# MoE-Infinity: Activation-Aware Expert Offloading for Efficient MoE Serving

## Abstract

The increasing complexity and size of Mixture-of-Experts (MoE) models present significant challenges for AI services. Models like Mixtral, Claude-3, OpenAI's GPT-4 represent state-of-the-art (SOTA) AI models, with architectures characterized by their vast parameter spaces, reaching the scale of trillions of parameters and occupying terabytes of memory. Serving (inference) with such models today necessitates scores of GPUs to accommodate the memory demands. This need arises due to low latency requirement for which the current practice is to store and process parameters near GPU high bandwidth memory (HBM).

In the quest to address the significant memory demands, there is a growing interest in exploring both model compression methods and memory offloading methods to reduce GPU memory consumption. Techniques such as quantization offer a limited amount of memory footprint reduction (4x from float16 to int4), while often coming with a loss in generalization and accuracy. Offloading approaches, on the other hand, target lossless behaviour and do so by storing experts in MoE models on external storage mediums and only loading them onto GPUs as needed. However, this solution introduces excessive traffic over bottleneck PCIe links. This is because offloading systems (e.g., Zero-Offload) are designed for dense models, whereas sparsely activated MoE models where only 20% of the parameters are used on average. Such systems treat each sparse layer as a dense layer by fetching all parameters to the GPU. Fetching a 1TB model results in a latency of 40 seconds on PCIe4, while the computation latency is only 2-3 seconds.

We present MoE-Infinity, a novel serving system design to mitigate the latency overhead associated with offloading MoE parameters. Our approach leverages two observed MoE characteristics: sparse activation of experts and temporal locality. MoE-Infinity features sequence-level expert activation tracing, a new approach adept at identifying sparse activations and capturing the temporal locality of MoE inference. By analyzing these traces, MoE-Infinity can detect the imminent expert activation, thus fetching only experts needed to GPU and reducing the traffic on PCIe. MoE-Infinity can also predict the expert activation for the subsequent layers for prefetching. Prefetching in MoE-Infinity prioritizes the loading of experts likely to be needed soon, based on their activation history. Caching in MoE-Infinity selectively retains experts in the cache based on their activation frequency and layer position during each generation.

Our evaluation of MoE-Infinity in a GPU cluster environment, serving MoE models such as Google's Switch Transformers, Facebook's NLLB and Mixtral, demonstrates significant performance improvements over current state-of-the-art (SOTA) serving systems (e.g., DeepSpeed, CUDA Unified Memory). Notably, MoE-Infinity achieves up to 4-20x improvements in latency, and 2-10x improvements in throughput, compared to existing offloading systems. MoE-Infinity also reduces 8x in GPU resources without performance degradation.