Primary Supervisor: Ahmed M. A. Sayed
Secondary Supervisor: TBA
Title: Efficient Machine Learning on Decentralized Data at Scale
Abstract: AI/ML systems are becoming an integral part of user products and applications as well as the main revenue driver for most organizations. This resulted in shifting the focus to bringing the intelligence towards where the data are produced including training the models on these data. Existing approaches operate as follows: 1) the data is collected on multiple servers and processed in parallel (e.g., Distributed Data-Parallel); 2) the server coordinates the training rounds and collects model updates from the clients (e.g., Federated Learning); 3) the server splits the model training between the clients and the server (e.g., Split Learning); or 4) the clients coordinate among themselves via gossip protocols (i.e., Decentralized Training). The challenges that manifest themselves are the highly heterogeneous learners, configurations, environment, communication and synchronization overheads, fairness and bias, and privacy and security. Therefore, existing approaches fail to scale with a large number of learners and produce models with low qualities and high bias at prolonged training times. It is imperative to build systems that provide high-quality models in a timely manner. This project addresses this gap by exploring novel ideas and proposing efficient and scalable ML systems for decentralized data.

Primary Supervisor: Ahmed M. A. Sayed
Secondary Supervisor: TBA
Title: Novel Optimisations Towards Accelerating Generative AI and LLMs
Abstract: In the rapidly evolving landscape of artificial intelligence, the development of sophisticated Generative AI and Large Language Models (LLMs) has become pivotal for various applications, ranging from natural language processing to creative content generation. However, the training of these models is computationally intensive, often requiring substantial time and resources. This project will study and propose system and algorithmic optimizations to accelerate the training process for Generative AI and LLMs, addressing the challenges posed by the complexity of these models. The core focus of this research lies in the exploration and implementation of advanced parallel computing techniques, leveraging the power of distributed systems and specialized hardware accelerators. By optimizing algorithms, employing parallelization strategies, and harnessing the capabilities of GPUs, TPUs, or emerging AI-specific hardware, this project aims to significantly reduce the training time of Generative AI and LLMs, making the process more efficient and cost-effective. Furthermore, the study delves into the realm of transfer learning and explores techniques to enhance model convergence and accuracy. By leveraging pre-trained models and developing novel transfer learning methodologies, the research intends to minimize the amount of data and computational resources required for training, thereby democratizing access to cutting-edge AI technologies.