Summary of “A Nine Year Study of File System and Storage Benchmarking” by Traeger et al

- Gary Miller

This report focused on the different ways in which benchmarking can be conducted as well as some of the major pitfalls associated with benchmarking. In the introduction of the paper two major themes were highlighted that dictate what the focus of a quality benchmark should be. The first of which states that the benchmark should be documented in as much detail as possible. For example, this includes but is not limited to allowing source code to be open to all validation of the benchmark’s correctness, listing system specs to better explain why certain results were found, and outlining the exact steps followed in the experiment. The next major theme that was highlighted to explain why things were done the way they were. The reason behind this is to once again allow for better validation of correctness for the benchmark. By stating a certain step was conducted in a specific fashion to satisfy a given requirement, those who were not part of the experiment are actually able to deduct more accurate and thorough results from the information that is presented because an understanding of purpose is established.

The second major part of this report is outlining and defining the three types of benchmarks that exist; macrobenchmarks, traces-based benchmarks, and microbenchmarks. The authors state that a macrobenchmark is a measurement or analysis of several operations and are best used for describing the overall performance of a system. The report later goes on into greater detail about the usefulness of macrobenchmarks, but also the fact that the results may not be truely representative of system performance due to lack of insight on exeogenous parameters on CPU resource allocation. Next the report defines what a trace-based benchmark is. A trace-based benchmark is one that documents a series of operation, and attempts at replaying them for a more accurate attempt at depcited representativeness of system performance. This is a highly desriable quality of a benchmark, but there are trade-offs that are associated with it. First of all, it there is a large amount of overhead in terms of implementation and storage cost, but more importantly it lacks flexibility. Trace-based benchmarks are only capable of desribing the specific sequence of operation that were docuemented, therfore any changes in the seqeucne of operation can lead to useless analysis of the system. The final type of benchmark the report outlines is called a microbenchmark. This is related to the analysis of a only one or two operations. These types of benchmarks are useful in desribing the effects or costs of a very small part of a sytem as well as analyzing worst-case outcomes. The downside associated with microbenchmarks is that often on their own they do not say too much about system performance and are generally only useful in supporting analysis with other benchmarks.

The remainder of research presented in this report generally focuses on further analysis the trade-offs of each type of benchmark by criticizing existing benchmarking systems. The examples presented in the later sections of this report outline exactly how a benchmark may not be representative of system performance, as well as outlines examples of how certain benchmarking systems did a good job at accurately representing system performance. One other important aspect of benchmarking that is presented in the earlier chapters of the report is how to properly present the results of an experiment. The authors explain concepts related to consistency of trials, such as keeping the CPU in a similar state for each trial as well as utilizing principles of statistics to verify results. Some of the statistical concepts identified are related to confirming a sample that satisfies a normal distribution of data by calculating things such as standard deviations and using confidence intervals. By following the guidelines presented in this article, implementing a benchmarking system will not only be easier due to an expanded knowledge in the field, it can also improve the results yielded from benchmarking experiments.