Energy-efficient and Scalable ML

Name of supervisor:

Ali Ramezani-Kebrya (actively publishing in top ML venues, PI in Norwegian Centre of Excellence Integreat and SFI Visual Intelligence, ELLIS member, NeurIPS Area Chair)

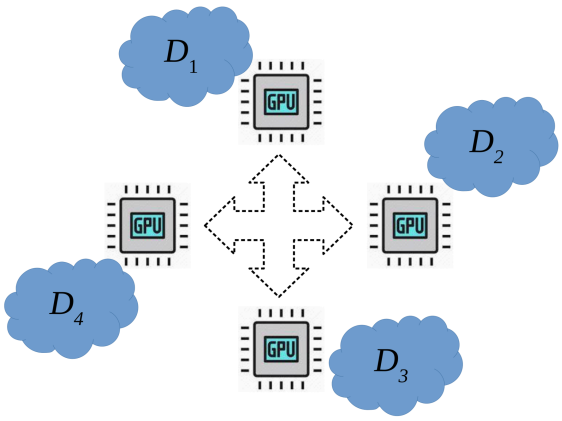
Preferred background of candidate(s):

Machine learning, Computer Science, Math, or Physics

Number of available projects : 2

Preferred project period: Flexible

Outline of project work including expected outcomes/deliverable



GPT-3 has been trained on 10,000 V100 GPUs in a Microsoft cloud data center with unprecedented communication costs and Carbon footprint. The total emissions of training Llama 2 family of models and GPT-3 are estimated to be 539 and 552 tons of CO2 equivalent, respectively, which is not sustainable. Quantization, sparisification, and pruning of the underlying neural networks and data movement minimization are proposed to make ML more efficient as data movements between global and shared memory dominates the energy consumption of hardware platforms and the total energy consumption is a monotonically increasing function of the model size.

We have significantly accelerated ML in multi-node systems, i.e., handled scalability issues by developing highly communication-efficient and large-scale training in supervised learning, and training deep generative adversarial networks without sacrificing accuracy [1–3].

In this work, you will continue my current work with our previous undergrad summer intern from the math department. Our intern spent a summer with me on a problem and did a super great job. We now plan to finalize the work for a top ML submission by adding experimental results.