UC Berkeley EECS Research Paper Summary

Alejandro Esteban Salazar

Computer Science Student

University of California, Berkeley

Berkeley, California, USA

May 7, 2015

Abstract

A brief summary and analysis of the UC Berkeley EECS research paper ”RAID-CUBE: The Modern Datacenter for RAID”, by Jayanta Basak and Randy H. Katz, submitted in April / 2015.

0.1 Summary

Big Data continues to grow as a field, as companies like Google, Amazon and Apple (and let's not forget our good friends at the NSA) continue to store ever increasing piles of information in what can only be described as massive in quantity. As this field's importance continues to grow in the Computing World, new issues start to arise from the manipulation and storage of such massive quantities of information. One such issue – the one addressed in this research paper – is that of Data Loss, and the capability of Data Retrieval and Recuperation when something goes wrong in large Data Factories. In this research paper, a new Data Storage configuration is presented, the RAID-CUBE, that according to the authors of the research paper, proves far more adequate against the loss of data in large data storage situations.

According to the research paper, current data protection mechanisms – mainly the standard RAID level redundancy mechanisms (RAID6 dual parity, RAID triple-parity) – suffer from a high decrease in Mean Time to Data Loss (MTTDL) as the number of data disks increases in an information cluster. By providing a standard methodology for measuring the MTTDL of different data storage configurations, the research paper goes on to demonstrate that RAID-CUBE (while still suffering from the inevitable data losses that come when the quantity of stored information becomes too big) proves to be more efficient against data loss due to its dampened decrease in MTTDL when compared to the MTTDL of the other standard RAID level redundancy mechanisms.

How this is accomplished is pretty tough to explain, let alone follow. Although the paper gave a rather easy to understand sequential process of how MTTDL's decrease depending on the quantity of data stored, the fact of the matter is, due to my lack of general knowledge in the area of Data Storage, there were times where I simply had to take the paper's word for it when it was describing complex simulations, as I simply lacking the basic knowledge to make a critical assessment of what was being presented. One thing I didn't fail to notice though, and that in fact I found absolutely fascinating, was how no matter the mechanism used, whether it'd be CUBE, Triple-Parity, Dual-Parity, or 2D-CUBE, all MTTDL's eventually approached the same constant as data quantity increased. (You can see this phenomenon in Figure 11 of the Research Paper). It's quite fascinating that all MTTDL's will never stop approaching 0 no matter their sophistication. Almost as if data scientists are fighting against data's natural state.

0.2 Conclusions

Since I don't know much about Data Science, I don't think I am in the right position to determine whether or not RAID-CUBE is as significant a contribution to Data Storage as the research paper seems to imply it is, but regardless of its actual significance or not, Mr. Basak and Mr. Katz have still managed to present a viable and more efficient mechanism against data loss in large scale data storage configurations than current standard RAID Level parity mechanisms.