Intelligent Resource Management for Processing Large Data Streams

HarmonicIO framework is designed to fulfil the needs of the scientific datasets. The conventional streaming frameworks are extremely efficient and robust for the large datasets based on very small individual object size. Whereas scientific datasets are also huge but the individual object size is significantly large (100s of KBs or MBs). Together with filling the gap of processing relatively large individual objects, the HarmonicIO also offers cloud-native execution model, auto-scaling and complete isolation for processing engines (PEs) using container technology. The article [1] highlights the essential features and initial experiments related to HarmonicIO framework. We have also presented a detailed comparison of HarmonicIO and Spark streaming framework to highlight its role and importance as a specialized streaming framework for scientific datasets.

Recently, we have developed an intelligent resource management (IRM) component for the HarmonicIO. The IRM component is responsible for the optimal placement of containerized PEs with a minimal performance loss and introduces autoscaling based on the pressure of streaming requests. We have used the online bin-packing algorithm for processing engine placement. The results are promising and we have achieved similar performance as presented in article [1] using 1/5 of the resources.

The newly developed IRM component adheres to a computer-centric approach. In this master thesis, the task is to include more information (network, storage, memory, operational cost and information about inhomogeneous environments) to the IRM component that will be one step further towards optimal resources utilization with the minimal performance loss. For this, the task is to study and implement multi-dimensional online bin-packing for the HarmonicIO framework.

Contact Person:

Salman.Toor@it.uu.se

References:

1. P. Torruangwatthana, H. Wieslander, B. Blamey, A. Hellander and S. Toor. HarmonicIO: Scalable Data Stream Processing for Scientific Datasets. *2018 IEEE 11th International Conference on Cloud Computing (CLOUD)*, San Francisco, CA, USA, 2018, pp. 879-882. doi:10.1109/CLOUD.2018.00126
2. B. Blamey, A. Hellander, S. Toor. Apache Spark Streaming and HarmonicIO: A Performance and Architecture Comparison. eprint arXiv:1807.07724