

I/O Workload Characterization on Supercomputers using Unsupervised Learning

Pablo J. Pavan¹, Jean Luca Bez¹, Matheus S. Serpa¹, Francieli Zanon Boito², Philippe O. A. Navaux¹

¹Institute of Informatics, Federal University of Rio Grande do Sul (UFRGS) — Porto Alegre, Brazil

²Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LIG, 38000 Grenoble, France

{pjpavan, jean.bez, msserpa, navaux}@inf.ufrgs.br, francieli.zanon-boito@inria.fr

Abstract—I/O operations are the bottleneck for an increasing number of applications due to the difference between the processing and data access speeds. The application access patterns reflect directly on their performance. Therefore, it is essential to characterize the I/O workload of large systems in order to identify problems and propose solutions. With this premise, we propose an I/O characterization approach that uses unsupervised learning to cluster jobs with similar I/O behavior using information from high-level aggregated traces. We apply our proposal to four months of activity — a total of 28.938 jobs — from the Intrepid machine at Argonne Laboratory. We show that our approach helps identify the most representative behaviors among the applications.

Index Terms—Application I/O Behavior, Parallel I/O, I/O workload characterization, unsupervised learning, clustering algorithms