Performance guarantees in dynamic networks and graph algorithms

Arvind Easwaran[1] · Sebastian Altmeyer[2]

Deep Neural Networks (DNNs) are steadily gaining relevance in several safety- and time-critical applications in the domain of cyber-physical systems. However, their resource intensive nature poses significant challenges to their deployment in resource-constrained embedded settings. Orthogonally, with this increasing complexity of workloads, real-time resource management techniques must also consider functional dependencies among them. Focussing on such research questions, this special issue titled "Performance Guarantees in Dynamic Networks and Graph Algorithms", presents two papers which are extended versions of some of the outstanding papers presented at IEEE Real-Time Systems Symposium (RTSS) 2022.

The IEEE RTSS is a premier conference in the field of real-time systems and is a venue for researchers and practitioners to showcase innovations covering all aspects of real-time systems. In 2022, RTSS celebrated its 43rd anniversary, and it continued the trend of making RTSS an expansive and inclusive event striving to embrace new and emerging areas of real-time systems research. RTSS 2022 received a total of 128 submissions, out of which 37 papers were accepted to appear at the conference, and a further selection of 4 outstanding papers from this shortlist was done by a best-paper committee. The authors of these outstanding papers were invited to submit an extension of their work to this special issue.

The first paper in this special issue, titled "Inference Serving with End-to-End Latency SLOs over Dynamic Edge Networks", was written by Vinod Nigade, Pablo Bauszat, Henri Bal and Lin Wang. It delves into deep learning (DL) inference serving for mobile and IoT applications, essential for technologies like augmented reality and autonomous driving. DL models, such as DNNs, require intensive computations, and are challenging to deploy on devices with limited capabilities. Techniques like model compression, quantization, and pruning help, but still face accuracy loss and memory

Arvind Easwaran
arvinde@ntu.edu.sg

[1] Nanyang Technological University, Nanyang Ave, Singapore

[2] University of Augsburg, Augsburg, Germany

constraints. This paper highlights the importance of considering both network and compute times, proposing new designs for timely DL inference serving.

The second paper in this special issue, titled “The Shape of a DAG: Bounding the Response Time Using Long Paths”, was written by Qingqiang He, Nan Guan, Mingsong Lv, Xu Jiang and Wanli Chang. It presents advancements in the analysis of Directed Acyclic Graph (DAG) tasks under work-conserving scheduling on multi-core platforms. Traditional methods, like Graham’s well-known bound, rely on the longest path and total workload, often leading to overly pessimistic estimates. This research introduces a more precise response time bound by considering multiple long paths, offering a clearer picture of parallel execution and reducing unnecessary interference assumptions. With new abstractions and innovative techniques, the presented bounds not only theoretically surpasses Graham’s, but also shows substantial empirical improvements.

We hope you enjoy this special issue!