Data Management in Hierarchical Storage using Reinforcement Learning
(Project in Computational Science course, 15 ECTS)

Tianru Zhang, Salman Toor *

Department of Information Technology, Uppsala university

Project Description

According to the available statistics, the overall amount of data created worldwide had reached 79 zettabytes in 2021 [1]. This data explosion brings numerous big data challenges for all industries, academia, and even for individuals. Large-scale data management is one of the most challenging prospects in the big data domain [2]. Over the years, a number of frameworks for data management have become available. Most of them are highly effective, but ultimately create data silos. It becomes difficult to move and work coherently with data as new requirements emerge. A possible solution is to use an intelligent hierarchical (multi-tier) storage system (HSS). A HSS is a meta solution that consists of different storage mediums organized as a jointly constructed storage pool. The main idea is to connect different independent storage solutions and move the data between them according to policy designed to meet a set of requirements. In one of our recent articles [3], we have comprehensively described the underlying challenges related to storage hierarchies and introduced an open-source hierarchical storage framework with a dynamic migration policy based on reinforcement learning (HSM-RL). We have also built a simulation and a fully functional cloud-based framework for general purpose datasets. We have conducted experiments based on static and dynamic datasets and the results have proved the effectiveness, efficiency and consistency of the RL-based policy.

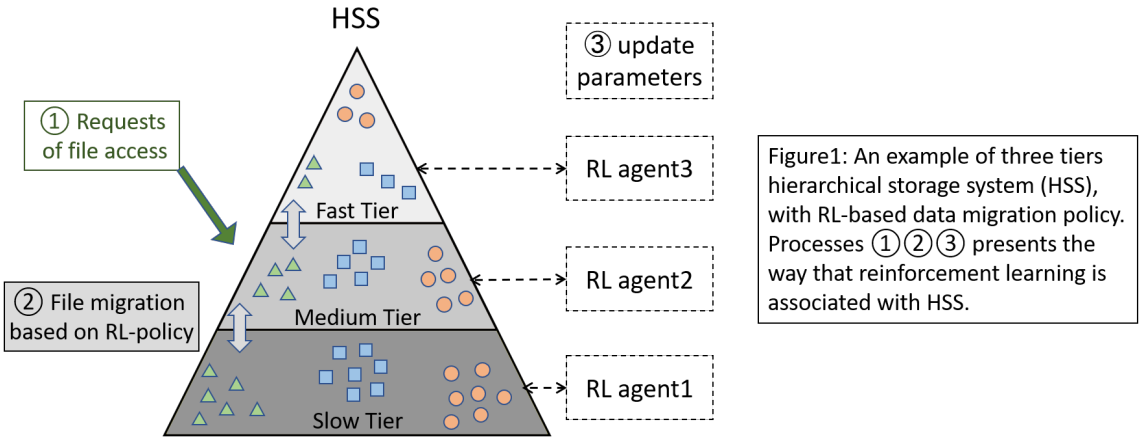

Figure1: An example of three tiers hierarchical storage system (HSS), with RL-based data migration policy. Processes ①②③ presents the way that reinforcement learning is associated with HSS.

In this project, the aim is to first understand the available HSM-RL framework. Then it is expected to explore the effect of different learning methods and parameters with respect to performance and efficiency. Good programming skills in Python are required, and knowledge of reinforcement learning and data management is preferred.

References

[1] *Statista, Big data - Statistics Facts*. https://www.statista.com/topics/1464/big-data/#dossierKeyfigures (cit. on p. 1).

[2] Vivien Marx. "The big challenges of big data". In: *Nature* 498.7453 (2013), pp. 255–260. DOI: https://doi.org/10.1038/498255a (cit. on p. 1).

[3] Tianru Zhang, Andreas Hellander, and Salman Toor. "Efficient Hierarchical Storage Management Empowered by Reinforcement Learning". In: *IEEE Transactions on Knowledge and Data Engineering* (2022), pp. 1–1. DOI: 10.1109/TKDE.2022.3176753 (cit. on p. 1).

*tianru.zhang@it.uu.se, salman.toor@it.uu.se