Deriving Storage Insights from the IO500

Luke Logan, Jay Lofstead, Anthony Kougkas, and Xian-He Sun
Department of Computer Science, Illinois Institute of Technology, Chicago, IL
llogan@hawk.iit.edu, gflofst@sandia.gov, {akougkas,sun}@iit.edu

I. Extended Abstract

Overview: Modern High-Performance Computing (HPC) applications generate and analyze massive amounts of data and are bottlenecked by I/O performance. The IO-500 [1] is a community-based benchmark that is designed to stress the I/O performance of HPC and Cloud storage systems in order to gauge the effectiveness of those systems for data-intensive workloads. The IO-500 collects various information on the design of the storage system, including the architecture and scale of the system, the software running on the system, and the vendors used to purchase hardware and software for the system. In this work, we aim to analyze the data collected by the IO-500 to gain insights on driving storage designs, purchasing decisions, potential bottlenecks, and the benefits/drawbacks of different hardware compositions (e.g. NVMe vs HDD). In our initial analysis, we found that storage systems that use NVRAM 3DXP technologies in their data storage nodes perform significantly (5x) better than those comprised of NVMe and SATA technologies.

Data Description: The IO-500 submissions dataset has 57 columns and 115 records. Submissions range from November of 2019 (for SC'19) to July of 2020 (for ISC'20). Various institutions, such as Intel, NVIDIA, and Red Hat, have made submissions, and systems including Tianhe-2E, Oracle Cloud, Oakforest-PACS, and Frontera were benchmarked; the largest deployment was on Oracle Cloud, with a total of 810 nodes. Information on the design of the storage system includes the number of nodes, operating system, kernel version, storage devices per node, RAM capacity per node, storage type, storage interface, and network used for metadata and data storage services. The type and name of the PFS used to store data is also collected.

Data Cleaning: Multiple issues were encountered during data cleaning: 1) entries to fields in the IO-500 were not standardized, resulting in multiple phrasings for the same value; 2) multiple submissions were missing information such as RAM capacity; and, 3) the meaning of certain fields were interpreted in different ways by different submitters, resulting in incorrect entries. The solution to the first issue was to manually standardize the data; for example, we removed kernel version from OS names since there was a separate field for that. For missing information, we made our best effort to discern the correct value based on other submissions by the same group. The solutions to the last issue depended on the field. For example, for storage interface, some wrote NVMe and SATA (which are buses), whereas others wrote iSCSI and NVMeof (which are network protocols). We defined this entry as the bus, and replaced values with iSCSI and NVMeof with the correct bus (NVMe, SATA, etc.).

ds_storage_type	ds_storage_interface	io500__score_avg	io500__score_std
3DXP	DIMM	605.492	129.9412387
3DXP	NVMe	142.4333333	15.07546793
NAND	NVMe	78.1	94.92771785
NAND	SATA	67.26666667	47.52310631
HDD	SAS	5.58	2.192464367

Fig. 1. Storage type/interface for data storage nodes vs IO500 score

Preliminary Analysis: We analyzed the data gathered for the 10-node challenge, where only 10 nodes were used to run the IO-500 job. In Figure 1, we compare the average IO-500 scores for submissions that used 3DXP, NAND, and HDD as the storage type in their data storage nodes. We found that, in general, data storage nodes that were comprised of NVRAM (DIMM) 3DXP technology performed significantly better (5x on average) than systems using NVMe-based technology, and that HDD performed significantly worse than all other technologies. However, we did not notice a significant difference between 3DXP and NAND when they are connected to the NVMe or SATA bus (they are within one standard deviation of each other).

Limitations: Additional information about the system design would be useful for determining the performance, cost, and power consumption of a particular system design, such as 1) the topology of the system, 2) information about the CPUs used for the different nodes (model number, family, core count, cache size, etc), 3) the average wattage-per-node, 4) the price for the nodes and their components, 5) model numbers for storage devices, and 6) the amount of storage per node.

Future Steps: We hope to include the results of SC'20 to add more historical data to the observations and to improve the data collection used for the IO500 by making the definition of the different fields clearer and providing more guidance on the format of inputs. Furthermore, we hope to develop a model that associate factors of system design (filesystem type, storage type, networking type, topologies, etc.) with the performance, power consumption, and cost of the storage system. This model could then be used to determine the power consumption and financial cost needed to perform a certain workload.

References

[1] IO500, August 2020. [Online]. Available: https://www.vi4io.org/io500/about/start