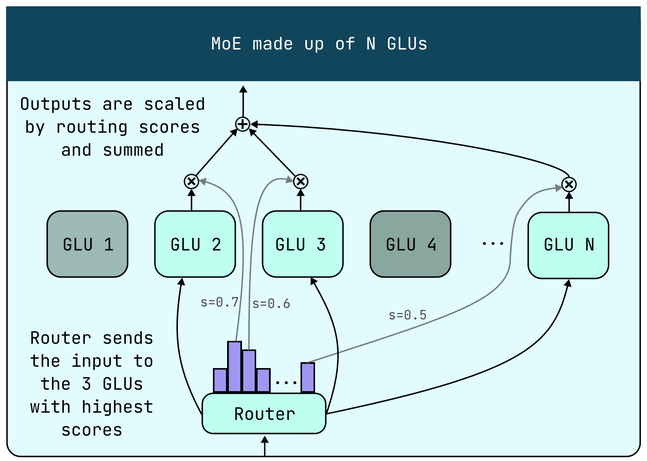
# GPU Kernels for Mixture of Experts

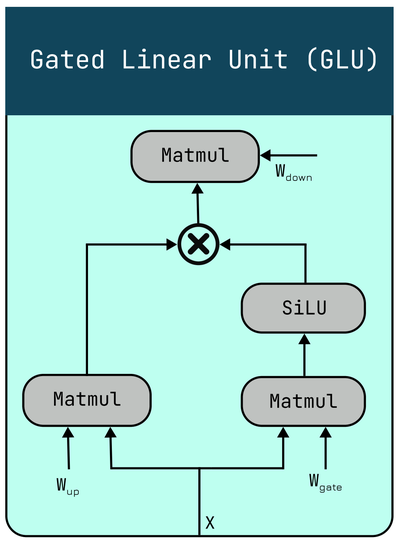
Arthur Feeney[1] Ying Wai Li[2] Aparna Chandramowlishwaran[1]

[1]University of California Irvine, [2]Los Alamos National Laboratory

**Background**: A Sparse Mixture of Experts (MoE) is a popular module to increase the number of parameters in a model, without a corresponding increase in FLOPs. This is typically implemented by using a router to select a subset of "experts" for different inputs:




It is common for each experts to be an MLP or GLU:

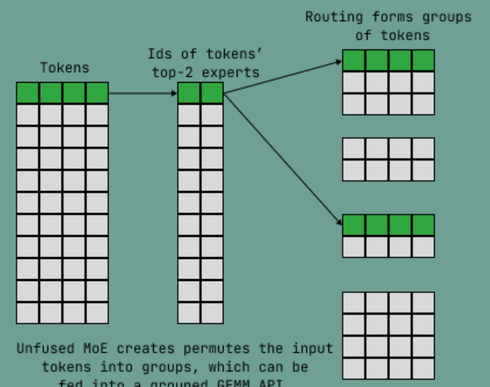


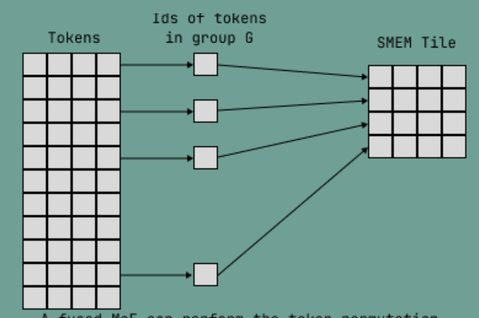

**Goal**: develop GPU kernels for Mixture of Experts, mainly for inference or fine-tuning.

An MoE with N experts performs N matrix multiplications with N weight matrices and N different subsets of tokens. This does not map well to typical batched GEMM APIs. Alternatives APIs to process multiple GEMMs of different sizes include:

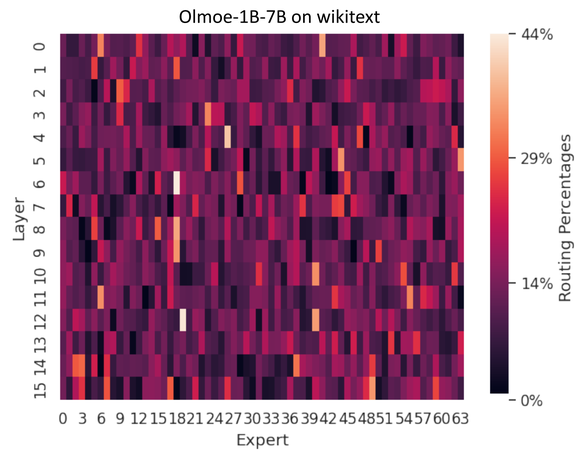
1. Magma variable batch gemm,
2. CUTLASS and cuBLAS Grouped GEMM.

An MoE includes additional operations for routing, which can be fused with the GEMM kernels.

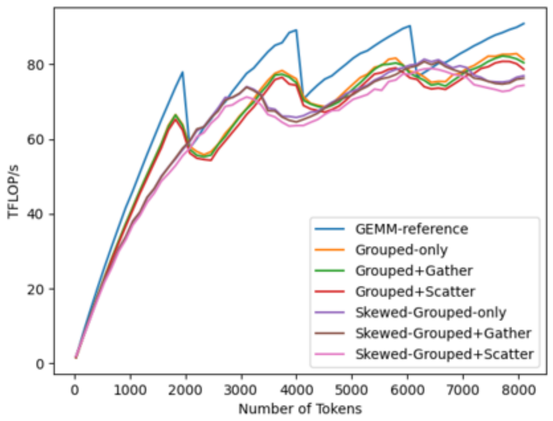


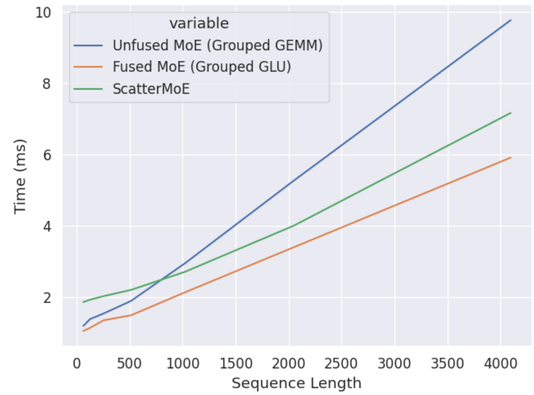

In Practice, routing in MoEs can be imbalanced so that experts may process different quantities of tokens.



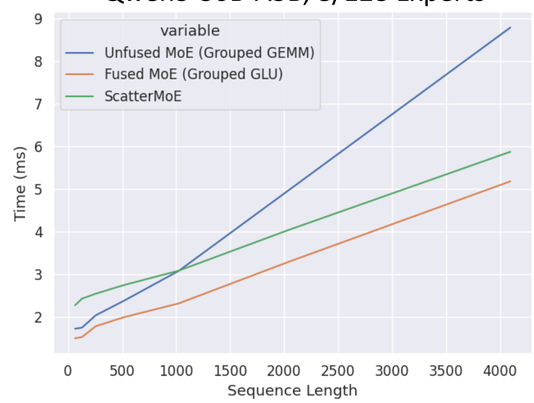

The fused routing does not affect performance relative to unfused grouped GEMM, though the routing distribution does affect performance.




Olmoe-1B-7B, 8/64 Experts




Qwen3-30B-A3B, 8/128 Experts




Ernie4.5-21B-A3B, 6/64 Experts