

Dr Maciej Besta, ETH Zürich

Temat:

Parallel and Distributed Graph Neural Networks: An In-Depth Concurrency Analysis

Streszczenie:

Graph neural networks (GNNs) are among the most powerful tools in deep learning. They routinely solve complex problems on unstructured networks, such as node classification, graph classification, or link prediction, with high accuracy. However, both inference and training of GNNs are complex, and they uniquely combine the features of irregular graph processing with dense and regular computations. This complexity makes it very challenging to execute GNNs efficiently on modern massively parallel architectures. To alleviate this, we first design a taxonomy of parallelism in GNNs, considering data and model parallelism, and different forms of pipelining. Then, we use this taxonomy to investigate the amount of parallelism in numerous GNN models, GNN-driven machine learning tasks, software frameworks, or hardware accelerators. We use the work-depth model, and we also assess communication volume and synchronization. We specifically focus on the sparsity/density of the associated tensors, in order to understand how to effectively apply techniques such as vectorization. We also formally analyze GNN pipelining, and we generalize the established Message-Passing class of GNN models to cover arbitrary pipeline depths, facilitating future optimizations. Finally, we investigate different forms of asynchronicity, navigating the path for future asynchronous parallel GNN pipelines. The outcomes of our analysis are synthesized in a set of insights that help to maximize GNN performance, and a comprehensive list of challenges and opportunities for further research into efficient GNN computations. Our work will help to advance the design of future GNNs.