A Case for Elastic Replication Information Services

Jose E. Torres Berrocal
Advisor: Bienvenido Velez

Electrical and Computer Engineering Department
University of Puerto Rico, Mayagüez Campus
Mayagüez, Puerto Rico 00681-5000
jetorres@ece.uprm.edu

Abstract

Elastic Replicated Information Services (ERIS) could be described as information services that by means of adaptive replication algorithms maintain a desired level of availability on clusters of workstations as the number of nodes changes. A multi-disk storage system simulator was programmed. The simulator included algorithms of replication of objects, based on a Bernoulli process to find the availability of the system while the nodes count increases. The simulation output data about system multi-disk system failures based on the MTBF of individual disks. This data is analyzed by graphical means and statistical formulations. The need for an ERIS is justified.

1. Introduction

Elastic Replicated Information Services (ERIS) could be described as information services that by means of adaptive replication algorithms maintain a desired level of availability on clusters of workstations as the number of nodes changes. These services arise because clusters grow in number of nodes as load increases, and the fact of adding nodes increases the expectation of the number of failures in a system. Since the nodes could be added at any time it also increases the need to automate the process of determining the changes required to compensate the consequent availability loss.

Replication is the process of storing multiple copies of objects on various nodes or discs. When a particular node is down for any reason, the data on that node is still available on the online nodes with the replicated data, so if the data is needed on the system, it can continue to offer service. The way or strategy in which the nodes replicate the data when a new node gets online are to store additional copies of the available objects or by migration (moving) of the objects without increasing the actual replication count of the objects. Notice that the replicated data has to be updated on each or considerably mayor group of replicas every time a write operation is done, so when a large count of write operations is done on the system, on a certain point the replication strategy is not acceptable, and the migration should be used instead. On this investigation will be demonstrated that migration alone should not be used. Leading to the need of a calculated balance between the two strategies, and this calculation is done by the ERIS.

In the investigation, the cluster of workstations is resemble as a group of discs, where the availability of a disc is measure as a function of the mean time to failure (MTTF) of the disc. In order to realize the experiments a simulator program is written.

2. Previous Work

2.1 Disadvantages of Current Competing Replication Algorithms

The replication problem has been treated in various ways, and we have found three different mayor types of methods to do replication. These methods are: Consensus Based [Özalp87][Gifford97], Data Trading [Cooper2002] [Stonebraker96], and RAID [Patterson88].

All these methods have advantages and disadvantages, depending on the situation used. In this paper we claim only disadvantages. In our opinion the disadvantages are the following. The Consensus Based requires extensive use of replication in order to reach to consensus taking much memory and processing resources. The Data Trading does not have enough control so does not warranty good replication for all nodes (you could finish getting nodes without replication and nodes with to many). The RAID as describe in RAID [Patterson88] all their parameters are constants and does not change according to the addition of nodes, giving under or over utilization of space. In all cases if it is used a fix replication percent, then, by our hypothesis, they also over estimates the reliability of the system.

2.2 Theoretical Background

This paper assumes that the reader knows and understands the basic definitions of the following subjects: Nondeterministic experiment, Random, and Probability.

Table 1. Calibration and System Confidence results used on Figure 1.

#obj.	disks	Avg. MTBF	Expected MTBF	%Error MTBF	System Confidence
100	1	194072	200000	2.964	0.999995
100	2	96688	100000	3.312	0.999990
100	3	60461	66666	9.307593	0.999985
100	4	49786	50000	0.428	0.999980
100	5	44196	39999	10.49276	0.999975
100	6	27837	33333	16.48816	0.999970
100	7	26832	28571	6.086591	0.999965
100	8	24680	24999	1.276051	0.999960
100	9	21764	22222	2.061021	0.999955
100	10	21178	19999	5.895295	0.999950
100	20	12205	9999	22.06221	0.999900
100	30	6271	6666	5.925593	0.999850
100	40	4900	4999	1.980396	0.999800
100	50	3426	3999	14.32858	0.999750
100	60	2808	3333	15.75158	0.999700
100	70	2298	2857	19.56598	0.999650
100	80	2485	2499	0.560224	0.999600
100	90	2174	2222	2.160216	0.999550
100	100	2001	1999	0.10005	0.999500

Currently accepted definitions and formulas needed to understand this investigation (obtain from [Drake88]):

1. Bernoulli trial - These could be described as a nondeterministic experiment which results in two possible independent outcomes. Each experiment is called a trial. The outcome of each trial depends on a predefined probability in such a way that its outcome could be seen as a success or a failure. Also each trial is independent of the outcome of the previous trial. A series of Bernoulli trials is known as a Bernoulli process. Some mathematical formulas related to the Bernoulli process are:

$$P[S_N = k] = \binom{N}{k} p^k (1-p)^{N-k} \qquad (1)$$

$for any k = 0,1,2,\ldots,N.$

$$\binom{N}{k} = \frac{N!}{[k!(N-k)!]},$$

where N is the number of succesive trials and k is the number of successful Bernoulli trials, p is the probability of success on each trial, and P is the probability of a particular Bernoulli process sequence.

3. Poisson process - this process defers from the Bernoulli process in which its outcomes describes the behavior of arrivals at points on a continuous line instead of discrete trials. Generally this line refers to a time axis. In a mathematical description, it could be considered to be the limit, as $\Delta t \to 0$ of a Bernoulli process, one trial every Δt, with the probability of a success on any trial given by $p = \lambda \Delta t$.

Some mathematical formulas related to the Poisson process are:

$$P(k) = \frac{(\lambda t)^k e^{-\lambda t}}{k!} \quad t \geq 0; k = 0,1,2,\ldots \qquad (2)$$

3. Theoretical and Calibration Curves

In order to program a Simulator to investigate the response in availability of a group of discs, we started by coding a program that simulated a system which has a round robin distribution of objects without using replication. This permits us to predict a theoretical or calibration curve, for the availability of a system compose of equally reliable discs, where all objects have equal importance. By taking the equation 2, with k = 0, t = 1, and λ = 1/MTTF, we obtain the probability of failure of one disc, F. Then the probability of availability of one disc p = 1 – F. Now we take the equation 1, with k = 0, and N equal the number of discs in the system, which gives $P_N = (1 - (1/e^{\lambda}))^N$. Since a successive Poisson process produces a Poisson process, we can take as the probability of availability of the system as an entity, then the λ_N of the system equals $\ln(1/(1-P_N))$ and $MTTF_{sys} = 1/\lambda_N$. After running the simulation with an MTTF = 200000, we obtain the values presented on Table 1. The plot in Figure 1 presents the expected and experimental MTBF columns.

More important than the fact of calibration is that the System Confidence goes rapidly down one 9 after the second disc is added, and goes down again after 20 discs added, as presented on Table 1.

Now that we have a valid simulator, and proved theoretically and experimentally that the System Confidence goes down while increasing discs in a system with the conditions previously mentioned we proceed with this investigation goal of an Elastic replication algorithm that maintain the Availability constant while adding discs.

4. Searching the Replication Algorithm

Previously we described that the System was a group of discs which have some particular distribution of objects.

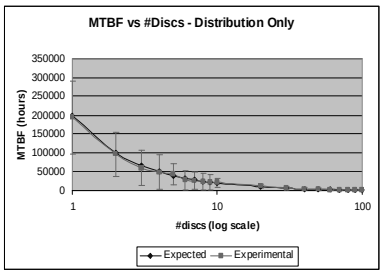

Figure 1. For clarity is plot in logarithmic scale.

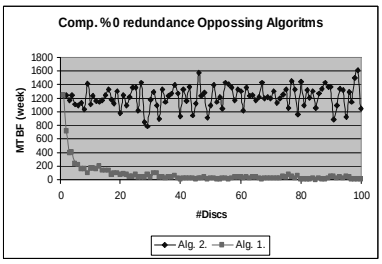

Figure 2. Alg. 1. – Fully distributed. Alg. 2. – No distribution.

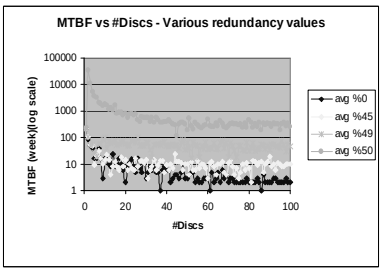

Figure 3. Fully distributed filling mechanism.

In more detail the System was modeled as a bi-dimensional Matrix where the columns represented the discs and the rows represented the objects. Then the simulation was divided in two primaries sub-process; first the filling of the Matrix, which represents the action of filling the discs, and the failure verification or availability of the objects. In fact, the hart of the simulation is the filling mechanism, which is where the actual elastic replication algorithm should be performed. Following the process of searching the desire replication algorithm we verify the response of failure on different filling mechanism.

4.1 Opposing Algorithms

On section 3, we detail the results for the theoretical and calibration curve, and that was obtain for a system with a round robin distribution of objects with no redundancy. Causally this distribution use on the filling mechanism gives a fully and evenly distributed Matrix (algorithm 1). On the other hand we could use a no distribution mechanism, where all objects will be located on only one disc, with or without replicas, because the replicas will be located on the same discs as the original objects (algorithm 2). These algorithms produce the plot on Figure 2 where we use an MTTF of 1250 weeks for the reliability of each individual disc. Notice that these two algorithms are opposites. The first have the maximum utilization of the system, since all discs have an evenly count of objects, but have the minimum Availability, while the second have the minimum utilization, because uses only one disc of the system, but have the maximum Availability; of curse they do not have any redundancy.

Clearly both previous algorithms are not the desire one, because the first fails to maintain the availability, and the second fails on practicality. Then we are looking for something in between. First, we started using replication with the use of the fully distributed mechanism. By running the simulation for various redundancy values we obtain the plot on Figure 3. These reflects that the MTBF for all replication values also presents the same overall response, that while the disc count increase from one disc to a certain count of discs the MTBF decrease.

4.2 Hybrid Algorithm

Previously we had that the full distribution mechanism is not good to maintain the availability constant although the usability is maximized, then we still need to find another filling mechanism. These mechanisms should be hybrids algorithms where the utilization will be sacrificed on behave of better availability.

4.2.1 Hybrid Algorithm Description

For these hybrids algorithms we use a logical visualization of the System as presented on Figure 4. Remember that we assume that all objects are equally important and all discs individually have equal MTTF, this means that no particular object nor disc give any advantage or disadvantage to the availability of the System. Observe that with a 60% redundancy, 6 out 10 original objects have replicas, and 4 out 10 does not have. These gives an Up and Down region, where the Up region has the objects with the more replicas. Each region can have its own distribution or filling mechanism.

	0	1	2	3	4	5	6	7	8	9	
0	1	0	1	0	0	0	0	0	0	0	60% Redundancy
1	0	1	0	1	0	0	0	0	0	0	40% Up
2	1	0	1	0	0	0	0	0	0	0	
3	0	1	0	1	0	0	0	0	0	0	
4	1	0	1	0	0	0	0	0	0	0	
5	0	1	0	1	0	0	0	0	0	0	
6	1	0	0	0	0	0	0	0	0	0	20% Down
7	0	1	0	0	0	0	0	0	0	0	
8	1	0	0	0	0	0	0	0	0	0	
9	0	1	0	0	0	0	0	0	0	0	

Figure 4. System or Matrix visualization.

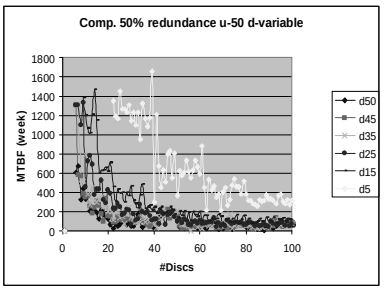

Figure 5.

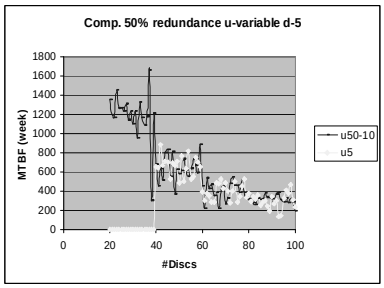

Figure 6.

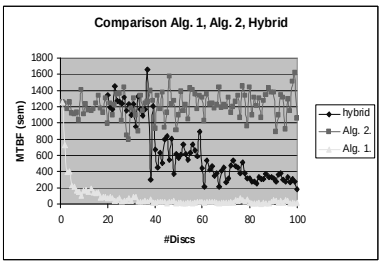

Figure 7. Hybrid plot is for 50% redundancy. u-50 d-5.

4.2.2 Hybrid Algorithm Results

On Figures 5 and 6 we present two plots for one particular redundancy value, taking the Up(u) and Down(d) utilization as parameters. On Figure 5 we change the Down utilization while keeping the Up utilization constant, and Figure 6 we do the opposite. For all cases the availability decreases while adding discs. For the 5 case there is a rapid decrease of availability on the curves for d5-d45 and for d50 it slows down, but also there is a clear difference between the curves trends, meaning that while increasing Down utilization the availability decreases. But for the 6 case there is not a clear difference between the curves trends, meaning that the Up utilization does not have much impact on the availability of the system.

In Figure 7 we present a comparison of the plots for the three algorithms. As expected the hybrid algorithm plot falls in between of the other two, but more important is that even that the hybrid algorithm uses replication it reflects a MTBF response lower than with no replication (Alg. 2), and lowest as the nodes are added. With this particular replication percent (50%) in the hybrid algorithm, it slows down the falling compare to the Alg. 1, but still goes significantly low.

5. Conclusions

1. We obtain a good curve for the MTBF values, that in fact the curve takes values very close to the expected MTBF values, letting us to conclude that the simulation program is valid.
2. By using a specific replication value on the System and maximum usability the availability decreases, proving the need for an Elastic Replication Algorithm.
3. The utilization of the System is a counter part of the reliability, meaning that at increasing utilization, decreasing reliability.
4. The group or region of discs where the fewer replicas are is the predominant point of failure of the System (The chain breaks on the weakest link).

6. Future Work

Currently overall utilization is equal to the highest utilization between Up and Down regions, but it could be higher by making the Down region to start on the last disc or from right to left, in opposite direction of the Up region, giving a total utilization to the sum of the two regions. After running the simulations with the previous filling mode, by this time we should be able to find a mathematical relationship between each parameter involved in the System Availability. With this obtained we should be able to construct the Elastic Replication Algorithm.

REFERENCES

[Drake88] Alvin W. Drake, *Fundamentals of Applied Probability Theory* (New York: McGraw-Hill Inc., 1988). G. Arthur Mihram, *Simulation: Statistical Foundations and Methodology* (New York: Academic Press, Inc., 1972).

[Özalp87] Özalp Babaoğlu, On the Reliability of Consensus-Based Fault-Tolerant Distributed Computing Systems, *ACM Transactions on Computer Systems*, 5(3), Nov 1987, 394-416.

[Gifford97] D.K. Gifford, Weighted Voting for Replicated Data, *Proceedings of the Seventh Symposium on Operating Systems Principles*, Pacific Grove, USA, Dec 1997, 150-162.

[Cooper2002] B.F. Cooper and H. Garcia-Molina, Peer-to-peer Resource Trading in a Reliable Distributed System, *Electronic Proceedings for the 1st International Workshop on Peer-to-Peer Systems*, Cambridge, USA, 2002.

[Stonebraker96] M. Stonebraker et al., Mariposa: A Wide-Area Distributed Database System, *VLDB Journal*, 5(1), Jan 1996, 48-63.

[Patterson88] D.A. Patterson, G. Gibson, and R.H. Katz, A Case for Redundant Arrays of Inexpensive Disks (RAID), *Proceedings of the ACM SIGMOD International Conference on Management of Data*, Chicago, USA, June 1988, 109-116.