Introduction

Data Availability Sampling (DAS) is a fundamental mechanism in Celestia, allowing light nodes to verify block data availability without downloading entire blocks. Although the theoretical underpinnings were established in the original research paper by @musalbas [1809.09044] Fraud and Data Availability Proofs: Maximising Light Client Security and Scaling Blockchains with Dishonest Majorities, we conducted reconstruction simulations

that show the real-world requirements for light nodes can be significantly lower than these initial estimates.

Our simulation code is publicly available in thesimulations GitHub repository.

Original Theory vs New Findings

The data availability scheme uses a 2D Reed-Solomon encoded matrix where k×k

original data is extended to a 2k×2k

matrix. The original paper analyzes the number of light nodes (LN) required for reconstruction. The algorithm assumes that for data to be unrecoverable, an adversary must withhold at least $(k+1)^2$

shares, leading to the formula k(3k-2)

for minimum required shares for reconstruction.

This case can be visualized as shown in the left picture, where red shares are assumed to be unavailable and white shares are available. While this represents the worst case for reconstruction, we've implemented simulation that allows to account a possible cases of complete reconstruction with fewer samples than in previous base assumption. The example on the right picture demonstrates this: white shares are required for reconstruction, while green shares are not necessary.

[

1368×666 15.1 KB

](https://forum.celestia.org/uploads/default/original/2X/2/2486d555c32af8443a84728bcda39f988a571b1f.png)

The original paper suggests that reconstruction should be possible when all light nodes collectively gather all white samples (as shown in the left picture). While true, this represents the worst case scenario. Many other permutations with fewer samples than k(3k-2)

can result in successful reconstruction. While a theoretical formula for defining the probability of reconstruction with a fixed number of samples and light nodes remains an open question, we can empirically analyze reconstruction probability through Monte Carlo simulations.

Methodology

We conducted Monte Carlo simulations with 1,000 iterations per configuration to determine the number of light nodes needed for 99% reconstruction probability. Each simulation:

1. Randomly distribute N samples across M light nodes
2. Attempt complete reconstruction
3. Record a binary success/failure outcome

Results and Analysis

Comparison with Previous Estimates

[

1600×1023 132 KB

](https://forum.celestia.org/uploads/default/original/2X/9/9981074e06defa95a21b79d213f2588a374ebaf7.png)

Our findings consistently show that we need approximately 2.5-2.8x fewer light nodes than initially estimated in the whitepaper. This ratio remains relatively stable across different sampling rates (s) for the same matrix size (k).

[

1600×319 59.4 KB

](https://forum.celestia.org/uploads/default/original/2X/1/151b44890669f02586f4211804428e2c165e17fd.png)

Light Node Requirements Analysis

To understand how these improvements translate to practical network deployments, we analyzed reconstruction requirements across different block sizes with preset sample amounts for each node. The relationship between block size and required light nodes follows a clear linear pattern, as demonstrated in log scale graphs.

[

1600×535 99 KB

](https://forum.celestia.org/uploads/default/original/2X/c/cb41b04b55e2ad1a9b1be4d7fbac2b3a49fc9e1b.png)

Key findings:

1. Sample Size vs Node Count

: Doubling the samples per light node reduces the required number of nodes by half, showing an inversely proportional relationship.

1. Block Size Impact

: When the original data size (k) doubles, the network needs 4x more light nodes to maintain the same reconstruction probability.

1. Linear Correlation

: There is a linear correlation between block size in MB and the required number of nodes with a fixed amount of samples. The ratio of required light nodes per MB remains constant regardless of block size and only depends on the samples per node.

Network Scaling Dynamics

One of the most interesting findings comes from analyzing networks with fixed amounts of light nodes. We simulated how many samples are required by each node for reconstruction to be possible.

[

1600×608 97.3 KB

](https://forum.celestia.org/uploads/default/original/2X/3/369c0080e6b9ffaafa27e087e4b093fc1f4e3fed.png)

This analysis reveals several important relationships:

1. Networks with more light nodes can maintain data availability with fewer samples per node. The relationship is inversely proportional - 4x network size reduces required samples proportionally.
2. Larger networks (>64k nodes) can maintain data availability with minimal sampling even for large blocks.
3. Sample Distribution Pattern (Diagonal Relationship): We found consistent sample requirements across different network configurations. For example, 40 samples per node enables reconstruction in these scenarios:
4. 256 light nodes with 2MB blocks
5. 1,024 light nodes with 8MB blocks
6. 4,096 light nodes with 32MB blocks
7. 16,384 light nodes with 128MB blocks
8. 256 light nodes with 2MB blocks
9. 1,024 light nodes with 8MB blocks
10. 4,096 light nodes with 32MB blocks
11. 16,384 light nodes with 128MB blocks

Relationship between the square size k and the number of samples s

The numbers found by our new simulation align well with the formula initially suggested in the blogpost (Increasing blocksize with more samples instead of light nodes) by @nashqueue:

$(c*s) / (2k)^2 = R$

Where:

- c = required amount of light nodes for reconstruction
- k = block size
- s = amount of samples each light node performs per block

The previous assumption showed that R = 1.37

; however, new approaches allowed us to determine a more precise number: $R \approx 0.6$

Practical Implications

The current Celestia implementation uses a fixed 16-sample requirement per light node. Using the provided heatmap, we can project how the light node count needs to grow with increasing block size to maintain security guarantees of possible reconstruction.

Variable Sampling Rates

Our findings suggest potential optimization opportunities. There's a clear trade-off between network size and sampling requirements. Smaller networks can compensate by increasing samples per node. Nodes could adjust their sampling based on:

- Current network size
- Block size
- Desired security level

Recommendations for Celestia

Based on these findings, we suggest to consider:

1. (Optional) Implementing dynamic sampling rates that adjust based on network conditions
2. Updating the current fixed 16-sample requirement to better match network size
3. Improving mechanisms to track light node count in the network
4. Considering block size limits based on active light node count

Conclusion

Our analysis demonstrates that Celestia's data availability scheme can operate effectively with significantly fewer resources than originally estimated. This opens up new possibilities for scaling while maintaining security guarantees.

We've validated a key relationship between network parameters:

$(c*s) / (2k)^2 = R$

Where:

c = number of light nodes

s = samples per node

k = block size

This refinement from the original estimate of C = 1.37

to C = 0.6

means the network needs less than half the originally estimated resources.

The findings suggest that as Celestia's network grows, it can support larger blocks with fewer samples per node, providing a natural scaling mechanism. However, careful consideration must be given to the relationship between block size, light node count, and samples per node to maintain reliable data availability verification.

Acknowledgments

Special thanks to @nashqueue for their invaluable collaboration in reviewing and engaging in many constructive discussions on this topic.