swPredicter: A Data-Driven Performance Model for Distributed Data Parallelism Training on Large-Scale HPC Clusters

Xianyu Zhu, Ruohan Wu, Hong An ✉, Junshi Chen ✉

***Abstract*—With the complexity of heterogeneous architecture and many nodes collaboration, large-scale HPC clusters pose challenges in resource utilization and performance optimization during distributed data parallelism (DDP) training. Performance modeling aims to analyze application bottlenecks and guide algorithm design, but existing performance models fail to consider the impact of system architecture on communication performance and make a systematic analysis of distributed training. To address the issues, the paper proposes swPredicter, a data-driven performance model devised for accurately predicting the performance of DDP training. First, an original performance dataset is developed according to the various communication patterns in the runtime to avoid systematic errors. Subsequently, a new multi-branch module FNO-Inception is proposed, combining FNO (Fourier Neural Operator) layer with Inception structure to simultaneously utilize various frequency features. Finally, by introducing the FNO-Inception module, a novel regression model FI-Net is constructed to fit complex nonlinear relationships. The experimental results demonstrate that FI-Net can accurately predict the performance of DDP training on the Sunway OceanLight supercomputer with an overall MAPE error of 0.93%, which outperforms the other baseline models.**

***Index Terms*—High-Performance Computing, Performance Modeling, Deep Learning, Distributed Training**

I. Introduction

THE integration of high-performance computing (HPC) and artificial intelligence (AI) facilitates the invocation of large-scale cluster resources to accelerate the AI applications for scientific research, such as weather forecasting [1], bioinformatics [2], materials modeling [3] and quantum mechanics [4]. Distributed data parallelism (DDP) is the most commonly used approach to accelerate AI training, which relies on distributed nodes collaboration to improve the training efficiency. However, the complex system architecture and diverse communication patterns of large-scale HPC clusters bring challenges in algorithm performance improvement and resource management [5]. In order to improve the performance of DDP training, it is necessary to identify and optimize the application bottlenecks of computation and communication [6]. Meanwhile, reasonable resource scheduling requires the accurate estimation of the algorithm performance, which can help in resource allocation and overall system efficiency. On the other hand, many AI model design approaches [7], [8] combine the model performance to construct a more efficient network. Performance modeling aims to analyze and predict the algorithm performance under various conditions, which can help optimize resource usage, guide algorithm design, and reduce development costs.

Existing performance modeling approaches can generally be classified into analytical modeling and data-driven modeling [9], [10]. The analytical modeling method evaluates the algorithm performance through the theoretical analysis models, such as queueing theory [11] and stochastic model [12]. When facing the complex system and diverse algorithms, it spends substantial costs in constructing an effective mathematical model. Since the model is designed specifically to describe a particular system or algorithm, its applicability is rather limited. With the advantages of wider applicability and lower costs, the data-driven modeling method has become popular. It utilizes collected extensive data to train AI models that can automatically fit the complex nonlinear relationships between algorithm parameters and performance [13].

Many data-driven modeling methods [14], [15] focus on devising more complex models to fit the relationships between algorithm parameters and performance within the dataset. However, they fail to consider the impact of the runtime system topology on the algorithm performance. For example, though the number of allocated computational nodes is the same, different scheduling methods and resource usage conditions will make the physical nodes different, leading to different communication bandwidth and latency [16]. Moreover, there are many performance models only built for model inference [17], [18], lacking the analysis of backward propagation, parameters update and distributed communication. To aid in the resource management and AI model design on large-scale HPC clusters, there is a lack of the performance model which supports the systematic analysis and prediction of DDP training.

To address these issues, the paper proposes a data-driven performance model swPredicter for accurately predicting the performance of DDP training on large-scale HPC clusters. First, a performance dataset has been developed, which in-

The paper is supported by the Strategic Priority Research Program of Chinese Academy of Sciences, Grant (XDB0500102). And this work is financially supported by Laoshan Laboratory (LSKJ202300305).

Xianyu Zhu and Ruohan Wu are with the School of Computer Science and Technology, University of Science and Technology of China, Hefei, China. (E-mails: {zhuxy, ruohanwu}@mail.ustc.edu.cn)

Hong An and Junshi Chen are with the School of Computer Science and Technology, University of Science and Technology of China, Hefei, China and also with Laoshan Laboratory, Qingdao, China. (E-mails: {han, cjuns}@ustc.edu.cn)

Hong An and Junshi Chen are corresponding author.