Montserrat Alonso
Taiki Tsukahara
Hailu Xu
CECS 326
May 12, 2024

Energy Saving With ENVPIPE

With the rapid consumption of AI in our technologies today, data centers have become one of the largest contributors to carbon emission in our environment. Researchers like Sangjin Choi suggest that the amount of carbon it produces is comparable to our worldwide airline industry today. Their research focuses on improving the energy savings in multi-GPU deep neural network training also known as DNN training, which is a method used in most ML workloads. Previously, the GPU Dynamic Voltage and Frequency Scaling (DVFS) method was used in reducing the energy consumption in DNN training.

DVFS works by adjusting the voltage and frequency of a GPU dynamically to balance the performance and energy consumption. However, the Zeus and other researchers suggest that this adjustment can unintentionally affect user-provided hyperparameters such as batch size and learning rate that may lead to statistical efficiency of the training process, ultimately diminishing the effectiveness in optimizing the energy consumption. Another constraint is that it can be a challenge to determine how much degradation in performance is acceptable at the cost of optimizing the energy. When training jobs, the completion time can be widely varied and it can be unpredictable so practitioners may be unable to gauge how much delay they can tolerate. It was explained that the minor performance degradation in a long running task can result in large delays; saying 10% degradation in performance for a month long training task can mean three-day delays.

In their study, they introduced ENVPIPE for energy savings with no accuracy and performance degradation. Unlike the previous approach that comes with side effects by modifying user-provided hyperparameters, ENVPIPE ensures that they remain unchanged. The method refrains from altering any hyperparameter such as batch size and they make sure that that data dependency also remains unchanged during the execution of the pipeline units. ENVPIPE relies only on a side-effect free control mechanism, the SM frequency, to optimize the energy usage. To avoid any decline in performance, ENVPIPE focuses on capitalizing the naturally occurring pipeline bubble encountered during the training of the pipeline parallelism. It will selectively decrease the SM frequency to reduce the energy consumption of pipeline units. As for the methodology, ENVPIPE is applied as a library within the ML frameworks so it can separate policy and mechanism for easier integration. When determining the optimal SM (Streaming Multiprocessor) frequency values and employing scheduling techniques, evaluation results demonstrate a significant amount of energy savings of up to 25.2% and 28.4% in a single and multi node setups, while also with the minimal performance degradation of less than 1%.

For the analysis part of this assignment, the article's proposed problem was optimizing energy consumption during the training of DEEP Neural Network(DNN) models by emphasizing the challenges. Some of the challenges that were mentioned were statistical Efficiency vs Energy Efficiency trade off, Modification of Hyperparameters, and Performance degradation vs Energy savings. The identified key constraints that are crucial when designing energy saving was no accuracy degradation and no performance degradation. ENVPIPE's design addresses the energy efficiency challenges by automating hyperparameter tuning, energy profiling and pipeline unit rescheduling which wouldn't compromise statistical efficiency. The system starts by profiling the energy consumption of DNN training jobs, the data is then used to get a better understanding of the trade-offs between performance and energy usage. Then the schedule of pipeline units is strategically adjusted to create usable bubbles to maximize energy saving opportunities. To further reduce energy consumption, SM frequencies are optimized. This design makes it adaptable to different ML platforms which allows users an easier experience for controlling multi-GPU pipeline scheduling and energy consumption.

When reading the article, we agreed that ENVPIPE'S has a great level of understanding of system-level optimization and resource management to address energy efficient DNN training. The algorithm that is presented in the article resonates with the principles of process scheduling, resource allocation and optimization. ENVPIPE's reconfigured algorithm resembles dynamic process scheduling techniques that adjusts resource allocation based on workload characteristics to optimize system performance. When we think about how pervasive technologies have become, especially with the rapid growth of AI usage in our environment, we never really realize or consider the carbon emissions when using our phone and computer devices. After reading the article, it reminded us of the usage of data centers or a server such as our google server where we usually store our class document for assignments or microsoft school emails. According to the very well known static data web page Statista, a company like Google emits about 10.2 million metric tons of carbon. This is equivalent to 22 million of people's carbon emission they produce on cars annually. To be honest, it's hard to visualize how big this is but we can all agree that data centers or servers that host data do have a big contribution to climate change. So reading this article was very interesting and made us recognize the importance of saving energy consumption when using our technology devices. Overall, ENVPIPE'S design shows an understanding of operating system principles that can be applied to the domain of deep learning and distributed system optimization. They also contribute to Operating systems, Machine learning and advancing energy efficient computing while maintaining high performance standards.

Although ENVPIPE has a great level of understanding and presents a compelling approach, an area of improvement would be further exploration of dynamic reconfiguration strategies based on real time workload characteristics. ENVPIPE currently employs an iterative approach by adjusting SM frequency in small steps, however exploring dynamic reconfiguration strategies based on real time workload characteristics would lead to further energy saving.For example, to enhance the system's adaptability and efficiency, would be an integration of machine learning models or heuristics to predict optimal frequency adjustments based on the workload pattern and system conditions.

Work Cited

Choi, S., Koo, I., Ahn, J., Jeon, M., & Kwon, Y. (1970, January 1). *{EnvPipe}: Performance-preserving {DNN} training framework for saving energy*. USENIX. https://www.usenix.org/conference/atc23/presentation/choi