Using Monte-Carlo Simulations to Estimate Reliability of Fault-Tolerant Storage System Designs

Dylan Tocci
ECE 544 - Fall 2021
University of Massachusetts Dartmouth
North Dartmouth, United States
dtocci1@umassd.edu

***Abstract*—Reliability is a large factor when designing large scale storage systems. The ability for a user to safely store and access their data online is essential, and thus designs must be chosen to meet these requirements. Two main methods of providing reliability in storage systems is by using triple modular redundancy (TMR) or a form of standby sparing. TMR requires three copies of the component, where two of the three must be functional. The output of these components feed into a voter which will take the majority output as correct and feed this data out. In contrast, standby sparing only requires a single spare. This spare can be hot, cold or warm. Hot implies the spare is always active, cold means that the spare will only activate once the primary module has failed, and warm serves as a middle-ground between these two types. This paper examines through Monte-Carlo simulations the effectiveness of TMR versus dynamic standby sparing, to see how the two fair against each other assuming the same components. By treating the failure rate of the hard drives as a random variable, simulations can be ran taking into account the variation in component quality. After simulations, the paper demonstrates how dynamic standby sparing was more effective despite the use of less components.**

***Index Terms*—Monte-Carlo; TMR; standy sparing**

I. Introduction

When designing systems that need to be accessed continually, reliability is a large factor that needs to be considered. Typically, the method for deciding what reliability measures to take depends on the financial state of the project. However, assuming there is sufficient funding, it can be difficult to say concretely why one design might be better than another. This is where using Monte-Carlo simulations becomes necessary. Monte-Carlo simulations are a method of simulating systems and events where a random variable exists. In the case of system reliability, the random variable would be the reliability or failure rate of the component. This can be done by generating random numbers based on a Gaussian distribution. This provides realistic variations in the components used, where some will be slightly less relaible and some more reliable. Using Monte-Carlo simulations, one can generate determine how the system would operate over thousands of trials to numerically see how effective the system is.

In terms of designing reliable systems, two main designs are applicable. These are dynamic standby sparing and triple modular redundancy (TMR) designs. Dynamic standby sparing implies that you have a primary component with a spare in a dynamic state. This means it could be hot (always running), cold (runs only after primary failure), or a middle-ground state known as warm sparing. In the case of this report, hot sparing is used. The main benefit of using a hot spare is that it can immediately be used upon primary failure. Thus, there is less downtime after a primary failure occurs. However, as a result of this the secondary component may fail at any time - including before the primary even fails. Thus, there is a trade-off between downtime upon primary failure, and the risk of the backup component failing. An alternative would be to use cold sparing components. Yet this comes with a trade-off as well as the cold spare needs time to turn on when the primary component fails. Depending on the system used, this down time may be unacceptable.

An alternative to dynamic standby sparing designs are triple modular redundancy designs. As the name implies, this method requires three copies of the primary component. These three components all feed into a voter which will output the majority of the input data. Thus, at least two of the three components need to be working in order for the system to be functioning. Additionally, this voter component needs to be active. This design is typically more expensive as it requires more components than using a single spare. A downside to this design is the single point of failure found by using a voter. This can significantly impact the reliability of the system, which will be demonstrated later on.

II. Previous Research

While Monte-Carlo simulations are a popular tool in the business and marketing field, less work has been done when it comes to applying Monte-Carlo simulations to system reliability. However, a few such cases have been studied. One example was a paper where Monte-Carlo simulations were used to examine the reliability of Distribution Generators (DG) connected Distribution Centers [3]. This allowed for a simulation that would address line capacity limit, and time-varying loads. The simulation would start with a sample year, and component faults would then be generated each time step based on a Gaussian distribution. The simulation would then check if these component faults would affect a customer at a given time point. Through their simulations, data could be derived demonstrating how impactful Distribution Generators were for system reliability.

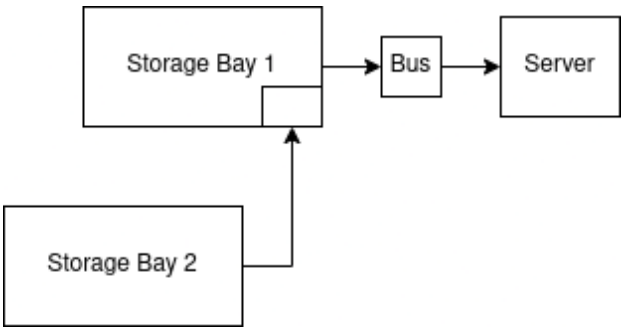

Fig. 1. The dynamic standby sparing system design

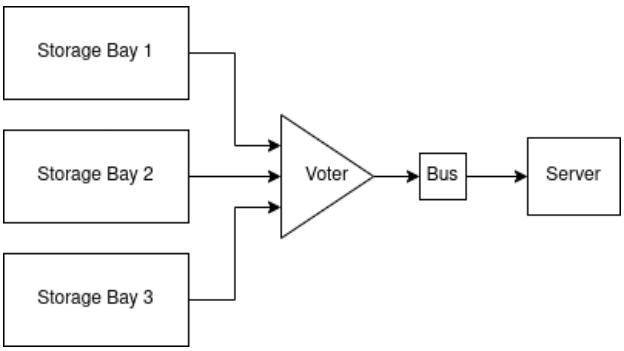

Fig. 2. The triple modular redundancy system design

A second use case in the field of reliability was seen in examining the reliability of emergency and standby power systems [2]. In this example, sequential sampling is used as a method of determining a components state. Thus, the system will generate a sequence of events using probability distributions of the random variables with some state duration. This was done to determine the time to failure (TTF), as well as the time to repair (TTR). The results of these simulations aimed to demonstrate the benefit of using emergency and standby power components in systems where critical components are sensitive to power fluctuations.

III. Testing Methodologies

For comparing dynamic standby sparing to TMR designed systems, two theoretical systems were designed for simulations. The basis of these designs was a simple server to storage system, where a single bus moved data between the storage and server interface. The design for dynamic standby sparing can be seen in Figure 1.

Figure 2 shows the design for triple modular redundancy. As can be seen, both designs use a single bus and server to better isolate the differences in the reliability design. The most notable difference between the two designs is the voter system used in TMR. This voter is considered to be a vital component, which will cause the system to fail if it fails. Otherwise, it will require two of the three storage bay components to be working for a functional system. In contract, the dynamic standby sparing design only requires one of the two storage bays to be working, without the need for a voter. It should be noted that this design assumes a hot sparing principle, where the spare is always running. As a result, there are cases where the spare may fail before the primary fails. This can result in system failure if the primary fails which will be discussed further in the Results.

For simulation, the failure rate of the storage components was the random variable. The failure rate was randomly chosen from a Gaussian distribution at 0.000178 failures per day with a Mu value of 0.00025. This data was chosen based off of failure rate statistics from BackBlaze [1]. This simulated choosing more or less reliable hard drives for each system. The other components all shared a reliability of 0.00001 statically as no realistic reliability data for these components could be found. This was kept at a low value to help the storage system design have more impact on system reliability. Keeping these components failure rates constant allowed for more focus to be put on the reliability of the storage designs themselves, rather than the entire system. The failure rates were exponential to simulate the degradation of components over time. Additionally, all components shared an exponential repair rate of 0.005. This would simulate delays in receiving replacements, or errors in repairing.

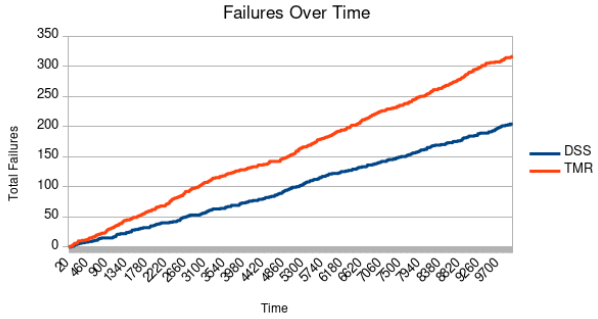

Fig. 3. Total system failures over time

To simulate the system, the simulation would run in time steps representing each day. 10,000 time steps were run per simulation, and 1000 simulations were run total. During each time step, the system would determine if a component would fail based off its Gaussian failure rate and up-time, and change the components state. If a component was in its failure state, the system would also determine if it would be repaired based off the repair rate and down-time. After these component failures and repairs occurred, the simulation would then check if the storage system was functioning, and if the entire system was functioning. Data was drawn for system and storage reliability, in addition to failure and repair rate statistics.

IV. Test Results

A. Failures and Repairs Over Time

The first statistic monitored in these simulations were the systems total failures and repairs over time. This would help demonstrate how likely components were to fail in reality, which gives background for future data. As can be seen

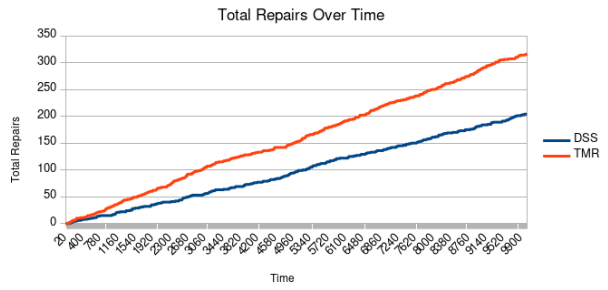

Fig. 4. Total system repairs over time

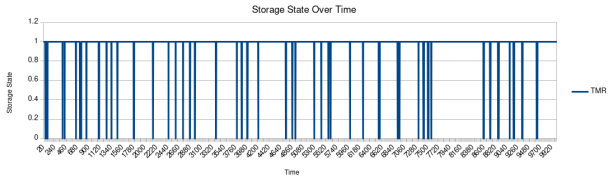

Fig. 5. TMR storage state over time

by referencing Figures 3 and 4, the TMR system design has significantly more failures over time when compared to dynamic standby sparing (labeled "DSS"). The TMR design had 316 total component failures, while the standby sparing design only had 204. This is because the TMR design requires an extra storage component, in addition to a voter component. As a result, it suffers from more failures and as a result has a higher amount of repairs. The amount of the repairs done was approximately equal to the failures that occurred in each design. These repairs affected the storage and system reliability which is examined later on. It should be noted that the graphs displaying data relative to time in this paper were from a single sample of the simulation. There will be averaged statistics later on over to help remove outliers which these graphs may present.

B. Storage State Over Time

Figures 5 and 6 show the storage system's state over time for each system. This represents how available each storage system was over time. A value of 1 represents the storage system functioning, while 0 represents the storage system not functioning. For dynamic standby sparing, only the primary or spare needed to be functional for the storage system to be

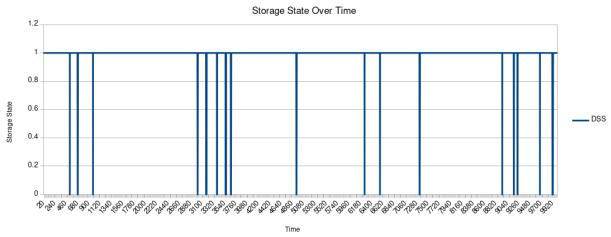

Fig. 6. Dynamic standby sparing storage state over time

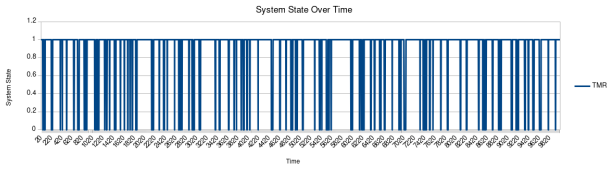

Fig. 7. TMR system state over time

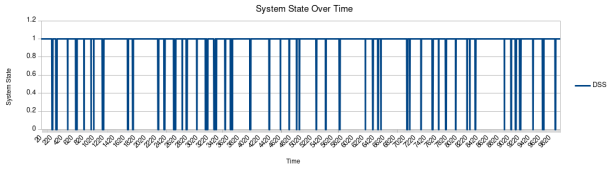

Fig. 8. Dynamic standby sparing system state over time

accessible. However, TMR required two of the three storage bays needed to be active, in addition to the voter. As a result, TMR had much less availability, and was frequently down. While TMR didn't have long down-times, they occurred very frequently which makes the system nearly as unreliable.

C. System State Over Time

System reliability was another statistic monitored. This shows how reliable the entire system was, including the bus and server components. By viewing Figures 7 and 8, the two appear more comparable. However, TMR still falls behind standby sparing in terms of total up time. Again, this is likely due to the storage system and the additional components it requires.

D. Overall System Performance

Figures 9 and 10 show the average system and storage subsystem downtime over 1000 simulations. This helps remove outliers in performance that may rarely occur from the Gaussian distribution. Additionally, it helps provide more context to the graphs shown in previous sections. In terms of storage downtime, standby sparing is significantly more efficient than the TMR design. The standby sparing design's storage system averaged 7% downtime each simulation, while

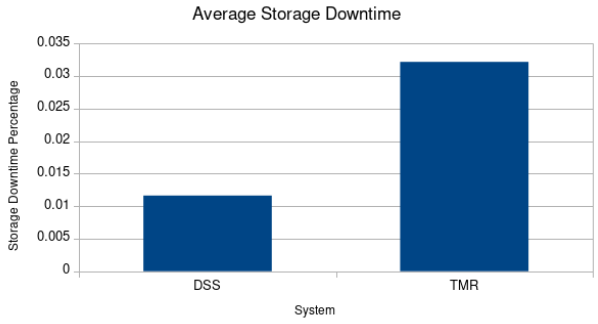

Fig. 9. Average Storage Subsystem Downtime

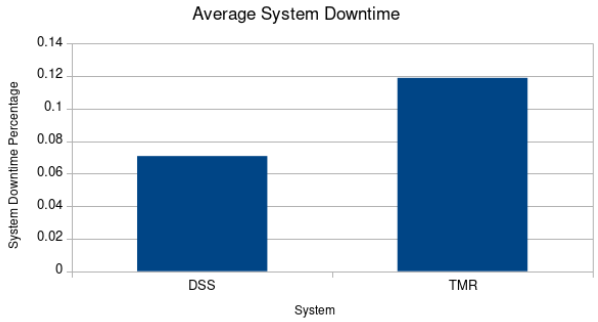

Fig. 10. Average System Downtime

the TMR design's storage system had 11% downtime each simulation. Additionally, this gap can be seen when comparing the system reliability of standby sparing and TMR. The standby sparing design is available approximately 20% more than TMR. The average system downtime for the TMR design is 32%, while the standby sparing design had an average system downtime of only 12%. This difference is likely due to the storage system design, as this is data averaged over 1000 simulations. Thus, any variation in the other components of the system should roughly equal out in both systems.

V. Conclusion

Monte-Carlo simulations provide a numerical method for analyzing systems in a more realistic context. By using Gaussian distributions to represent a random variable, the simulation can account for components and their replacements being better or worse than before - resulting in varying failure rates. Additionally, one can simulate with these components for thousands of years if desired, providing very long term statistics. When it comes to comparing TMR and standby sparing, it is apparent that standby sparing appears to be the more financial and reliable option. It had higher up time, less failures, and less components than TMR. This likely occurred as the TMR design had an additional single point of failure, which would bring the entire system down upon failure. Using Monte-Carlo simulations allowed for visualization of this flaw in the design, and provided numerical data to support it's effects.

VI. Future Work

For future work, simulating different standby sparing methods against each other may provide insights as to where each has its strengths. For example, one could determine if the downtime of having a cold spare out weights the risk of a hot spare failing. Additionally, one could test different parameters for a warm spare to find the most ideal settings for the system being tested. Testing 5MR and higher modular redundancies could also provide data to show how effective adding in additional redundancies is in terms of system downtime and theoretical cost. Tests could also be done for triplicated voters. This would remove the single point of failure in the TMR design tested in this paper, which may have resulted in the majority of system and storage failures. Perhaps using this design would make TMR more reliable than standby sparing at a higher cost and design complexity.

References

[1] A. Klein, "Hard drive failure rates for Q1 2021," Backblaze Blog — Cloud Storage & Cloud Backup, 23-Jul-2021. [Online]. Available: https://www.backblaze.com/blog/backblaze-hard-drive-stats-q1-2021/. [Accessed: 26-Nov-2021].

[2] C. Singh and J. Mitra, "Monte Carlo simulation for reliability analysis of emergency and standby power systems," IAS '95. Conference Record of the 1995 IEEE Industry Applications Conference Thirtieth IAS Annual Meeting, 1995, pp. 2290-2295 vol.3, doi: 10.1109/IAS.1995.530594.

[3] F. Li and N. Sabir, "Monte Carlo Simulation to Evaluate the Reliability Improvement with DG connected to Distribution Systems" Power '08. 8th WSEAS International Conference on Electric Power Systems, High Voltages, Electric Machines, 2008, pp. 177-182

Appendix

Project Code and Raw Data: GitHub Link