Paper Review - USHER: Holistic Interference Avoidance for Resource Optimized ML Inference

Sai Venkat Malreddy

22 oktober 2024

1 Idea and Contribution

The authors addresses the challenge of optimizing GPU resource utilization for machine learning (ML) inference in data centers. The increasing size and popularity of deep learning models have led to higher computational demands and significant financial costs due to the inefficient use of GPU resources. Existing techniques, such as batching and spatial multiplexing, face limitations like inter-model interference, inflexible resource allocation, and suboptimal use of GPU memory and compute capacity. These inefficiencies result in low GPU utilization—often between 25% and 50%—and inflated operational costs, with ML inference accounting for more than 90% of production expenses. USHER provides a holistic solution that maximizes GPU utilization while minimizing interference between models through three main components: a GPU kernel-based resource estimator that accurately determines resource needs without costly profiling, a heuristic-based scheduler that manages model placement and replication to reduce interference, and an operator graph merger that combines multiple models' computation graphs to minimize cache interference. These strategies enable USHER to significantly improve GPU utilization, increase the number of inference requests completed within service-level objectives (SLOs), and reduce the overall cost of inference serving. Experiments demonstrate that USHER achieves up to 2.6× higher goodput and 3.5× better cost-efficiency compared to existing methods.

2 Three Positive Comments

- **Resource Optimization**
  A holistic approach to optimize both computation and memory utilization of GPUs for ML inference.
- **Interference-Aware Scheduling:**
  USHER introduces a lightweight, interference-aware scheduling mechanism that balances different models based on their resource needs.
- **Cache Interference Reduction:**
  The operator graph merger in USHER is an innovative approach that minimizes cache interference by merging similar computation graphs.

3 Three Negative Comments

- **Generalizability:**
  While the results for USHER are promising in the controlled experiments, there may be challenges in generalizing the solution across a wide range of GPU architectures or in more diverse, real-world data center environments.

- **Costs:**
  It Introduces some overheads, such as operator graph merging and workload rescheduling. Although these overheads are justified by improved resource efficiency, the paper acknowledges that rescheduling may not always be suitable for environments with very frequent workload changes, potentially diminishing the system's overall efficiency in extremely dynamic settings.

- **Heterogeneous Environments:**
  USHER's evaluation focuses primarily on homogeneous or relatively straightforward heterogeneous GPU environments. In real-world data centers, resource pools are often heterogeneous, with a mix of different hardware generations and varying capabilities.

4 Questions unanswered about the paper

- **Energy Impact**
  What is the impact of USHER on energy consumption in data centers?