Centralized Repair of Multiple Node Failures

Ankit Singh Rawat[†], O. Ozan Koyluoglu[§], and Sriram Vishwanath[‡]
[†] Computer Science Dept., Carnegie Mellon University, Pittsburgh, PA 15213, USA.
[§] Dept. of ECE, The University of Arizona, Tucson, AZ 85721, USA.
[‡] Dept. of ECE, The University of Texas at Austin, Austin, TX 78712, USA.
Email: asrawat@andrew.cmu.edu, ozan@email.arizona.edu, sriram@austin.utexas.edu.

***Abstract*—This paper considers a distributed storage system, where multiple storage nodes can be reconstructed simultaneously at a centralized location. This centralized multi-node repair (CMR) model is a generalization of regenerating codes that allow for bandwidth-efficient repair of a single failed node. This work focuses on the trade-off between the amount of data stored and repair bandwidth in this CMR model. In particular, repair bandwidth bounds are derived for the minimum storage multi-node repair (MSMR) and the minimum bandwidth multi-node repair (MBMR) operating points. The tightness of these bounds are analyzed via code constructions. The MSMR point is characterized through codes achieving this point under functional repair for general set of CMR parameters, as well as with codes enabling exact repair for certain CMR parameters. The MBMR point, on the other hand, is characterized with exact repair codes for all CMR parameters for systems that satisfy a certain entropy accumulation property.**

***Index Terms*—Codes for distributed storage, regenerating codes, centralized multi-node regeneration.**

I. Introduction

The ability to preserve the stored information and maintain the seamless operation in the event of permanent failures and (or) transient unavailability of the storage nodes is one of the most important issues that need to be addressed while designing distributed storage systems. This gives rise to the so called 'code repair' or 'node repair' problem which requires a storage system to enable mechanism to regenerate (repair) the content stored on some (failed/unavailable) storage nodes with the help of the content stored on the remaining (live/available) nodes in the system. A simple replication scheme where one stores multiple copies of each data block on different nodes clearly enables the node repair as one can regenerate the data blocks stored on a node by obtaining one of their copies from the other nodes in the system. However, replication suffers from the decreasing rate as one increases the replication factor in order to enhance the resilience of the system. This motives the use of erasure codes as they efficiently trade-off the storage space for the ability to tolerate failure/unavailability of storage nodes. However, the better utilization of the storage space should also be accompanied by a resource-efficient node repair process and efficiency of the node repair becomes a yardstick for implementing one erasure code over another.

Towards this, Dimakis et al. propose repair bandwidth, the amount of data downloaded from the contacted nodes during the repair of a single node, as a measure of the efficiency of the repair process in [1]. Considering n storage nodes where any set of k nodes are sufficient to reconstruct the entire information, Dimakis et al. further characterize an information-theoretic trade-off among the storage space vs. the repair bandwidth for such codes. The codes which attain any point on this trade-off are referred to as regenerating codes. Over the past few years, the problem of designing regenerating codes has fueled numerous research efforts which have resulted into the constructions presented in [1]–[5] and the references therein.

In this paper, we explore the problem of enabling bandwidth efficient repair of multiple nodes in a centralized manner. In particular, we consider a setting where one requires the content of any k out of n nodes in the system to be sufficient to reconstruct the entire information (as a parameter for the worst case fault-tolerance of the system). As for the centralized repair process, we consider a framework where the repair of $t \geq 1$ node failures is performed by contacting any d out of the $n - t$ remaining storage nodes. We also assume that β amount of data from each of the d contacted nodes are downloaded. We aim to characterize the storage vs. repair-bandwidth trade-off under this centralized multi-node repair (CMR) framework.

We believe that this framework is more suitable for the setting of large scale storage systems where there is a need to perform repairs at a central location. Our CMR model is perhaps useful for the following scenarios: **a) Architectural and implementational issues:** Architectural constraints could make it more efficient to regenerate the content in a centralized manner. For instance, in a rack-based node placement architecture, a top-of-the-rack (TOR) switch failure would imply failure of nodes in the corresponding rack to be unaccessible, and regenerating entire content of the failed rack on a per-node basis, i.e., independently one by one, would be less efficient as compared to regenerating the content at a central location, e.g., at a leader node in that rack. **b) Threshold-based data maintenance:** These schemes regenerate servers after a threshold number of them fail. After regenerating the content stored on the failed nodes, the administrator can recruit t newcomers as replacements of the failed nodes and re-distribute the data to the newcomers in order to restore the state of the system prior to the failures. **c) Availability:** In the event of transient unavailability of the t storage nodes, the centralized repair process allows the user to get access the content stored on the unavailable nodes in a bandwidth efficient manner.

Related work: We note that the repair of multiple nodes in a bandwidth-efficient manner has previously been considered under the cooperative repair model introduced in [6], [7]. There are two major differences between the cooperative and centralized repair frameworks: a) Under cooperative repair framework [6], [7], all t newcomer nodes are not constrained to contact the same set of d out of $n-t$ surviving nodes. The framework allows each newcomer to contact any d surviving nodes independent of the nodes contacted by other $t-1$ newcomers. b) Under cooperative repair framework, after downloading data from the surviving nodes, the newcomers exchange certain amount of data among themselves. On the other hand, since a centralized entity (e.g., the administrator or a master server node) has access to all the downloaded information, such information exchange is not required in the centralized repair model. Our hope is that removing the additional restriction imposed by the cooperative repair framework will enable designing codes for a broader range of system parameters.

The problem of centralized bandwidth-efficient repair of multiple node failures in a DSS employing has previously been considered by Cadambe et al. [8]. However, they restrict themselves to only MDS codes and they show existence of such codes only in the asymptotic regime where node size (amount of data stored on a node) tends to infinity. We also note that the CMR model proposed here is also equivalent to the broadcast model [9], where repair transmissions are overheard by all the nodes under repair.

In addition, locality, the number of nodes contacted during repair of a single node, is another measure of node repair efficiency which have been extensively studied in the literature [10]. Various minimum distance bounds and constructions achieving trade-offs are presented in [10]–[13] and the references therein. In particular, recent works [14]–[16] have studied locality problem with multiple node repairs, which is a model relevant to the framework studied in this paper.

Contributions: In this paper, we develop general repair bandwidth bounds for the CMR model at minimum per-node storage and minimum repair bandwidth regimes. Then, we investigate tightness of the derived bounds with appropriate code constructions, and characterize the fundamental limits of the CMR model.

II. Centralized multi-node repair model

We introduce a new model for simultaneous repair of multiple node failures in a distributed storage system (DSS), namely *centralized multi-node repair (CMR) model.* Consider an (n,k)-DSS, *i.e.*, the system comprises n storage nodes and the content stored on any k nodes is sufficient to reconstruct the information stored on the system. For an (n,k)-DSS, under (d,t)-CMR model, any set of t failed nodes in the system can be repaired by downloading data from any set of d out of $n-t$ surviving nodes. Let α denote the size of each node (over a finite field $\mathbb{F}$) and β denote the amount of data downloaded from each of the contacted d nodes under the (d,t)-CMR model. In order to denote all the relevant system parameters, we also expand the notation for the CMR model as (n,k,d,t,α,γ)-CMR model or (d,t,α,γ)-CMR model. After downloading $\gamma = d\beta$ symbols from the contacted nodes, the content stored on all t failed nodes is recovered simultaneously in a centralized manner[1].

III. A file size bound for the CMR model

In this section, we initiate the study of the trade-off between the per-node storage α and repair bandwidth γ for the CMR model. We first provide a file size bound for the CMR model.

Let the system store a uniformly distributed file $\mathbf{f}$ of size $|\mathbf{f}| = \mathcal{M}$ (over a finite field $\mathbb{F}$). Consider the case when the nodes indexed by a set $\mathcal{K} \subseteq [n]$ such that $|\mathcal{K}| = k$ are used to reconstruct the file $\mathbf{f}$. Further, assume that this set of nodes are partitioned into g number of distinct subsets $\mathcal{S}_i$ with $|\mathcal{S}_i| = n_i \leq t$ such that $\sum\limits_{i=1}^{g} n_i = k$. We have the following bound.

Lemma 1. *The system parameters necessarily satisfy*

$$\mathcal{M} \leq \sum_{i=1}^{g} \min\Big\{n_i\alpha, \big(d - \sum_{j=1}^{i-1} n_j\big)\beta\Big\}. \tag{1}$$

Proof. Denoting the symbols stored on the nodes indexed by the set $\mathcal{S}$ by $\mathbf{x}_{\mathcal{S}}$, we have

$$\mathcal{M} = H(\mathbf{f}) \overset{(a)}{=} H(\mathbf{f}) - H(\mathbf{f}|\mathbf{x}_{\mathcal{K}}) = I(\mathbf{x}_{\mathcal{K}};\mathbf{f}) \leq H(\mathbf{x}_{\mathcal{K}}) \tag{2}$$

$$\overset{(b)}{=} \sum_{i=1}^{g} H(\mathbf{x}_{\mathcal{S}_i}|\mathbf{x}_{\mathcal{S}_1:\mathcal{S}_{i-1}}) \tag{3}$$

$$\overset{(c)}{\leq} \sum_{i=1}^{g} \min\Big\{H(\mathbf{x}_{\mathcal{S}_i}), \big(d - \sum_{j=1}^{i-1} n_j\big)\beta\Big\} \tag{4}$$

$$\overset{(d)}{\leq} \sum_{i=1}^{g} \min\Big\{n_i\alpha, \big(d - \sum_{j=1}^{i-1} n_j\big)\beta\Big\}, \tag{5}$$

where (a) is due to recoverability constraint $H(\mathbf{f}|\mathbf{x}_{\mathcal{K}}) = 0$ as $|\mathcal{K}| = k$, (b) is due to $\mathcal{K} = \cup_{i=1}^{g}\mathcal{S}_i$, (c) & (d) are due to the following bounds for each term in the sum: $H(\mathbf{x}_{\mathcal{S}_i}|\mathbf{x}_{\mathcal{S}_1:\mathcal{S}_{i-1}}) \leq H(\mathbf{x}_{\mathcal{S}_i}) \leq n_i\alpha$, and

$$\begin{aligned} H(\mathbf{x}_{\mathcal{S}_i}|\mathbf{x}_{\mathcal{S}_1:\mathcal{S}_{i-1}}) &\overset{(e)}{=} H(\mathbf{x}_{\mathcal{S}_i}|\mathbf{x}_{\mathcal{S}_1:\mathcal{S}_{i-1}}) \\ &\quad - H(\mathbf{x}_{\mathcal{S}_i}|\mathbf{x}_{\mathcal{S}_1:\mathcal{S}_{i-1}}, \mathbf{d}_{\mathcal{H}_i-\mathcal{S}_1:\mathcal{S}_{i-1}}) \\ &\leq H(\mathbf{d}_{\mathcal{H}_i-\mathcal{S}_1:\mathcal{S}_{i-1}}) \leq \Big(d - \sum_{j=1}^{i-1} n_j\Big)\beta \end{aligned}$$

where set of helper nodes to regenerate symbols in $\mathcal{S}_i$ is denoted as $\mathcal{H}_i$, this set of d nodes is constructed by using the sets $\mathcal{S}_1 \cdots \mathcal{S}_{i-1}$ and additional nodes not belonging to these sets (this is possible as $\sum_{i=1}^{g} n_i = k \leq d$), downloaded symbols from these additional nodes are denoted as $\mathbf{d}_{\mathcal{H}_i-\mathcal{S}_1:\mathcal{S}_{i-1}}$ with $|\mathcal{H}_i - \mathcal{S}_1 : \mathcal{S}_{i-1}| = d - \sum_{j}^{i-1} n_j$, and (e) follows as $H(\mathbf{x}_{\mathcal{S}_i}|\mathbf{x}_{\mathcal{S}_1:\mathcal{S}_{i-1}}, \mathbf{d}_{\mathcal{H}_i-\mathcal{S}_{1:i-1}}) = 0$ as $H(\mathbf{x}_{\mathcal{S}}|\mathbf{d}_{\mathcal{H}})$=0 for any $\mathcal{S}$ such that $|\mathcal{S}| \leq t$ and any $\mathcal{H}$ such that $|\mathcal{H}| = d$. □

[1] The CMR model also allow for the distributed/parallel repair of all the t failed nodes by t newcomers independently. However, it is assumed that each of the t newcomers have an access to all the γ downloaded symbols.

Given the bound in Proposition 1, we differentiate between two operating regimes of the system: *Minimum storage multi-node regeneration (MSMR)* and *minimum bandwidth multi-node regeneration (MBMR)*. The MSMR point corresponds to having an MDS code which requires that $\alpha = \mathcal{M}/k$. Codes that attain minimum possible repair bandwidth under this constraint, i.e., $\alpha = \mathcal{M}/k$, are referred to as MSMR codes. On the other hand, the MBMR point restricts that $H(\mathbf{x}_{\mathcal{S}}) = \gamma = d\beta$ for every $\mathcal{S} \subseteq [n]$ such that $|\mathcal{S}| = t$, i.e., the amount of data downloaded during the centralized repair of t node failures is equal to the amount of information stored on the lost t nodes. MBMR codes achieve the minimum possible repair bandwidth under this restriction, i.e., $H(\mathbf{x}_{\mathcal{S}}) = \gamma = d\beta$. In the following, we focus on the problem of characterizing these two operating points of the CMR model.

IV. MSMR Codes

We first utilize Lemma 1 to obtain a bound on the repair bandwidth at the MSMR point, and then focus on achievability.

A. Repair bandwidth bound

Proposition 1. *Consider an (n, k)-DSS that stores a file of size $\mathcal{M}$ and enables repair of t failed nodes under a $(d, t, \alpha_{MSMR} = \frac{\mathcal{M}}{k}, \gamma)$-CMR model. Then, we have*

$$\gamma_{MSMR} \geq \frac{\mathcal{M}dt}{k(d-k+t)}. \tag{6}$$

Proof. Let $a = \lfloor k/t \rfloor$ and $b = k - at$. We set $n_1 = b$ and $n_i = t$ for $i = 2, \cdots, g = a + 1$. From (1), we obtain

$$\mathcal{M} \leq \min\{b\alpha, d\beta\} + \sum_{i=1}^{a} \min\{t\alpha, [d - (i-1)t - b]\beta\}. \tag{7}$$

Note that we have $\alpha = \frac{\mathcal{M}}{k}$ which implies that $d\beta \geq b\alpha$ and

$$[d - (i-1)t - b]\beta \geq t\alpha, \forall i = 1, \cdots, a,$$

From this, we obtain $\beta \geq \frac{b\alpha}{d}$ and $[d - (a-1)t - b]\beta \geq t\alpha$, i.e., $\beta \geq \frac{t\alpha}{[d-at-b+t]} = \frac{t\alpha}{[d-k+t]}$. This along with the fact that we have $b < t \leq k$ establish (6). □

Remark 1. Note that the same bound is also obtained by Cadambe et al. in [8] where the authors consider repair of multiple failures in an MDS code.

Remark 2. A code that allows for repair of t failed nodes with the parameters $\left(d, t, \alpha = \frac{\mathcal{M}}{k}, \gamma = \frac{\mathcal{M}dt}{k(d-k+t)}\right)$-CMR is an MSMR code.

B. Constructions and the characterization of the MSMR point

1) Constructions from existing MSCR codes: Minimum storage cooperative regenerating (MSCR) codes allow for simultaneous repair of t storage nodes with the following scheme: Each newcomer node contacts to d nodes and downloads β symbols from each. (Different nodes can contact to different live nodes.) Then, each newcomer node sends β' symbols to each other. Under this setup, the repair bandwidth *per failed node* is $d\beta + (t-1)\beta'$. MSCR codes operate at $\alpha_{MSCR} = \mathcal{M}/k$ and $\beta_{MSCR} = \beta'_{MSCR} = \frac{\mathcal{M}}{k(d-k+t)}$.

Proposition 2. *A code $\mathcal{C}$ that operates as an MSCR code is also an MSMR code for the CMR model.*

Proof. Consider that each failed node contact to the same set of d nodes in the MSCR code $\mathcal{C}$. Then, each failed node downloads β_{MSCR} symbols from these d helper nodes, resulting in a total of at most $\gamma = td\beta_{MSCR} = \frac{\mathcal{M}dt}{k(d-k+t)}$ symbols. These symbols can recover each failed node, hence regenerates t failed nodes in the CMR model. Therefore, code $\mathcal{C}$ is an MSMR code with $\alpha = \frac{\mathcal{M}}{k}$ and $\gamma = \frac{\mathcal{M}dt}{k(d-k+t)}$. □

We remark that random linear network coding attains MSCR point [6], hence it provides an MSMR code with functional repair. Explicit code constructions for the MSCR setup while ensuring exact-repair, on the other hand, are known for a small set of parameters. The only such constructions that we are aware of are provided in [17] for $k = t = 2$, in [18] for $t = 2$ (for parameters (n, k, d) at which $(n, k, d+1)$ MSR codes exist), and in [6] for $d = k$. We believe that moving from the cooperative repair model [6], [7] to the CMR model would allow us to construct MDS codes (MSMR codes) that enable repair-bandwidth efficient repair of t nodes for an expanded set of system parameters. We exhibit this by designing a scheme to perform centralized repair of multiple nodes in a distributed storage system employing a zigzag code [3].

2) Centralized repair of multiple node failures in a zigzag code [3]: The zigzag codes, as introduced in [3], are MDS codes that allow for repair of a single node failure among systematic nodes by contacting $d = n-1$ (all of the) remaining nodes. The zigzag codes are associated with the MSR point [1] (or MSMR point with $t = 1$ (cf. (6))) as each of the contacted $d = n - 1$ nodes contributes $\beta = \frac{\alpha}{d-k+1} = \frac{\alpha}{n-k}$ symbols during the repair of a single failed node. This amounts to the repair bandwidth of $\gamma = d\beta = \frac{n-1}{n-k}\alpha$. Here, we show that the framework of zigzag codes also enable repair of multiple nodes in the CMR model.

Given the space limitation we just state the achievable parameters in the following result. We then illustrate the proposed centralized repair scheme with the help an example of an $(n = 6, k = 3)$-zigzag code where we can simultaneously repair any 2 systematic nodes. We refer the reader to a longer version of this paper for details [19].

Theorem 1. *For an $(n = k + r, k)$ zigzag code with $r = n - k \geq 2$, it is possible to repair any $1 \leq t \leq 3$ systematic nodes in a centralized manner with the optimal repair-bandwidth (cf. 6) by contacting $d = n - t$ helper nodes.*

Example 1 (Repairing $t = 2$ systematic nodes in a $(6, 3)$-zigzag code)**.** Let's consider a zigzag code with the parameters $n = 6, k = 3$ and $\alpha = 9$ from [3]. This code is illustrated in Table 1 where each column (indexed from 1 to 6) represents a storage node. Recall that, in the event of a single node failure, this code allows for the repair of any systematic node failure by contacting $\hat{d} = 5$ remaining nodes and downloading $\beta = \frac{\alpha}{n-k} = 3$ symbols from each of these nodes. We now show that we can use this same construction (with required modifications of the non-zero

1	2	3	4	5	6
$x_{0,0}$	$x_{0,1}$	$x_{0,2}$	$x_{0,0}+x_{0,1}+x_{0,2}$	$x_{0,0}+x_{6,1}+x_{2,2}$	$x_{0,0}+x_{3,1}+x_{1,2}$
$x_{1,0}$	$x_{1,1}$	$x_{1,2}$	$x_{1,0}+x_{1,1}+x_{1,2}$	$x_{1,0}+x_{7,1}+x_{0,2}$	$x_{1,0}+x_{4,1}+x_{2,2}$
$x_{2,0}$	$x_{2,1}$	$x_{2,2}$	$x_{2,0}+x_{2,1}+x_{2,2}$	$x_{2,0}+x_{8,1}+x_{1,2}$	$x_{2,0}+x_{5,1}+x_{0,2}$
$x_{3,0}$	$x_{3,1}$	$x_{3,2}$	$x_{3,0}+x_{3,1}+x_{3,2}$	$x_{3,0}+x_{0,1}+x_{5,2}$	$x_{3,0}+x_{6,1}+x_{4,2}$
$x_{4,0}$	$x_{4,1}$	$x_{4,2}$	$x_{4,0}+x_{4,1}+x_{4,2}$	$x_{4,0}+x_{1,1}+x_{3,2}$	$x_{4,0}+x_{7,1}+x_{5,2}$
$x_{5,0}$	$x_{5,1}$	$x_{5,2}$	$x_{5,0}+x_{5,1}+x_{5,2}$	$x_{5,0}+x_{2,1}+x_{4,2}$	$x_{5,0}+x_{8,1}+x_{3,2}$
$x_{6,0}$	$x_{6,1}$	$x_{6,2}$	$x_{6,0}+x_{6,1}+x_{6,2}$	$x_{6,0}+x_{3,1}+x_{8,2}$	$x_{6,0}+x_{0,1}+x_{7,2}$
$x_{7,0}$	$x_{7,1}$	$x_{7,2}$	$x_{7,0}+x_{7,1}+x_{7,2}$	$x_{7,0}+x_{4,1}+x_{6,2}$	$x_{7,0}+x_{1,1}+x_{8,2}$
$x_{8,0}$	$x_{8,1}$	$x_{8,2}$	$x_{8,0}+x_{8,1}+x_{8,2}$	$x_{8,0}+x_{5,1}+x_{7,2}$	$x_{8,0}+x_{2,1}+x_{6,2}$

Fig. 1: Repair of the first two systematic nodes in a $(6,3)$-zigzag code. (Coding coefficients of the parity symbols are not specified.) Blue (red) colored symbols contribute in the repair of only node 1 (respectively, 2) in the case of single node failure. Green colored symbols contribute in the repair of both node 1 and node 2 in the case of single node failure. Magenta colored symbols denote the additional symbols that need to be downloaded to enable the centralized repair of both the nodes.

coefficients in coded symbols) to repair 2 systematic node failures by contacting $d = n - 2 = 4$ remaining nodes. We download $t\frac{\alpha}{d-k+2} = 2\frac{\alpha}{n-k} = 6$ symbols from each of the $d = 4$ contacted nodes.

Assume that node 1 and 2 are in failure. We download the colored symbols from node 3 to node 6 in Figure 1 to repair these two nodes. Using the downloaded symbols, we get the following 18 combinations in the 18 unknown information symbols. (We suppress the coefficients of the linear combinations here.)

$$\begin{aligned}
&x_{0,0} + x_{6,1},\; x_{1,0} + x_{4,1},\; x_{2,0} + x_{2,1},\; x_{3,0} + x_{0,1},\\
&x_{4,0} + x_{7,1}, x_{5,0} + x_{5,1},\; x_{6,0} + x_{0,1},\; x_{7,0} + x_{7,1},\\
&x_{8,0} + x_{5,1}, x_{2,0} + x_{8,1},\; x_{1,0} + x_{7,1},\; x_{6,0} + x_{6,1},\\
&x_{2,0} + x_{5,1},\; x_{7,0} + x_{4,1}, x_{0,0} + x_{3,1},\; x_{8,0} + x_{2,1},\\
&x_{1,0} + x_{1,1},\; x_{0,0} + x_{0,1}. \qquad (8)
\end{aligned}$$

Now, we need to show that it is possible to choose the coding coefficients in such a manner that these 18 equations allow us to recover the desired 18 symbols. Assuming that A denotes the 18×18 coefficient matrix of the aforementioned 18 combinations, it is a necessary and sufficient (with large enough field size) condition for the matrix A to be full rank that the natural bipartite graph associated with the matrix A contains a perfect matching[2] [20], [21]. We illustrate one such perfect matching in (8), where the colored unknown symbol in a combination represents the unknown symbol matched by that combination. The similar argument can be performed for the remaining combinations of 2 failed systematic nodes.

3) MSMR point: The achievability results above together with the repair bandwidth bound reported in the previous section, see Remark 2, results in the following characterization.

Theorem 2. *The MSMR point for the* $(n, k, d, t, \alpha, \gamma)$*-CMR model is given by* $\alpha_{\mathrm{MSMR}} = \frac{\mathcal{M}}{k}$ *and* $\gamma_{\mathrm{MSMR}} = \frac{\mathcal{M}dt}{k(d-k+t)}$.

V. MBMR Codes

In this section, we focus on the other extremal point of the storage vs. repair-bandwidth trade-off, namely the MBMR point.

A. Repair bandwidth bound

For the MBMR point, depending on whether $t|k$ or $t \nmid k$, we state the following two results.

Proposition 3. *Assume that* $t|k$*. Consider an* (n, k)*-DSS that stores a file of size* $\mathcal{M}$ *and enables repair of* t *failed nodes under a* $(d, t, \alpha_{\mathrm{MBMR}}, \gamma_{\mathrm{MBMR}})$*-CMR model. Then, denoting the entropy of* t *nodes as* H_t*, we have*

$$t\alpha_{\mathrm{MBMR}} \geq H_t = \gamma_{\mathrm{MBMR}}, \qquad (9)$$

$$\gamma_{\mathrm{MBMR}} \geq \frac{\mathcal{M}2dt}{k(2d-k+t)}. \qquad (10)$$

Proof. Note that the MBMR point has $H(\mathbf{x}_{\mathcal{S}}) = \gamma_{\mathrm{MBMR}}$ for every $\mathcal{S} \subseteq [n]$ such that $|\mathcal{S}| = t$. Therefore, we have

$$\gamma_{\mathrm{MBMR}} = H(\mathbf{x}_{\mathcal{S}}) \leq \sum_{i \in \mathcal{S}} H(\mathbf{x}_i) \leq t\alpha_{\mathrm{MBMR}}.$$

In order to establish the lower bound on γ_{MBMR} in (10), we use $n_i = t, \forall i \in [a]$ in the bound (1), we obtain

$$\mathcal{M} \leq \sum_{i=1}^{k/t} \big(d - (i-1)t\big)\beta = \frac{k}{t}\left(\frac{2d-k+t}{2}\right)\beta. \qquad (11)$$

This implies that $\gamma_{\mathrm{MBMR}} = d\beta \geq \frac{\mathcal{M}2dt}{k(2d-k+t)}$. □

Proposition 4. *Consider an* (n, k)*-DSS that stores a file of size* $\mathcal{M}$ *and enables repair of* t *failed nodes under a* $(d, t, \alpha_{\mathrm{MBMR}}, \gamma_{\mathrm{MBMR}})$*-CMR model. Then, the bounds given in* (9) *and* (10) *hold for the case of* $t \nmid k$*, if* $H_b \geq \binom{\beta}{t}\left[b\left(\frac{2d+t-1}{2}\right) - \binom{b}{2}\right]$*, where* $b = k \pmod t$*, and* H_b *denotes entropy of* b *nodes in the system.*

Proof. The bound in (9) follows from the similar analysis as presented in the proof of Proposition 4. In order to establish

[2] The left and the right nodes in the bipartite graph correspond to the combinations and the unknowns, respectively.

(10), we select $g = \lfloor k/t \rfloor + 1 = a + 1$ disjoint sets of nodes indexed by the sets $\mathcal{S}_1, \mathcal{S}_2, \dots, \mathcal{S}_g$ such that $n_1 = |\mathcal{S}_1| = b$ and $n_i = |\mathcal{S}_i| = t$ for $i \in \{2, 3, \cdots, g = a + 1\}$. Note that we have $\sum_i n_i = k$. Utilizing this particular sequence of sets in (4) along with the fact that we have $H(\mathbf{x}_{\mathcal{S}_i}) = d\beta$ for $2 \leq i \leq g$, we obtain

$$\mathcal{M} \leq H(\mathbf{x}_{\mathcal{S}_1}) + \sum_{i=1}^{a} \big(d - (i-1)t - b\big)\beta. \tag{12}$$

Note that the choice of the set $\mathcal{S}_1$ is arbitrary and all the nodes in the system are equivalent in terms of their information content. Therefore, $H_b = H(\mathcal{S}_1)$ (the amount of information stored on b nodes indexed by the set $\mathcal{S}_1$) only depends on b. It follows from (12) that

$$\mathcal{M} \leq H_b + \left(\frac{2d - k + (t - b)}{2}\right) a\beta \tag{13}$$

In order to have the bound in (10) we need the RHS of (13) to be at least the RHS of (11), i.e.,

$$H_b + \left(\frac{2d - k + (t - b)}{2}\right) a\beta \geq \frac{k}{t}\left(\frac{2d - k + t}{2}\right)\beta.$$

This implies that

$$H_b = \left(\frac{\beta}{t}\right)\left[b\left(\frac{2d + t - 1}{2}\right) - \binom{b}{2}\right]. \tag{14}$$

□

Remark 3. A code that allows for repair of t failed nodes with $H_t = \gamma = \frac{\mathcal{M}2dt}{k(2d-k+t)}$ is an MBMR code for the case of $t|k$ and $t \nmid k$, if for the latter case the system also operates at $H_b \geq \left(\frac{\beta}{t}\right)\left[b\left(\frac{2d+t-1}{2}\right) - \binom{b}{2}\right]$.

B. Constructions and the characterization of the MBMR point

1) Constructions from existing MBCR codes: MBCR codes have $\alpha_{MBCR} = \frac{\mathcal{M}}{k}\frac{2d+t-1}{2d+t-k}$, $\beta_{MBCR} = \frac{\mathcal{M}}{k}\frac{2}{2d+t-k}$, and $\beta'_{MBCR} = \frac{\mathcal{M}}{k}\frac{1}{2d+t-k}$. A construction of MBCR codes for all parameters is provided in [22], where the entropy accumulation for MBCR codes is also characterized. In particular, entropy of $b \leq k$ nodes is given by $H_b = \left(b\left(\frac{2d+t-1}{2}\right) - \binom{b}{2}\right)\beta$.

Proposition 5. *A code $\mathcal{C}$ that operates as an MBCR code is also an MBMR code for the CMR model that operates at $\alpha = \frac{\mathcal{M}(2d+t-1)}{k(2d+t-k)}$ and $H_b \geq \left(\frac{\beta}{t}\right)\left[b\left(\frac{2d+t-1}{2}\right) - \binom{b}{2}\right]$.*

Proof. Consider that each failed node contact to the same set of d nodes in the MBCR code $\mathcal{C}$. This results in a repair bandwidth of at most $\gamma = td\beta_{MBCR} = \frac{\mathcal{M}2dt}{k(2d+t-k)}$. Entropy of t nodes in this code is given by $H_t = \left(t\left(\frac{2d+t-1}{2}\right) - \binom{t}{2}\right)\frac{\mathcal{M}}{k}\frac{2}{2d+t-k} = \frac{\mathcal{M}2dt}{k(2d+t-k)} = \gamma$. These and also the entropy of b nodes meet the conditions stated in Remark 3, establishing the claimed result. □

Remark 4. In general, for MBMR codes, we have the condition that $t\alpha \geq H_t = \gamma_{MBMR}$. It is not clear if α can be further reduced than that in Proposition 5, e.g., when $b = 0$.

2) MBMR point: The achievability results above together with the repair bandwidth bound reported in the previous section results in the following characterization.

Theorem 3. *Let* $k \pmod{t} = b$. *Then, for the CMR models satisfying* $H_b \geq \left(\frac{\beta}{t}\right)\left[b\left(\frac{2d+t-1}{2}\right) - \binom{b}{2}\right]$, *the MBMR point is given by* $H_t = \gamma_{MBMR} = \frac{\mathcal{M}2dt}{k(2d+t-k)}$.

Acknowledgement

The authors would like to thank Salim El Rouayheb for pointing us to the work by Cadambe et al. [8].

References

[1] A. G. Dimakis, P. Godfrey, Y. Wu, M. Wainwright, and K. Ramchandran. Network coding for distributed storage systems. *IEEE Trans. Inf. Theory*, 56(9):4539–4551, 2010.

[2] K. Rashmi, N. Shah, and P. Kumar. Optimal exact-regenerating codes for distributed storage at the MSR and MBR points via a product-matrix construction. *IEEE Trans. Inf. Theory*, 57:5227–5239, 2011.

[3] I. Tamo, Z. Wang, and J. Bruck. Zigzag codes: MDS array codes with optimal rebuilding. *IEEE Trans. Inf. Theory*, 59(3):1597–1616, 2013.

[4] D. Papailiopoulos, A. G. Dimakis, and V. Cadambe. Repair optimal erasure codes through hadamard designs. *IEEE Trans. Inf. Theory*, 59(5):3021–3037, 2013.

[5] B. Sasidharan, G. K. Agarwal, and P. V. Kumar. A high-rate MSR code with polynomial sub-packetization level. *CoRR*, abs/1501.06662, 2015.

[6] K. W. Shum and Y. Hu. Cooperative regenerating codes. *IEEE Trans. Inf. Theory*, 59(11):7229–7258, 2013.

[7] A.-M. Kermarrec, N. Le Scouarnec, and G. Straub. Repairing multiple failures with coordinated and adaptive regenerating codes. In *Proc. of 2011 NetCod*, pages 1–6, 2011.

[8] V. R. Cadambe, S. A. Jafar, H. Maleki, K. Ramchandran, and C. Suh. Asymptotic interference alignment for optimal repair of mds codes in distributed storage. *IEEE Transactions on Information Theory*, 59(5):2974–2987, May 2013.

[9] P. Hu, C. W. Sung, and T. H. Chan. Broadcast repair for wireless distributed storage systems. *CoRR*, abs/1603.00154, 2016.

[10] P. Gopalan, C. Huang, H. Simitci, and S. Yekhanin. On the locality of codeword symbols. *IEEE Trans. Inf. Theory*, 58(11):6925–6934, 2012.

[11] A. S. Rawat, O. O. Koyluoglu, N. Silberstein, and S. Vishwanath. Optimal locally repairable and secure codes for distributed storage systems. *IEEE Trans. Inf. Theory*, 60(1):212–236, 2014.

[12] G. M. Kamath, N. Prakash, V. Lalitha, and P. V. Kumar. Codes with local regeneration and erasure correction. *IEEE Trans. Inf. Theory*, 60(8):4637–4660, Aug 2014.

[13] I. Tamo and A. Barg. A family of optimal locally recoverable codes. *IEEE Trans. Inf. Theory*, 60(8):4661–4676, Aug 2014.

[14] A. S. Rawat, A. Mazumdar, and S. Vishwanath. Cooperative local repair in distributed storage. *EURASIP J. Adv. Signal Process.*, pages 1–17, 2015.

[15] N. Prakash, V. Lalitha, and P. V. Kumar. Codes with locality for two erasures. In *Proc. of 2014 IEEE International Symposium on Information Theory (ISIT)*, pages 1962–1966, June 2014.

[16] W. Song and C. Yuen. Locally repairable codes with functional repair and multiple erasure tolerance. *arXiv preprint arXiv:1507.02796*, 2015.

[17] N. Le Scouarnec. Exact scalar minimum storage coordinated regenerating codes. In *Proceedings of 2012 IEEE International Symposium on Information Theory (ISIT)*, pages 1197–1201, 2012.

[18] J. Li and B. Li. Cooperative repair with minimum-storage regenerating codes for distributed storage. In *Proc. of 2014 IEEE INFOCOM*, pages 316–324, 2014.

[19] A. S. Rawat, O. O. Koyluoglu, and S. Vishwanath. Centralized repair of multiple node failures with applications to communication efficient secret sharing. *CoRR*, abs/1603.04822, 2016.

[20] L. Lovász. On determinants, matchings, and random algorithms. In *Fundamentals of Computing Theory*. Akademia-Verlag, Berlin, 1979.

[21] N. Alon. Combinatorial nullstellensatz. *Comb. Probab. Comput.*, 8(1-2):7–29, 1999.

[22] A. Wang and Zhang. Exact cooperative regenerating codes with minimum-repair-bandwidth for distributed storage. In *Proc. of 2013 IEEE INFOCOM*, pages 400–404, 2013.