Note:** All indexes i are determined by the expression $i \in D_{\rho, d_j}$

- 1: **for** $v = 0, v < r$ **do**
- 2: Access and transfer $(k-1)\lceil \alpha/r^2 \rceil$ symbols $a_{i,j}$ from the v -th instance of all $k-1$ non-failed systematic nodes and $\lceil \alpha/r^2 \rceil$ non-paired symbols $p_{i,j}$ from the v -th instance of the parity nodes;
- 3: Repair $a_{i,j}$ from the v -th instance;
- 4: **end for**
- 5: **for** $v = 0, v < r$ **do**
- 6: Access and transfer the symbols $a_{i,l}$ from the v -th instance listed in the i -th row of the arrays $\mathbf{P}_0, \dots, \mathbf{P}_{r-1}$ that have not been read in Step 2;
- 7: Access and transfer $(r-1)\lceil \alpha/r^2 \rceil$ paired symbols $p_{i,j}$ from the v -th instance;
- 8: Repair $a_{i,j}$ by solving paired $r \times r$ linear systems of equations.
- 9: **end for**

C. Repair of parity nodes

Repair of a single parity node is given in Alg. 3.

Algorithm 3 Repair of a parity node p_l where $l = 0, \dots, r-1$ **Input:** l ;**Output:** p_l .

- 1: Access and transfer all symbols from instance l of the systematic nodes and the non-failed parity nodes;
- 2: Repair the symbols from p_l .

Without a proof (just a reference to Theorem 2 and 3 from [9]) we give the following Proposition:

Proposition 2: The repair bandwidth for a single parity node γ_p reaches the lower bound given in Eq. (1) for any sub-packetization level α including small α , i.e.,

$$\gamma_p = \frac{(n-1)}{\alpha} \lceil \frac{\alpha}{r} \rceil. \quad (5)$$

D. Performance Analysis

We compare the average amount of data read and downloaded during a repair of a single node taking into account all nodes (systematic and parity nodes). HashTag+ codes outperform both Piggyback 2 and HashTag codes for any code parameters as it is shown in Fig. 5. Compared to HashTag codes, the lower repair bandwidth comes at the cost of an increased sub-packetization of factor r . HashTag+ codes offer savings of up to 40% in the average amount of data accessed and transferred during repair compared to Piggyback 2.

III. CONCLUSIONS

We presented a general construction of a family of systematic MDS codes called HashTag+ codes that reaches the lower bound of the repair bandwidth for any single failure of all nodes when $\alpha = r^{\lceil n/r \rceil}$. HashTag+ codes have a high-rate and they have a flexible sub-packetization level ($4 \leq \alpha \leq r^{\lceil n/r \rceil}$). They also achieve the MSR point for repair of single parity node for sub-packetization levels lower than or equal to the

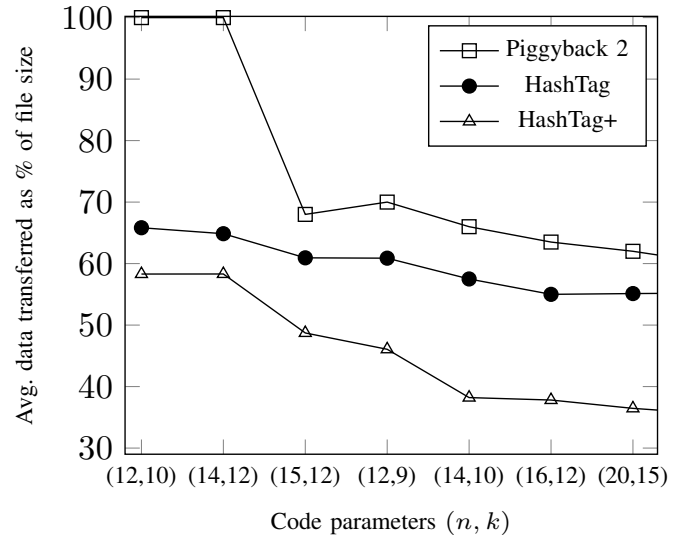


Fig. 5. Average data read and transferred for repair of any single node with Piggyback 2 [3] for $\alpha = 4 \times (2r-3)$, HashTag [5] for $\alpha = 8$, and HashTag+ for $\alpha = 8 \times r$.

maximal exponential value of $r^{\lceil n/r \rceil}$. Additionally they are access-optimal i.e. they access and transfer the same amount of data.

HashTag+ codes are the first explicit construction in the literature that repairs optimally the parity nodes even for small sub-packetization levels. The repair bandwidth for the systematic nodes is as close as possible to the lower bound when $\alpha < r^{\lceil n/r \rceil}$.

REFERENCES

- [1] A. G. Dimakis, P. B. Godfrey, Y. Wu, M. J. Wainwright, and K. Ramchandran, "Network coding for distributed storage systems," *IEEE Trans. Inf. Theory*, vol. 56, no. 9, pp. 4539–4551, Sept. 2010.
- [2] I. Tamo, Z. Wang, and J. Bruck, "Access versus bandwidth in codes for storage," *IEEE Transactions on Information Theory*, vol. 60, no. 4, pp. 2028–2037, April 2014.
- [3] K. V. Rashmi, N. B. Shah, and K. Ramchandran, "A piggybacking design framework for read-and download-efficient distributed storage codes," *IEEE Trans. on Inf. Theory*, vol. 63, no. 9, pp. 5802–5820, Sept 2017.
- [4] K. Kralevska, D. Gligoroski, and H. Øverby, "General sub-packetized access-optimal regenerating codes," *IEEE Comm. Letters*, vol. 20, no. 7, pp. 1281–1284, July 2016.
- [5] K. Kralevska, D. Gligoroski, R. E. Jensen, and H. Øverby, "Hashtag erasure codes: From theory to practice," *IEEE Trans. on Big Data*, vol. PP, no. 99, pp. 1–1, 2017.
- [6] A. S. Rawat, I. Tamo, V. Guruswami, and K. Efremenko, "MDS code constructions with small sub-packetization and near-optimal repair bandwidth," *CoRR*, vol. abs/1709.08216, 2017.
- [7] M. Vajha, V. Ramkumar, B. Puranik, G. Kini, E. Lobo, B. Sasidharan, P. V. Kumar, A. Barg, M. Ye, S. Narayanamurthy, S. Hussain, and S. Nandi, "Clay codes: Moulding MDS codes to yield an MSR code," in *16th USENIX Conf. on File and Storage Technologies (FAST 18)*, 2018, pp. 139–154.
- [8] K. Rashmi, N. B. Shah, D. Gu, H. Kuang, D. Borthakur, and K. Ramchandran, "A 'hitchhiker's' guide to fast and efficient data reconstruction in erasure-coded data centers," in *Proceedings of the 2014 ACM Conference on SIGCOMM*, 2014, pp. 331–342.
- [9] J. Li, X. Tang, and C. Tian, "A generic transformation for optimal repair bandwidth and rebuilding access in mds codes," in *IEEE Int. Symposium on Inf. Theory (ISIT)*, June 2017, pp. 1623–1627.