Cloud Reamer:

Enabling Inference Services in Training Clusters

Osama Khan [1], Gwanjong Park[2], Junyeol Yu[2] and Euiseong Seo[2*]

[1] *AI System Engineering, Sungkyunkwan University, Republic of Korea*

[2] *College of Computing and Informatics, Sungkyunkwan University, Republic of Korea*

Corresponding author (Electronic mail: euiseong@skku.edu)

Abstract~ CPU cores in GPU servers are often underutilized during DNN training. Co-locating CPU-based inference tasks with DNN training offers an opportunity to utilize these idle CPU cycles. However, three technical challenges must be addressed: avoiding disruption to training workloads, meeting different performance requirements for online and offline inference, and swiftly adjusting inference configurations based on available resources. This paper proposes Cloud Reamer, a scheme to co-locate training and inference tasks on GPU servers, optimizing unused CPU cycles without disrupting training. Cloud Reamer prioritizes training tasks to minimize interference. For online inference, it allocates cores to ensure predictable performance, while for offline inference, it uses all available cores to maximize throughput. Cloud Reamer enhances online and offline inference performance by dynamically adjusting configurations based on surplus CPU resources. Evaluations show that Cloud Reamer improves inference throughput with minimal impact on training, maintaining training interference below 3.2%. It meets latency requirements for 46% more requests for online inference and achieves a 61x throughput increase for offline inference compared to conventional methods.

Acknowledgments This research was partially supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2024-RS-2023-00258649) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation), and by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2023-00321688).