UNIVERSITY OF CALIFORNIA, SAN DIEGO

Distributed Storage with Communication Costs

A thesis submitted in partial satisfaction of the
requirements for the degree of
Master of Science

in

Electrical & Computer Engineering
(Communication Theory and Systems)

by

Craig Armstrong

Committee in charge:

Professor Alexander Vardy, Chair
Professor Young-Han Kim
Professor Paul H. Siegel

2011

The thesis of Craig Armstrong is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

Chair

University of California, San Diego

2011

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGEMENTS

I firstly wish to thank my supervisor, Professor Alex Vardy, for all of his support during my studies and research as well as his patience, guidance and invaluable direction.

I would next like to extend my deepest gratitude to Eitan Yaakobi for his generous assistance and support in finding my way in graduate research, his untiring positivity, encouragement and also for proposing to me the problem that would become this thesis.

I am also sincerely grateful to the other members of my committee, Professors Young-Han Kim and Paul Siegel, for their advice and ideas as well as their exceptional teaching.

My appreciation and acknowledgement is expressed as well at this opportunity for the research support provided by the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Powell Foundation.

And finally, a heartfelt thanks to my family and all of my friends for their unwavering support and faith in me in everything that I do.

ABSTRACT OF THE THESIS

Distributed Storage with Communication Costs

by

Craig Armstrong

Master of Science in Electrical & Computer Engineering
(Communication Theory and Systems)

University of California, San Diego, 2011

Professor Alexander Vardy, Chair

Distributed storage systems provide reliable storage of data by dispersing redundancy across multiple nodes. As individual nodes are unreliable this protects the integrity of the data against the failure of nodes. In order to maintain this reliability new nodes must be introduced into the system whenever nodes are lost which restore the redundancy. This process involves having a new node download information from remaining nodes and is known as the *repair problem*.

In this thesis, we consider networks with communication costs associated to each link and explore means to minimize the cost of performing these repairs. We do this by considering a generalized method of repair wherein the amount of information downloaded to a new node varies amongst the other nodes in the network. We find that when nodes store the minimum amount of data that the minimum cost can be achieved by *quasi-uniform repair*, where the same amount of data is downloaded from each node communicated with. We also consider systems with the additional freedom that the amount of storage is allowed to vary from node to node and look at repair cost minimization there as well.

Chapter 1

Motivation and Overview

The need for efficient and reliable means to store large amounts of data across a collection of devices has become increasingly important in recent years. In systems where the individual nodes can be be unreliable a method of introducing redundancy to create a reliable system as a whole that also takes into consideration the limitations of the system is an important problem. In particular, individual nodes may have limited storage capacity and the transfer of data throughout the network may be a costly or time-consuming procedure. Examples of distributed storage systems include data centers, peer-to-peer storage applications, wireless sensor networks and distributed file systems.

A particular issue for these systems that has been studied is how to replace a node after it fails or leaves the network. It is a critical feature of distributed storage systems that they maintain their reliability over a long period of time and thus when a node is lost we must create a new node in its place that restores the lost redundancy and prevents degradation of the system. Methods of coding that allow this node regeneration to be performed efficiently were first introduced in [3] by Dimakis et al., which we will present the results of as well as other works in Chapter 2. The repair model covered here is what we call *quasi-uniform repair* where a node is required to download the same amount of information from all existing nodes to which it connects during repair.

In this thesis we seek to model deployable systems more accurately by imposing communication costs on the links in the network as variable communication

link qualities are to be expected in real systems. We then study techniques to minimize the cost of data transfers when operating such systems.

In Chapter 3 we present results on a generalization of the repair model seen in Chapter 2 where we allow variable download quantities and use these to analyze the behavior of the system when imposing the communication costs. We then present our main result on the optimality of the restriction to the quasi-uniform repair model from [3] for minimizing cost during node repair when nodes store the minimum possible amount of data.

Finally, in Chapter 4 we add an additional level of generality to the system model by allowing the storage capacities of the nodes to vary. We explore the requirements on repairs in this scenario and observe the changes in behavior of repair cost minimization.

Chapter 2

An Introduction to Coding for Distributed Storage

In this chapter we formally state the repair problem and present some relevant results from other authors. The main result that we cover is a theorem from [3] which establishes the fundamental tradeoff between the amount of data stored at each node and the total amount of bandwidth required for a repair. Throughout this thesis the data quantities will be referred to as bits, but we will allow the values to be real numbers as files that would actually be considered are quite large and fractional amounts can be approached by dividing them into relatively small fragments.

2.1 The Repair Problem

We now establish the mathematical model and setup for the repair problem. A file of size M bits needs to be reliably stored across a network consisting of n storage nodes, where each storage node has a non-negligable risk of failure. In the event of a failure, the entirety of the data stored at the node is lost. So, the questions we would like to answer are how can we code and store this file across the distributed storage network so as to incur as little storage expansion as possible while also maximally avoiding data loss in the event of node failures and how do we restore the lost redundancy in such events?

It is apparent that in answer to the latter question we should create a new node in place of the failed node, a process that we call node *repair*. Thus, any proposed coding scheme must be designed with this process as a fundamental consideration.

A naive solution to the problem is to store the entire file at every node in the network so that at any point where we have at least one node remaining we can recover the file. Here, redundancy can be restored via full replication of the file into a new node each time a failure occurs. This solution requires the storage of the entire file size, M, at every node and a total storage of $n \cdot M$ bits in the network. Additionally, the repair of any lost node requires the transfer of a full M bits.

A more efficient approach is to use an (n, k)-MDS (maximum distance separable) code in order to distribute the file across the network. In this way, each node will store a block of size $\frac{M}{k}$ bits and any k nodes will be sufficient for reconstruction of the original file. Thus, this system can tolerate $n - k$ failures, without repair, while still maintaining reliability. This scheme also optimizes the redundancy-reliability tradeoff for any system that requires at least k nodes for reconstruction as each individual node here requires the minimum amount of storage to satisfy this condition.

In terms of storage this approach is thus very desirable, but we also need to establish a procedure for node repair. A very straightforward method would be to have the new node connect to any k of those remaining and download all of the data from each. This node could then completely reconstruct the original file and compute a newly encoded block of data to store. But what if network resources are limited and we need to incur as little bandwidth consumption as possible during the repair process? This method, again, would require each repair to transfer the entirety of the file size M. This leads to the question as to if it is possible to improve upon this, which we may, in fact, as demonstrated in the following example.

Example provided in [2]:

In this example we use a $(4, 2)$-MDS code and also allow sub-packetization. That is, the data at each node is split into multiple blocks, in this case 2. Thus, $M = 4$ blocks here.

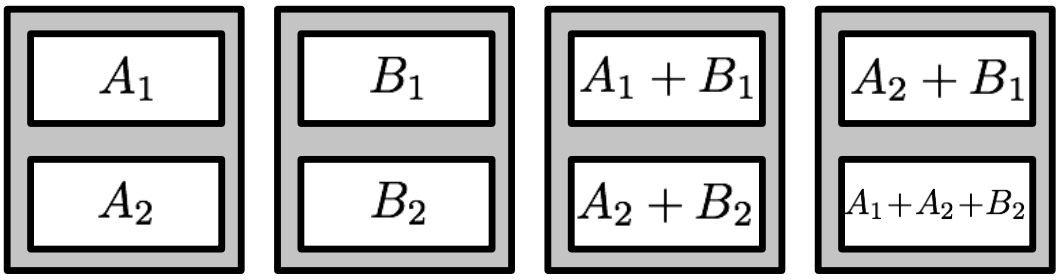

Figure 2.1: Original contents of the nodes.

Figure 2.1 shows the original contents of the storage nodes. Assume then that node 1 fails and must be repaired by downloading information from the remaining 3 nodes. From Figure 2.2 we see that only 3 blocks (as opposed to a full $M = 4$) need to be communicated.

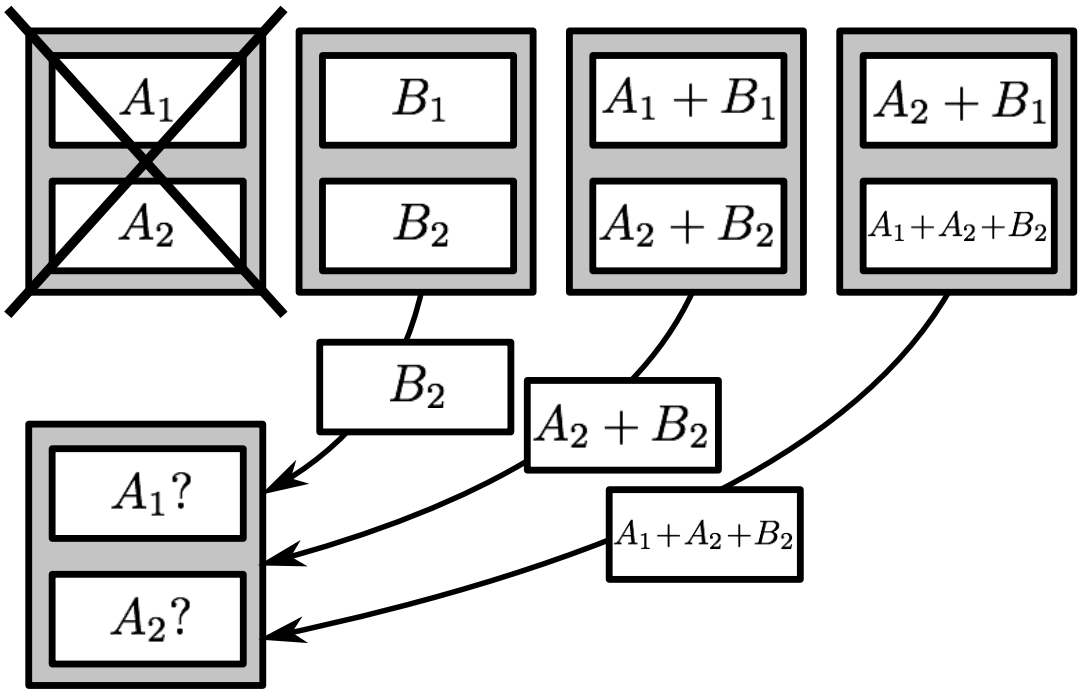

Figure 2.2: Repair of node 1.

Now assume that node 4 fails. In Figure 2.3 we see that again only 3 blocks are needed for repair, but now we require that nodes be able to compute linear combinations of their data before transmission. It turns out that this example is optimal in terms of the minimum required repair bandwidth, which we will establish in the next section.

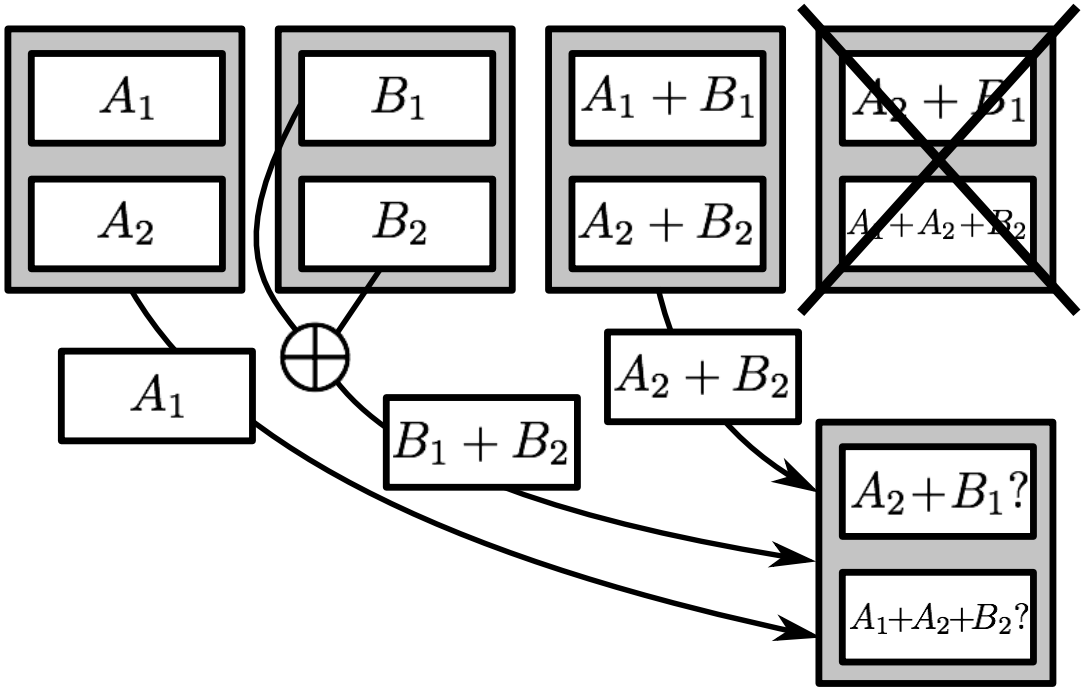

Figure 2.3: Repair of node 4.

The repairs in this example demonstrate *exact repair*. Exact repair is where a failed node is replaced by a new node with the exact data as was previously stored. Alternatively, we could consider *functional repair* where the constraint on the data in the new node is only that the reconstruction requirement is maintained in the system so that any k nodes remain sufficient for reconstruction. In this case, we may not restore exactly what was lost, but the system maintains the same level of reliability. There is also a third type of repair known as *hybrid repair*, which is a combination of the previous two types. There, the data is stored using a systematic code and the systematic parts are repaired exactly while the remaining parts need only be repaired functionally.

In this thesis, we will allow sub-packetization and the computation of linear combinations within nodes, as was seen in the previous example, and we will be considering exclusively functional repair.

2.2 Minimum Repair Bandwidth

The problem of finding the minimum bandwidth necessary for a successful repair was completely characterized in [3] for quasi-uniform functional repair, where quasi-uniform repair imposes the requirement that a node download the same quantity from all of the nodes to which it connects during repair. As the other forms of repair are encompassed within functional repair this also provides a lower bound on the bandwidth necessary in those cases.

First, let's establish the parameters for the problem:

- we have a complete network of n nodes
- every k nodes are required to suffice for reconstruction
- the size of the file to be stored is M bits
- each node stores α bits
- when a node fails it is repaired by downloading β bits each from any d ($\geq k$) of the remaining $(n-1)$ nodes
- the repair bandwidth is then $d\beta$

It was shown that this problem could be solved by making use of multicasting results by introducing *information flow graphs*. An information flow graph is a graph representing the network and its progression as failures and repairs occur. It consists of storage nodes, a source (S) and Data Collectors (DCs). Each storage node is represented by an input node and an output node which are connected by an edge of capacity α, the storage capacity of a node. Each time a node fails it will become inactive and a new node will be added to the system.

Initially, the source node, S, is connected to the original n nodes via links of infinite capacity. When a node fails and becomes inactive, a new node enters the system and is connected to any d of the currently active nodes with links of capacity β. Data Collectors represent all possible requests for reconstruction and

at any point in the evolution of the graph may connect to any set of k active nodes with infinite capacity links. Figure 2.4 shows an example.

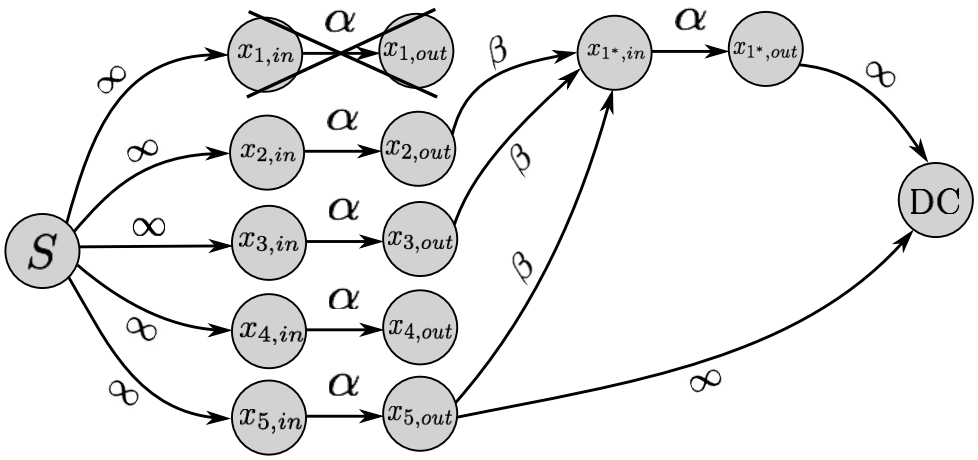

Figure 2.4: Example information flow graph where $n = 5, k = 2, d = 3$.

A set of values n, k, d, α, β is achievable if the minimum cut between S and all possible DCs are $\geq M$ for all possible evolutions of node failures and repairs. This provides both an information theoretic lower bound and is achievable through deterministic network codes. The following result was established in [3].

Theorem: For any $\alpha \geq \alpha^*(n, k, d, \gamma = d\beta)$, the points $(n, k, d, \alpha, \gamma)$ are feasible and linear network codes suffice to achieve them. It is information theoretically impossible to achieve points with $\alpha < \alpha^*(n, k, d, \gamma)$. The threshold function is the following:

$$\alpha^*(n, k, d, \gamma) = \begin{cases} \frac{M}{k}, & \gamma \in [f(0), \infty) \\ \frac{M - g(i)\gamma}{k - i}, & \gamma \in [f(i), f(i-1)) \end{cases} \tag{2.1}$$

where

$$f(i) \triangleq \frac{2Md}{(2k-i-1)i+2k(d-k+1)}, \tag{2.2}$$

$$g(i) \triangleq \frac{(2d-2k+i+1)i}{2d} \tag{2.3}$$

$$\text{for} \quad i \in \{0,...,k-1\} \tag{2.4}$$

It was also shown that for given n, k, d the minimum bandwidth is

$$\gamma_{\min} = f(k-1) = \frac{2Md}{2kd-k^2+k} \tag{2.5}$$

and that this minimum bandwidth is a decreasing function of d. Thus, the repair bandwidth is smallest when $d = n-1$, where the new node communicates with all remaining nodes during repair.

2.3 Extremal Points

The optimal tradeoff curve for achievable codes given in the above theorem has two extremal points which are of particular interest:

- The Minimum Storage Regenerating (MSR) codes have minimum possible α, and
- The Minimum Bandwidth Regenerating (MBR) codes with minimum possible γ

It is clear that we must have $\alpha \geq \frac{M}{k}$ to satisfy the reconstruction requirement and so the MSR point, which we get by minimizing γ after fixing this α in the theorem (giving $i = 0$), corresponds to

$$(\alpha_{MSR}, \gamma_{MSR}) = \left(\frac{M}{k}, \frac{Md}{k(d-k+1)}\right)$$

Then, letting $d = n-1$ to minimize γ, we get

$$(\alpha_{MSR}, \gamma_{MSR}^{\min}) = \left(\frac{M}{k}, \frac{M}{k} \cdot \frac{n-1}{n-k}\right)$$

and so we see that at the MSR point a factor of at least $\frac{n-1}{n-k}$ more data must be downloaded during repair than will be ultimately stored in the node.

The MBR point, which we get by first minimizing γ in the theorem then α (giving $i = k - 1$), on the other extreme is

$$(\alpha_{MBR}, \gamma_{MBR}) = \left(\frac{2Md}{2kd - k^2 + k}, \frac{2Md}{2kd - k^2 + k} \right)$$

With $d = n - 1$,

$$(\alpha_{MBR}, \gamma_{MBR}^{\min}) = \left(\frac{M}{k} \frac{2n-2}{2n-k-1}, \frac{M}{k} \frac{2n-2}{2n-k-1} \right)$$

Note how here the repair bandwidth is the same as the amount of stored data, although more data must be stored at each node and these codes are no longer optimal in terms of the redundancy-reliability tradeoff.

Chapter 3

Minimization of the Cost for Repair

We now begin the original work of this thesis where we explore the minimization of repair cost. In order to accomplish this we must first establish conditions on a generalized method of repair, which will provide flexibility to the system over which we can optimize. We then establish our cost function and show that repair cost minimization can be achieved by quasi-uniform repairs when the nodes contain minimum storage.

3.1 Generalized Repair

The majority of existing work has considered the quasi-uniform model of repair, as seen in the previous chapter, and explored how to achieve minimal repair bandwidth within each of the repair types. In this paper, we will be considering a more general method of repair where the amounts of data downloaded from the remaining nodes are variable. That is, if node 1 fails then it is repaired by downloading $\beta_{i,1}$ bits from each of the remaining nodes $i \in \{2, .., n\}$. We will need to find a characterization of achievable rate tuples, $(\beta_{2,1}, \beta_{3,1}, ..., \beta_{n,1})$, for successful functional repair in this scenario.

This repair model was also considered in [4], but only a bound on the total repair bandwidth required was presented there. In [5], variable rate repair was also

analyzed, but under a more general requirement for reconstruction referred to as *'flexible reconstruction,'* where any DC that is connected to all active nodes and downloads μ_i from each node i such that $\sum_{i=1}^{n} \mu_i \geq M$ must be able to reconstruct the file.

Here, we continue to use the reconstruction requirement that at all times any k nodes are sufficient for reconstruction. The following 2 lemmas provide a partial characterization of achievable repair rates.

Lemma 1: For functional repair of a node, i, and with storage capacity α for all nodes, we have the following necessary condition on repair rates $\{\beta_{j,i}\}$ communicated from the remaining nodes $j \in [n]\backslash\{i\}$:

$$\min_{\substack{\mathcal{R}\subseteq[n]\backslash\{i\},\\ |\mathcal{R}|=n-k}} \sum_{j\in\mathcal{R}} \beta_{j,i} \geq M - (k-1)\alpha \tag{L1}$$

and any set $\{\beta_{j,i}\}$ satisfying this condition is achievable using network codes for a single repair.

Proof of Lemma 1:

We will consider first a single functional repair. This proof will use information flow graphs as discussed in Section 2.2. WLOG let node 1 fail and be replaced via repair with a new node 1^*. We must ensure that the repair rates $\{\beta_{j,1^*}\}$ are sufficient so that the minimum cut in the graph is $\geq M$. This is both an information theoretic minimum bound and achievable with deterministic network codes, as shown in [3]. See Figure 3.1 for a single repair information flow graph.

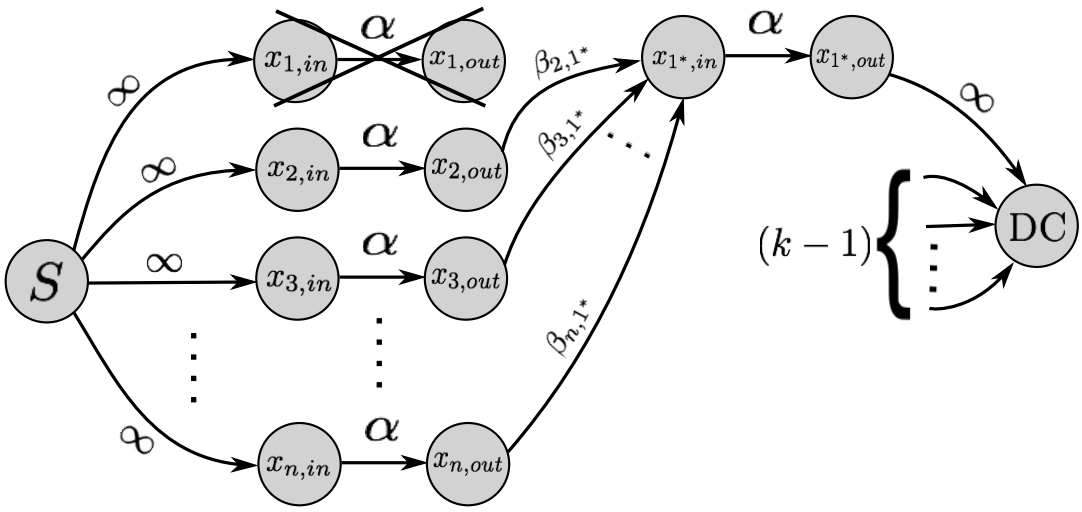

Figure 3.1: Information flow graph for 1 repair.

Consider any DC that is connected to our newly repaired node 1^* and $(k-1)$ others (as reconstruction for any other DC is trivial) and call this set of $(k-1)$ storage nodes $\mathcal{D}$. Let the set $\mathcal{S}$ of nodes in the information flow graph be a graph cut, where the source S must be in $\mathcal{S}$ and DC $\in \mathcal{S}^C$. For each of the $(k-1)$ nodes $i \in \mathcal{D}$, we must have $x_{i,out} \in \mathcal{S}^C$ for any minimum cut as $x_{i,out}$ connects to DC with infinite capacity and these edges will not be included in a minimum cut. So, each of these $(k-1)$ nodes will contribute α to the min cut. Note also that for the same reason $x_{1^*,out} \in \mathcal{S}^C$. A cut with $x_{1^*,in} \in \mathcal{S}$ would include an additional α and thus have value at least $k\alpha$, which is always $\geq M$ (as $\alpha \geq \frac{M}{k}$) and satisfies the cut requirement. So, let $x_{1^*,in} \in \mathcal{S}^C$.

For the nodes $j \in \mathcal{D}^C$ it just remains to be checked if $x_{j,out} \in \mathcal{S}$ or $\mathcal{S}^C$ for the min cut. It is clear that we must have $\beta_{i,j} \leq \alpha,\ \forall i,j$ as a node can not transmit more information than it is storing and so, as $\beta_{j,1^*} \leq \alpha\ \forall j$, we get that $x_{j,out}$ must be in $\mathcal{S}$. These storage nodes then provide altogether $\sum_{j\in\mathcal{D}^C} \beta_{j,1^*}$ to the cut. Taking the minimum over all possible sets $\mathcal{D} \subseteq \{2,...,n\}$ gives:

$$\min_{\substack{\mathcal{D}\subseteq[n]\setminus\{1\},\\ |\mathcal{D}|=k-1}} \sum_{j\in\mathcal{D}^C} \beta_{j,1^*} + (k-1)\alpha$$

Requiring this to be $\geq M$ gives the condition (L1) for a single repair.

Now that we have that the condition is necessary and sufficient for the first repair, it remains to be argued that the condition is necessary in general. For this we simply note that for any subsequent repair as the minimum cut between S and DC must be $\geq M$ then every possible cut must be $\geq M$. So, for any repair, if we consider cuts such that for every currently active node i (original or previously repaired) we have $x_{i,in} \in \mathcal{S}$ the potential minimum cuts of this type will be independent of previous repairs and follow the exact analysis as was just given for the first repair in the system. Therefore, the stated condition must be necessary for any repair.

□

Next we present a result on sufficient conditions when extending the system to a second repair. This lemma pertains to codes at the *minimum storage point* where the nodes all store the minimum required amount of $\frac{M}{k}$.

Lemma 2: For functional repair of any node i at the minimum storage point, $\alpha = \frac{M}{k}$, the condition from Lemma 1 is both necessary and sufficient for achievability for a second repair in the system as well as the first. That is,

$$\min_{\substack{\mathcal{R}\subseteq[n]\setminus\{i\},\\ |\mathcal{R}|=n-k}} \sum_{j\in\mathcal{R}} \beta_{j,i} \geq \frac{M}{k} \tag{L2}$$

characterizes the sets $\{\beta_{j,i}\}$ of repair rates that are achievable for at least 2 failures/repairs at the minimum storage point.

Proof of Lemma 2:

In this proof we will begin with a general α and later introduce the restriction $\alpha = \frac{M}{k}$.

We have already covered the first repair in Lemma 1 and so now consider the second.

WLOG let node 1 be the first failure and node 2 be the second. Note that having the same node failing twice follows the same constraints as a single failure

above as the repair would be performed from the same active nodes as before.

Let the new nodes be 1^* and 2^*, respectively. The original nodes 1 and 2 will become inactive and thus no DC may connect to them here. Also, we must already have that (L1) holds for the first repair, 1^*. We again need to ensure that the every cut on the graph is $\geq M$.

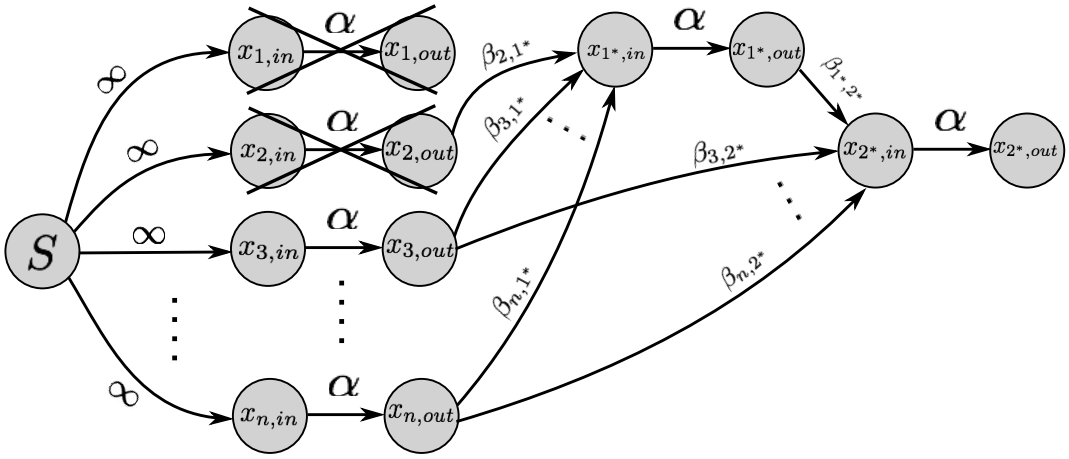

Figure 3.2: Information flow graph for 2 repairs.

Case 1: DC connects to 1^*, but not 2^*

Constraint:

$$\min_{\substack{\mathcal{D}\subseteq\{3,..,n\},\\ |\mathcal{D}|=k-1}} \beta_{2,1^*} + \sum_{i\in\mathcal{D}^C} \beta_{i,1^*} + (k-1)\alpha \geq M$$

which holds by (L1) for 1^*.

Case 2: DC connects to 2^*, but not 1^*

Constraints:

$x_{2^*,in} \in \mathcal{S}$:

$$(k-1)\alpha + \alpha \geq M \tag{3.1}$$

$x_{2^*,in} \in \mathcal{S}^C, x_{1^*,out} \in \mathcal{S}$:

$$(k-1)\alpha + \beta_{1^*,2^*} + \min_{\substack{\mathcal{D}\subseteq\{3,..,n\},\\ |\mathcal{D}|=k-1}} \sum_{i\in\mathcal{D}^C} \beta_{i,2^*} \geq M \tag{3.2}$$

$x_{2^*,in} \in \mathcal{S}^C, x_{1^*,out} \in \mathcal{S}^C, x_{1^*,in} \in \mathcal{S}$:

$$(k-1)\alpha + \alpha + \min_{\substack{\mathcal{D}\subseteq\{3,..,n\},\\ |\mathcal{D}|=k-1}} \sum_{i\in\mathcal{D}^C} \beta_{i,2^*} \geq M \tag{3.3}$$

$x_{2^*,in} \in \mathcal{S}^C, x_{1^*,out} \in \mathcal{S}^C, x_{1^*,in} \in \mathcal{S}^C$:

$$(k-1)\alpha + \beta_{2,1^*} + \min_{\substack{\mathcal{D}\subseteq\{3,..,n\},\\ |\mathcal{D}|=k-1}} \sum_{i\in\mathcal{D}^C} \min\{\alpha, \beta_{i,1^*} + \beta_{i,2^*}\} \geq M \tag{3.4}$$

Case 3: DC connects to 2^* and 1^*

Constraints:

$x_{2^*,in} \in \mathcal{S}, x_{1^*,in} \in \mathcal{S}^C$:

$$(k-1)\alpha + \beta_{2,1^*} + \min_{\substack{\mathcal{D}\subseteq\{3,..,n\},\\ |\mathcal{D}|=k-2}} \sum_{i\in\mathcal{D}^C} \beta_{i,1^*} \geq M \tag{3.5}$$

$x_{2^*,in} \in \mathcal{S}^C, x_{1^*,in} \in \mathcal{S}$:

$$(k-1)\alpha + \min_{\substack{\mathcal{D}\subseteq\{3,..,n\},\\ |\mathcal{D}|=k-2}} \sum_{i\in\mathcal{D}^C} \beta_{i,2^*} \geq M \tag{3.6}$$

$x_{2^*,in} \in \mathcal{S}^C, x_{1^*,in} \in \mathcal{S}^C$:

$$(k-2)\alpha + \beta_{2,1^*} + \min_{\substack{\mathcal{D}\subseteq\{3,..,n\},\\ |\mathcal{D}|=k-2}} \sum_{i\in\mathcal{D}^C} \min\{\alpha, \beta_{i,1^*} + \beta_{i,2^*}\} \geq M \tag{3.7}$$

These constraints can be reduced by the following:

- Contraints (3.1) and (3.3) are satisfied as $\alpha \geq \frac{M}{k}$
- (3.4) and (3.5) follow from (L1) for node 1^*
- (3.2) and (3.6) combined are equivalent to (L1) for node 2^*

Lastly, we consider constraint (3.7):

Since the constraint must hold for the minimization over all sets $\mathcal{D}$ then it must hold for each $\mathcal{D}$. So consider any fixed $\mathcal{D}^*$. If $\min\{\alpha, \beta_{i,1^*} + \beta_{i,2^*}\} = \alpha$ for any $i \in \mathcal{D}^{*C}$ then the inequality becomes the same as (3.4), which holds. So, assume $\beta_{i,1^*} + \beta_{i,2^*} < \alpha,\ \forall i \in \mathcal{D}^{*C}$. Then,

$$\begin{aligned}
\text{LHS of (3.7)} &= (k-2)\alpha + \beta_{2,1^*} + \sum_{i \in \mathcal{D}^{*C}} (\beta_{i,1^*} + \beta_{i,2^*}) \\
&\overset{\text{by (L1) for } 1^* \text{ and } 2^*}{\geq} (k-2)\alpha + \beta_{2,1^*} + 2(M - (k-1)\alpha) \\
&= 2M - k\alpha + \beta_{2,1^*} \\
&\geq 2M - k\alpha
\end{aligned}$$

which is guaranteed $\geq M$ when $\alpha \leq \frac{M}{k} \Rightarrow \alpha = \frac{M}{k}$. Therefore, when $\alpha = \frac{M}{k}$, constraint (L1) $\equiv$ (L2) on both repairs is both necessary and sufficient for achievability.

□

We will now show with the following example the strictness of the requirement that $\alpha = \frac{M}{k}$ for the sufficiency of (L1) in the achievability of the second repair.

Example 1:

As we are considering a system with two repairs we must have $k > 2$ so that we may consider the case where the DC connects to both newly repaired nodes as well as original nodes. We will construct a scenario where the constraint (L1) is satisfied for both repairs, but the constraint (3.7) is not satisfied for the second repair. Thus, (3.7) will not be redundant in this case and is required in the characterization of achievable repair rates. Moreover, this example will use $\alpha > \frac{M}{k}$ such that α may be made arbitrarily close to $\frac{M}{k}$ to illustrate the strictness of this requirement in Lemma 2.

Let $(n-k)$ be arbitrarily large by setting $n >> k$.

Now, consider $\alpha = \frac{M}{k} + \frac{M}{(k(n-k)+k-1)k} + \epsilon \qquad (\rightarrow \frac{M}{k} + \epsilon \text{ as } (n-k) \rightarrow \infty)$

Then, define $\tau :=$ RHS of (L1) $= M - (k-1)\alpha = \frac{M(n-k)}{k(n-k)+k-1} - (k-1)\epsilon$

If we let all of the repair rates for both repairs be

$$\frac{\tau}{(n-k)} = \frac{M}{k(n-k)+k-1} - \frac{(k-1)}{(n-k)}\epsilon \quad (\downarrow 0 \text{ as } (n-k) \to \infty)$$

then, by design, (L1) is satisfied for both repairs.

But,

$$\begin{aligned} \text{LHS of (3.7)} &= (k-2)\alpha + \beta_{2,1^*} + \min_{\substack{\mathcal{D}\subseteq\{3,..,n\},\\ |\mathcal{D}|=k-2}} \sum_{i\in\mathcal{D}^C} \min\{\alpha, \beta_{i,1^*} + \beta_{i,2^*}\} \\ &= (k-2)\alpha + \frac{\tau}{(n-k)} + 2\tau \\ &= M - \epsilon\left(\frac{k(n-k)+k-1}{(n-k)}\right) \\ &< M \end{aligned}$$

And so, constraint (3.7) is not satisfied here. Therefore, $\alpha = \frac{M}{k}$ is a strict requirement in Lemma 2.

Based on the results of Lemmas 1 and 2, we are now led to conjecture that the condition we found will hold in the general case of arbitrarily many failures and repairs, not just the first 2. This conjecture, which will be formally stated below, then provides a complete characterization of achievable generalized repair rates for functional repair at the minimum storage point that we may use to explore further properties of codes for distributed storage networks.

Conjecture: The following condition on repair rates when functionally repairing any node i is necessary and sufficient for achievability via network coding at the minimum storage point for any number of node repairs

$$\min_{\substack{\mathcal{R}\subseteq[n]\setminus\{i\},\\ |\mathcal{R}|=n-k}} \sum_{j\in\mathcal{R}} \beta_{j,i} \geq \frac{M}{k} \tag{C1}$$

3.2 The Cost Function

We will now associate a cost with the use of communication links between nodes in the network. This scenario models more accurately the variability in capacity among the various links in a network that is likely to take place in practice. These links will then incur varying penalties when used for repair. We will again be considering a fully connected network. Here, though, we will associate a cost of $p_{i,j}$ per bit for using link (i, j). We will assume that when a node fails, it is replaced by a new node which inherits all of the cost relationships of the failed node with respect to the other $(n-1)$ nodes. This is a reasonable assumption as the node would likely be repaired at the same location and this will prevent the cost functions from evolving over time. The cost of repairing node i when communicating $\beta_{j,i}$ from each node j is then

$$\sum_{j \in [n] \setminus \{i\}} p_{j,i} \beta_{j,i}$$

The introduction of data transfer costs in distributed storage networks has been previously explored in [6]. The scenario that was presented in this paper was that of the nodes being partitioned into two disjoint sets, where all nodes in each set have the same download cost, C_1 and C_2, respectively. So any node that is downloading information from a node in set 1, regardless of the receiving nodes location, would incur a cost per bit of C_1. Repair is performed by downloading β_1 bits from each of d_1 nodes in the first set and β_2 bits from each of d_2 nodes in the second set. The effect of the choice of these parameters on the repair cost was then analyzed.

Cost functions were also used in [7], where general functional repair conditions with varying repair rates were given by matrices and minimizing repair cost for a single repair was explored numerically.

In this thesis, we will use our conjecture stated at the end of section 3.1 to explore functional repair cost minimization for a system allowing arbitrarily many repairs at the minimum storage point and derive our result on the optimality of quasi-uniform repairs under these conditions.

Definition: Under the assumption of our conjecture, we can define the minimum cost of functionally repairing node i at the minimum storage point as

$$C(i) := \min_{\substack{\sum_{j\in\mathcal{R}} \beta_{j,i} \geq \frac{M}{k} \\ \forall \mathcal{R} \subseteq [n]\setminus\{i\}, |\mathcal{R}|=n-k}} \sum_{j\in[n]\setminus\{i\}} p_{j,i}\beta_{j,i}$$

With this definition in hand, we could then consider the following properties of the network:

- Maximum repair cost:

$$C_{max} = \max_{i\in[n]} C(i)$$

- Average repair cost:

$$\bar{C} = \frac{1}{n}\sum_{i\in[n]} C(i)$$

- or Expected repair cost if varying failure probabilities exist:

$$E[C] = \sum_{i\in[n]} P(\text{node } i \text{ failure}) \cdot C(i)$$

3.3 Minimizing Repair Cost

Let's now see how the minimization of the cost function for a single, specific repair behaves. Say node 1 fails, WLOG, and consider the simplified equation by dropping unnecessary subscripts, giving

$$C(1) = \min_{\substack{\sum_{j\in\mathcal{R}} \beta_{j} \geq \frac{M}{k} \\ \forall \mathcal{R} \subseteq \{2,..,n\}, |\mathcal{R}|=n-k}} \sum_{j\in\{2,..,n\}} p_j\beta_j$$

This is just a linear optimization problem which can easily be solved using linear programming for any given set of parameter values. Consider the following examples:

Let $\alpha = 5$

- $n = 6, k = 3, p[i] = [0, 1, 2, 3, 4]$:
  $C = 15, \beta_1 = \beta_2 = \beta_3 = \beta_4 = 2.5, \beta_5 = 0$
- $n = 6, k = 3, p[i] = [0, 0, 1, 1, 1]$:
  $C = 5, \beta_1 = \beta_2 = \beta_3 = \beta_4 = 2.5, \beta_5 = 0$
- $n = 6, k = 3, p[i] = [2, 9, 5, 9, 7]$:
  $C = 53.\overline{33}, \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 1.\overline{66}$
- $n = 10, k = 6, p[i] = [3, 1, 7, 2, 1, 5, 10, 3, 10]$:
  $C = 52.5, \beta_i = 1.25, \forall i$

We can immediately notice of these examples that every solution is of the quasi-uniform repair type as seen in Chapter 2. This suggests that quasi-uniform repair may, in fact, always be able to achieve the minimum repair cost which leads us to the following result.

Theorem: At the minimum storage point $\alpha = \frac{M}{k}$, the minimum functional repair cost $C(i)$ can always be achieved by a quasi-uniform set of repair rates. Thus, the minimization for $C(i)$ is equivalent to a cost minimization over the parameter d in the quasi-uniform repair model (and minimum associated $\beta = \frac{\alpha}{(d-k+1)}$).

Before we prove this theorem, it should be noted that linear optimization over symmetric constraints does not always result in a quasi-uniform solution and so this result is non-trivial. The following example demonstrates this fact.

Example 2: Consider the objective function to be minimized

$$p_1x_1 + p_2x_2 + p_3x_3 + p_4x_4 \text{ where } p_1 < p_2 < p_3 < p_4$$

subject to

$$\begin{aligned} x_1 + x_2 + x_3 + x_4 &\geq m_1 \\ 0 \leq x_i &\leq m_2 \end{aligned}$$

if $m_2 < m_1 < 2 \cdot m_2$ then we get the optimal solution

$$\begin{aligned} x_1 &= m_2 \\ x_2 &= m_1 - m_2 \\ x_3, x_4 &= 0 \end{aligned}$$

which is not quasi-uniform.

Proof of Theorem:

Without loss of generality the theorem will be proved for $C(n)$.
We first note a few key facts that we will use later:

(a) The constraints being minimized over in $C(n)$ are symmetric in the β_i's. That is to say if any two β_i, β_j are swapped then the constraints remain the same.

(b) d must be $\geq k$ for any feasible repair.

(c) There is an optimal solution achieving $C(n)$ that has the sum over the β_i's in the minimum $\mathcal{R}$ set achieving the lower bound α.

Let an optimal solution be given and let d be the number of non-zero β_i's that it has.

Re-label these β_i's so that the corresponding p_i's are non-decreasing ($p_1 \leq \cdots \leq p_{n-1}$). We must then have that the β_i's are non-increasing, otherwise by (a) we could swap the values of some β_i and β_j so as to decrease the cost, but this is not possible as the solution is already optimal.

We will now show that an optimal solution exists with d' non-zero β_i's such that the smallest is $\geq \frac{\alpha}{d'-k+1}$. If such a solution exists then it must be quasi-uniform as $\beta_1 = \beta_2 = \cdots = \beta_{d'} = \frac{\alpha}{d'-k+1}$ satisfies the constraints and has cost $\leq$ to any other such solution.

Assume then that $\beta_d < \frac{\alpha}{d-k+1}$, as otherwise the given solution would already be quasi-uniform and we would be done.

With the β_i's ordered in non-increasing fashion it is easy to see the smallest set $\mathcal{R}$ for the constraints:

$$\underbrace{\beta_1, ..., \beta_{k-1}}_{\mathcal{R}^C}, \underbrace{\beta_k, \beta_{k+1}, ..., \beta_d, 0, 0, ..., 0}_{\mathcal{R}}$$

As $\beta_d < \frac{\alpha}{d-k+1}$, then at least one of $\beta_{i^*} \in \{\beta_k, ..., \beta_{d-1}\}$ must be $> \frac{\alpha}{d-k+1}$.

We also now observe another general fact:

(d) for any optimal solution we must have that $\beta_1 = \beta_2 = \cdots = \beta_k$ as β_k is the largest rate in the minimum constraint set and thus the largest necessary rate for feasibility.

Case 1: $\beta_1 = \beta_2 = \cdots = \beta_k = \cdots = \beta_{d-1} > \frac{\alpha}{d-k+1} > \beta_d$

Let $\frac{\alpha}{d-k+1} - \beta_d =: D > 0$. Then, since $\beta_1 = \cdots = \beta_{d-1} = \beta$ and the given solution is optimal, by (c) we get

$$\begin{aligned}(d-k)\beta + \beta_d &= \alpha \\ \Rightarrow \beta &= \frac{\alpha}{d-k+1} + \frac{D}{d-k}\end{aligned}$$

We will now compare the given optimal solution to two closely related quasi-uniform solutions and show that one must be at least as good.

Quasi-Uniform Solution #1: Number of non-zero repair rates is $d' = d - 1$

$$\begin{aligned}\beta &= \frac{\alpha}{d-k+1} + \frac{D+\beta_d}{d-k} \\ &= \frac{\alpha}{d-k+1} + \frac{\alpha}{(d-k+1)(d-k)} \\ &= \frac{(d-k+1)\alpha}{(d-k+1)(d-k)} \\ &= \frac{\alpha}{d-k}\end{aligned}$$

Change to objective function from given solution:

$$\Delta_1 := \big(\frac{\beta_d}{d-k}\big)\big(\sum_{i=1}^{d-1} p_i\big) - \beta_d \cdot p_d \tag{3.8}$$

Quasi-Uniform Solution #2: Number of non-zero repair rates is d

$$\beta = \frac{\alpha}{d-k+1}$$

Change to objective function from given solution:

$$\Delta_2 := D \cdot p_d - \big(\frac{D}{d-k}\big)\big(\sum_{i=1}^{d-1} p_i\big) \tag{3.9}$$

Now,

$$\begin{aligned}\Delta_1 &= \beta_d\Big(\frac{1}{d-k}\big(\sum_{i=1}^{d-1} p_i\big) - p_d\Big) = -\big(\frac{\beta_d}{D}\big)\Delta_2 \\ \Delta_2 &= D\Big(p_d - \frac{1}{d-k}\big(\sum_{i=1}^{d-1} p_i\big)\Big) = -\big(\frac{D}{\beta_d}\big)\Delta_1\end{aligned}$$

As both $\beta_d > 0$ and $D > 0$ we have that either Δ_1 or Δ_2 must be ≤ 0. Therefore, either Quasi-Uniform Solution #1 or #2 is at least as good as the given optimal solution $\Rightarrow$ in Case 1 an optimal quasi-uniform solution exists.

Case 2: $\exists$ a smallest $j \in \{k+1, ..., d-1\}$ such that $\beta_j < \beta_k$

In this case, we may shift some rate from β_d to β_j, without increasing β_j beyond the value of β_k, and we will still have a valid solution as the minimum set $\mathcal{R}$ will retain the same total value and set of β_i's.

If $p_j < p_d$, then this transfer will create a strictly better solution than the one given, which contradicts the optimality. Thus, it must be the case that $\beta_d \geq \frac{\alpha}{d-k+1}$ in the given solution $\Rightarrow$ the solution is quasi-uniform.

If $p_j = p_d$, this transfer will not strictly improve the solution, but we will be in the following situation

$$\begin{array}{ll} \overbrace{\beta_1 = \cdots = \beta_{k-1}}^{\mathcal{R}^C} & = \overbrace{\beta_k = \cdots = \beta_{j-1} > \beta_j \geq \cdots \geq \beta_d > 0, 0, ..., 0}^{\mathcal{R}} \\ p_1 \leq \cdots \leq p_{k-1} & \leq p_k \leq \cdots \leq p_{j-1} \leq p_j = \cdots = p_d \leq \cdots \leq p_{n-1} \end{array}$$

Here, shifting rate around within $\beta_j, ..., \beta_d$ will not affect the value of the cost function and the set of repair rates are guaranteed to remain feasible as long as none of $\beta_j, ..., \beta_d$ are made $> \beta_k$.

So, we will shift rate within this set to the left so that

$$\beta_j = \beta_{j+1} = \cdots = \beta_{j+l} = \beta_k, \quad \beta_{j+l+1} < \beta_k \text{ and } \beta_{j+l+2} = \cdots = \beta_d = 0$$

As the cost function has not changed, this is still an optimal solution, but with new $d' = j + l + 1 \, (\leq d)$ and now of the form handled by Case 1. Therefore, a quasi-uniform optimal solution exists.

□

This theorem shows that for any repair the cost can be minimized by a quasi-uniform repair. Moreover, even without the assumption of our conjecture, the definition of $C(i)$ is a minimization over a condition we showed was necessary in Lemma 1 and thus $C(i)$ is a lower bound on the minimum possible functional repair cost. As the form of the optimal quasi-uniform solution we get from the proof of the theorem is shown to be feasible in [3] for any number of repairs, it follows that we have proven that for any individual functional repair, at any point in the evolution of the system, a quasi-uniform repair where we have the freedom to choose the parameter d will minimize the cost.

3.4 Minimum Costs for Reconstruction and Flexible Reconstruction

Minimizing reconstruction cost is a very simple problem that we will now cover. With our condition of complete download from any k nodes being required and sufficient for reconstruction it is clear that to minimize the cost of reconstruction a DC should connect to the k storage nodes from which it has least cost per bit to download. With $p_{i,\mathrm{DC}}$ being the cost of using link (i, DC), the minimum reconstruction cost is then

$$\min_{\mathcal{D}\subseteq[n],|\mathcal{D}|=k} \alpha \cdot \sum_{i\in\mathcal{D}} p_{i,\mathrm{DC}}$$

Similarly, for the case of flexible reconstruction presented in [5], a greedy cost minimization is also optimal. As the reconstruction requirement is to download μ_i bits from each node i such that $\sum_{i=1}^{n} \mu_i \geq M$ then downloading the maximum amount through the lowest cost links will minimize reconstruction cost. If we let $(p_{(1),\mathrm{DC}}, ..., p_{(n),\mathrm{DC}})$ be a non-decreasing ordering of the set of costs $\{p_{i,\mathrm{DC}}\}$ then this minimum cost is

$$\alpha \cdot \sum_{i=1}^{\lfloor \frac{M}{\alpha} \rfloor} p_{(i),\mathrm{DC}} + \mathrm{frac}\big(\frac{M}{\alpha}\big) \cdot p_{(\lfloor \frac{M}{\alpha} \rfloor+1),\mathrm{DC}}$$

Also, the associated repair rate requirement in this flexible framework for repairing node j is of the form

$$\sum_{i=1(i\neq j)}^{n} \beta_{i,j} \geq \gamma, \quad \text{with } 0 \leq \beta_{i,j} \leq \beta_{max}$$

and so again a greedy cost minimization is optimal resulting in a repair cost of

$$\beta_{max} \cdot \sum_{i=1}^{\lfloor \frac{\gamma}{\beta_{max}} \rfloor} \beta_{(i),j} + \mathrm{frac}\big(\frac{\gamma}{\beta_{max}}\big) \cdot p_{(\lfloor \frac{\gamma}{\beta_{max}} \rfloor+1),j}$$

Chapter 4

Varying the Capacity of Storage Nodes

In this chapter, we will consider the repair model of Chapter 3, but generalize it further by allowing storage nodes to have varying storage capacity. This extended model broadens the flexibility of the distributed storage network even further to cover a greater number of real systems. We characterize the achievable repair rates for up to 2 repairs and explore repair cost minimization for this model numerically.

4.1 Characterizing Repair Rates

We will now let α_i denote the available storage in node i.

We first require that $\alpha_i \geq \frac{M}{k}, \ \forall\, i$ as this is the standard minimum storage assumption and ensures that any k nodes contain $\geq M$ bits for reconstruction. Also, we make the reasonable assumption that repaired nodes will have the same storage capacity as the originals they are replacing. We will now establish results on the achievable repair rates for these networks.

Lemma 3: For functional repair of a node i, we have the following necessary condition for repair rates $\{\beta_{j,i}\}$ communicated from the remaining nodes $j \in [n]\backslash\{i\}$:

$$\min_{\substack{\mathcal{D}\subseteq[n]\backslash\{i\},\\ |\mathcal{D}|=k-1}} \Big\{\sum_{l\in\mathcal{D}} \alpha_l + \sum_{m\in\mathcal{D}^C} \beta_{m,i}\Big\} \geq M \tag{L3}$$

and this condition is also achievable using network codes for a single repair.

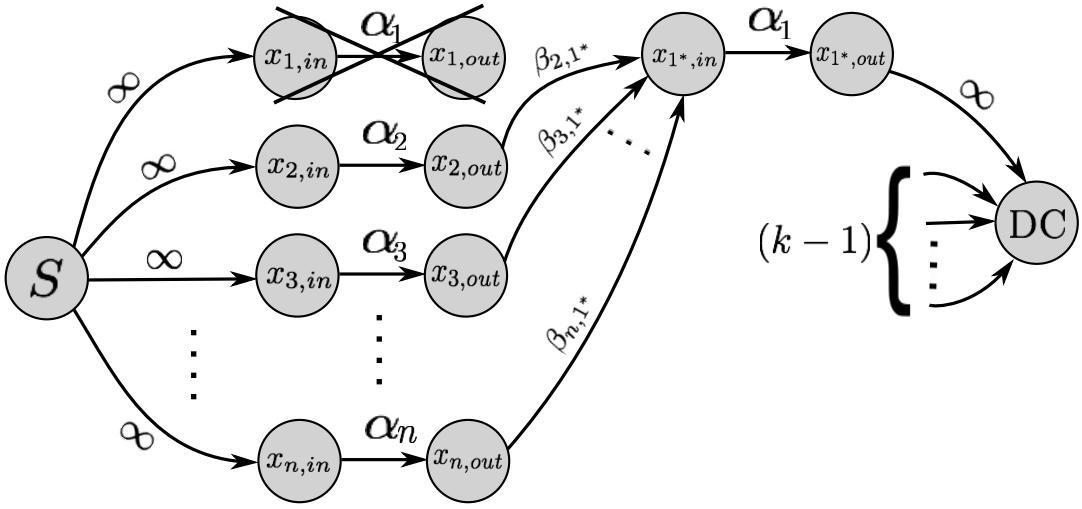

Figure 4.1: Information flow graph for 1 repair with varying storage capacity nodes.

Proof of Lemma 3:

For a single repair, let node 1 fail and be replaced by repaired node 1^* and let the DC connect to $\{1^*\} \cup \mathcal{D}$ where we will minimize over all possible sets $\mathcal{D} \subseteq \{2, ..., n\}, |\mathcal{D}| = k-1$.

Letting $\mathcal{S}$ be a cut set, for each $j \in \mathcal{D}$ we must have $x_{j,out} \in \mathcal{S}^C$ for a minimum cut, each thus providing α_j. As $\beta_{i,1^*} \leq \alpha_i \ \forall i$, if $x_{1^*,in} \in \mathcal{S}^C$ then the nodes in $\mathcal{D}^C$ will all contribute $\beta_{i,1^*}$, otherwise if $x_{1^*,in} \in \mathcal{S}$ we get α_1.

Thus, we require

$$\min_{\substack{\mathcal{D}\subseteq\{2,\dots,n\},\\ |\mathcal{D}|=k-1}} \Big\{ \underbrace{\sum_{i\in\mathcal{D}} \alpha_i + \alpha_1}_{\text{always } \geq M},\ \sum_{i\in\mathcal{D}} \alpha_i + \sum_{j\in\mathcal{D}^C} \beta_{j,1^*} \Big\} \geq M$$

which leaves us the contraint:

$$\min_{\substack{\mathcal{D}\subseteq\{2,\dots,n\},\\ |\mathcal{D}|=k-1}} \Big\{ \sum_{i\in\mathcal{D}} \alpha_i + \sum_{j\in\mathcal{D}^C} \beta_{j,1^*} \Big\} \geq M$$

The general necessity of this constraint follows by the same argument as was presented in the proof of Lemma 1, where we consider only cuts with $x_{i,in} \in \mathcal{S}$ for all active nodes i.

□

4.2 Single Repair Cost Minimization

We now have a necessary condition for functional repair in this scenario that is also achievable for a single repair.

So let's now define the following cost minimization for repair that we get by using (L3).

$$C'(i) := \min_{\substack{\sum_{l\in\mathcal{D}} \alpha_l + \sum_{m\in\mathcal{D}^C} \beta_{m,i} \geq M\\ \forall \mathcal{D}\subseteq[n]\setminus\{i\}, |\mathcal{D}|=k-1}} \sum_{j\in[n]\setminus\{i\}} p_{j,i}\beta_{j,i}$$

This function satisfies the following properties, by Lemma 3:

i) $C'(i)$ is the minimum possible functional repair cost for node i in a system allowing a single repair

ii) $C'(i)$ is a lower bound on the minimum repair cost of node i for a system allowing any number of repairs

This function is again a linear optimization problem that we can solve using linear programming for any chosen parameter values, so let us consider a few examples:

Let $M = 12, n = 10, k = 3$

- $\alpha[i] = [8, 5, 6, 5, 5, 7, 6, 6, 4], p[i] = [7, 10, 6, 2, 4, 6, 5, 1, 8]$:
  $C' = 8, \beta[i] = [0, 0, 0, 1, 1, 0, 0, 2, 0]$

- $\alpha[i] = [6, 8, 9, 10, 10, 6, 8, 7, 6], p[i] =$ anything :
  $C' = 0, \beta[i] = 0$
  Note that every $(k-1)$ $\alpha_i's$ sum to $\geq M$

- $\alpha[i] = [5, 5, 6, 5, 7, 10, 5, 4, 8], p[i] = [8, 10, 3, 1, 1, 3, 5, 5, 10]$:
  $C' = 4, \beta[i] = [0, 0, 0, 1, 3, 0, 0, 0, 0]$

We find in this case, where the storage capacities may vary, that the optimal repair solution is no longer always quasi-uniform. Thus, a general means of repair is required to achieve minimum repair cost in a system with individual freedom in node storage capacities.

4.3 Multiple Repair Rate Characterization

In Lemma 3 we found a necessary condition on achievable functional repair rates, but we would like to know what conditions are sufficient when allowing more than a single repair, as would be required of any practical system. The following lemma covers a second repair and shows the dependence between repairs which exists making a general characterization difficult.

Lemma 4: For functional repair of two node failures, say i first then j being replaced by i^* and j^* respectively, necessary and sufficient conditions for achievable repair rates are (L3) for both the first and second repair and the following additional condition on the second:

$$\beta_{j,i^*} + \min_{\substack{\mathcal{D}\subseteq[n]\setminus\{i,j\},\\ |\mathcal{D}|=k-2}} \Big\{ \sum_{l\in\mathcal{D}} \alpha_l + \sum_{m\in\mathcal{D}^C} \min(\alpha_m, \beta_{m,i^*} + \beta_{m,j^*}) \Big\} \geq M \qquad \text{(L4)}$$

Proof of Lemma 4:

WLOG, let node 1 be the first failure and node 2 be the second. As before, let the new nodes be 1^* and 2^*, respectively. We must check for the conditions that ensure every cut on the graph is $\geq M$.

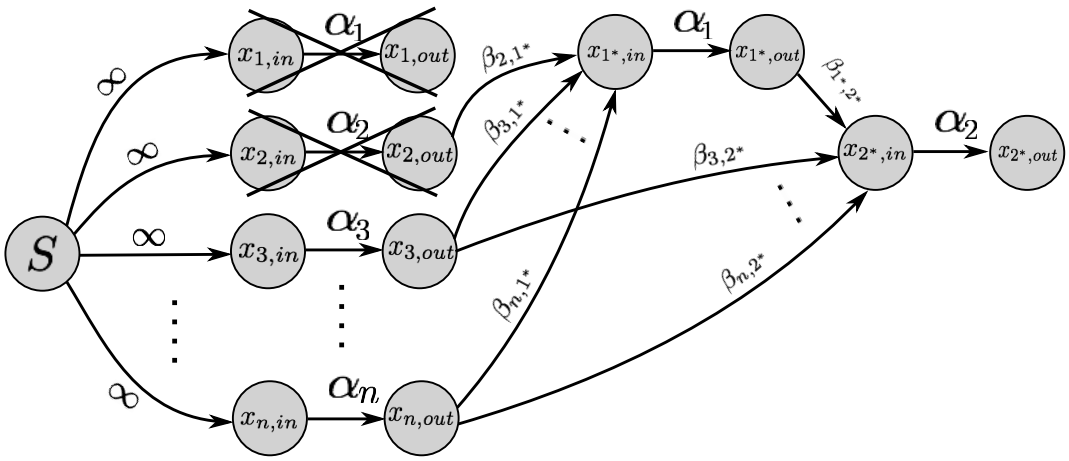

Figure 4.2: Information flow graph for 2 repairs with varying storage capacity nodes.

From Lemma 3 we have that condition (L3) is necessary and sufficient for repair of the first node failure. So, for node 1^* we have

$$\min_{\substack{\mathcal{D}\subseteq[n]\setminus\{1\},\\ |\mathcal{D}|=k-1}} \Big\{ \sum_{l\in\mathcal{D}} \alpha_l + \sum_{m\in\mathcal{D}^C} \beta_{m,1^*} \Big\} \geq M$$

As before, let the set $\mathcal{S}$ be a graph cut where the source S must be in $\mathcal{S}$ and $DC \in \mathcal{S}^C$ and notice that for a min cut we must have for each of the nodes, l, connected to the DC that $x_{l,out} \in \mathcal{S}^C$ as $x_{l,out}$ connects to DC with infinite capacity.

Case 1: DC connects to 2^*, but not 1^*

Constraints:

$x_{2^*,in} \in \mathcal{S}$:

$$\alpha_2 + \min_{\substack{\mathcal{D}\subseteq\{3,..,n\},\\ |\mathcal{D}|=k-1}} \sum_{i\in\mathcal{D}} \alpha_i \geq M \tag{4.1}$$

$x_{2^*,in} \in \mathcal{S}^C, x_{1^*,out} \in \mathcal{S}$:

$$\beta_{1^*,2^*} + \min_{\substack{\mathcal{D}\subseteq\{3,..,n\},\\ |\mathcal{D}|=k-1}} \Big\{ \sum_{i\in\mathcal{D}} \alpha_i + \sum_{j\in\mathcal{D}^C} \beta_{j,2^*} \Big\} \geq M \tag{4.2}$$

$x_{2^*,in} \in \mathcal{S}^C, x_{1^*,out} \in \mathcal{S}^C, x_{1^*,in} \in \mathcal{S}$:

$$\alpha_1 + \min_{\substack{\mathcal{D}\subseteq\{3,..,n\},\\ |\mathcal{D}|=k-1}} \Big\{ \sum_{i\in\mathcal{D}} \alpha_i + \sum_{j\in\mathcal{D}^C} \beta_{j,2^*} \Big\} \geq M \tag{4.3}$$

$x_{2^*,in} \in \mathcal{S}^C, x_{1^*,out} \in \mathcal{S}^C, x_{1^*,in} \in \mathcal{S}^C$:

$$\min_{\substack{\mathcal{D}\subseteq\{3,...,n\},\\ |\mathcal{D}|=k-1}} \Big\{ \sum_{i\in\mathcal{D}} \alpha_i + \sum_{j\in\mathcal{D}^C} \min(\alpha_j, \beta_{j,1^*} + \beta_{j,2^*}) \Big\} \geq M \tag{4.4}$$

Case 2: DC connects to 2^* and 1^*

Constraints:

$x_{2^*,in} \in \mathcal{S}, x_{1^*,in} \in \mathcal{S}^C$:

$$\alpha_2 + \beta_{2,1^*} + \min_{\substack{\mathcal{D}\subseteq\{3,..,n\},\\ |\mathcal{D}|=k-2}} \Big\{ \sum_{i\in\mathcal{D}} \alpha_i + \sum_{j\in\mathcal{D}^C} \beta_{j,1^*} \Big\} \geq M \tag{4.5}$$

$x_{2^*,in} \in \mathcal{S}^C, x_{1^*,in} \in \mathcal{S}$:

$$\alpha_1 + \min_{\substack{\mathcal{D}\subseteq\{3,..,n\},\\ |\mathcal{D}|=k-2}} \Big\{ \sum_{i\in\mathcal{D}} \alpha_i + \sum_{j\in\mathcal{D}^C} \beta_{j,2^*} \Big\} \geq M \tag{4.6}$$

$x_{2^*,in} \in \mathcal{S}^C, x_{1^*,in} \in \mathcal{S}^C$:

$$\beta_{2,1^*} + \min_{\substack{\mathcal{D}\subseteq\{3,...,n\},\\ |\mathcal{D}|=k-2}} \Big\{ \sum_{i\in\mathcal{D}} \alpha_i + \sum_{j\in\mathcal{D}^C} \min(\alpha_j, \beta_{j,1^*} + \beta_{j,2^*}) \Big\} \geq M \tag{4.7}$$

These constraints can be reduced in the following ways:

- Contraints (4.1) and (4.3) are satisfied as $\alpha_i \geq \frac{M}{k}$
- (4.4) and (4.5) follow from (L3) for node 1^*
- (4.2) and (4.6) combined are equivalent to (L3) for node 2^*

Leaving constraint (4.7) $\equiv$ (L4).

□

Recall that in the case of non-varying α we found that the additional constraints were made redundant at the minimum storage point of $\alpha = \frac{M}{k}$. We find here that any constraint of the form $\alpha_i \leq B,\ \forall i$, other than the minimum storage point constraint of $B = \frac{M}{k}$ as explored earlier, fails to make constraint (L4) redundant in the second repair. We show this with the following example, a slight modification of Example 1.

Example 3:

Let any upper bound $B > \frac{M}{k}$ be given.

Let $\alpha_i = \frac{M}{k} + \frac{M}{(k(n-k)+k-1)k} + \epsilon,\ \forall i$ and let $(n-k)$ be large enough and ϵ sufficiently small such that $\alpha_i \leq B$.

By our choice for the α_i's, we now have the same non-varying storage capacity as in Example 1.

Then, comparing the constraints we see (L3) $\equiv$ (L1) and (L4) $\equiv$ (3.7) and by the same calculations given in Example 1 we get that (L3) holds for both repairs, but (L4) is not satisfied.

We have found that for the model with varying storage capacities the necessary condition is again only sufficient for > 1 repair when we are at the minimum storage point. In contrast to the uniform storage capacity model, though, cost minimization over the necessary condition can no longer be achieved by quasi-uniform repair.

Chapter 5

Conclusion

We explored the minimization of repair cost in distributed storage systems with communication costs associated to links in the network. In order to do this we explored conditions on repair rates for achievable codes under a generalized functional repair model. Allowing the amount of information downloaded from different nodes to vary and minimizing the cost for a repair we found that the quasi-uniform repair method of downloading the same quantity from each of d nodes attains the minimum cost when storing the minimum amount of information at each node. This result applies to any of the functional repairs in the lifetime of the system.

We then generalized the model further by allowing the storage capacities of the nodes to vary and found a general necessary condition for achievable repair rates as well as sufficient conditions for the first and second repairs in the system. We then established that quasi-uniform repairs are not optimal for cost minimization in this case.

Bibliography

[1] A. Dimakis, K. Ramchandran, Y. Wu, C. Suh, "A Survey on Network Codes for Distributed Storage," *Proceedings of the IEEE*, vol. 99, no. 3, pp. 476-489, March 2011.

[2] Z. Wang, R. Mateescu, A.G. Dimakis, J. Bruck, "Array codes for distributed storage: Results and open problems," *Information Theory and Applications (ITA)*, 2010.

[3] A. G. Dimakis, P. G. Godfrey, Y. Wu, M. J. Wainwright, K. Ramchandran, "Network coding for distributed storage systems," *IEEE Transactions on Information Theory*, vol. 56, pp. 4539-4551, September 2010.

[4] Y. Wu, "A construction of systematic MDS codes with minimum repair bandwidth," *IEEE Trans. Inf. Theory. [Online]*, August 2009. Available: http://arxiv.org/abs/0910.2486.

[5] N.B. Shah, K.V. Rashmi, P. Vijay Kumar, "A flexible class of regenerating codes for distributed storage," *Information Theory Proceedings (ISIT), 2010 IEEE International Symposium on*, pp. 1943-1947, 13-18 June 2010.

[6] S. Akhlaghi, A. Kiani, M. R. Ghanavati, "Cost-bandwidth tradeoff in distributed storage systems, arXiv:1004.0785v2 [cs.IT], 14 April 2010.

[7] M. Gerami, M. Xiao, M. Skoglund, "Optimal-cost repair in multi-hop distributed storage systems," *Information Theory Proceedings (ISIT), 2011 IEEE International Symposium on*, (To Appear).