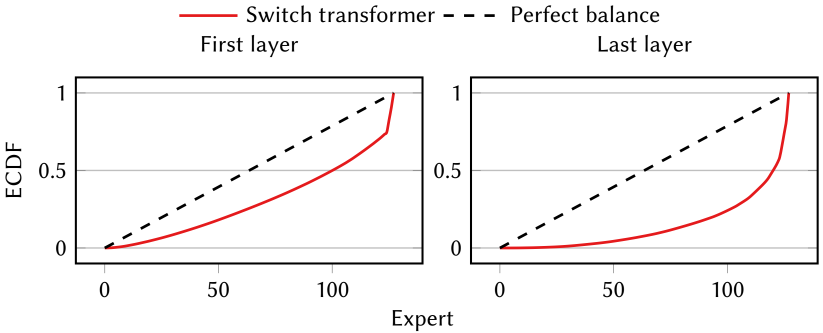
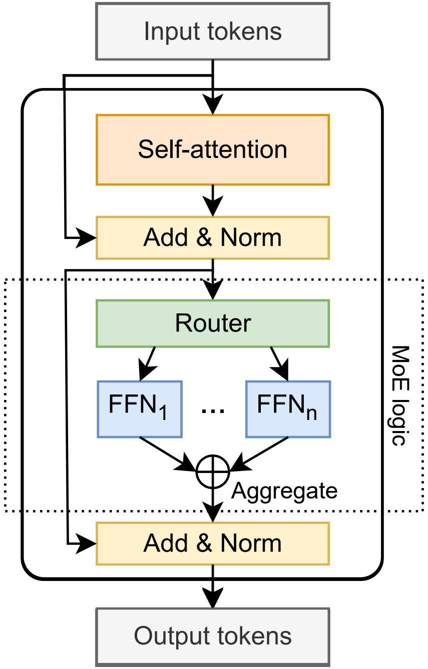
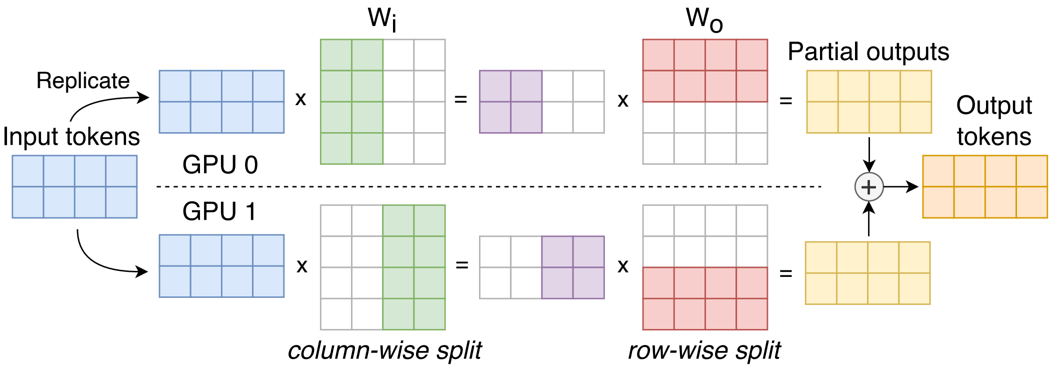
Mixture Of Experts

- MoE blocks in place of Linear layers
- Expert parallelism
- Expert imbalance ➤ workload imbalance





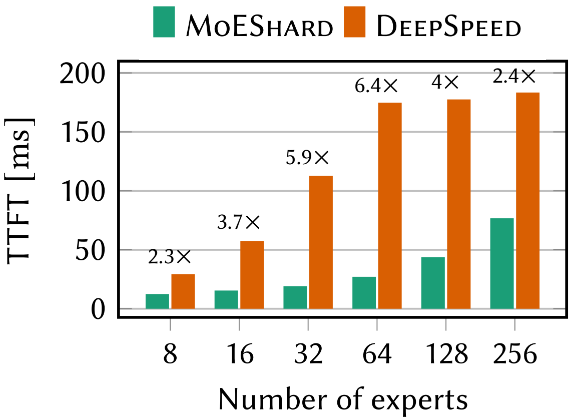
MoEShard



- Tensor Sharding
- Block Sparse Matrix Multiplication[1]

- Perfect load balancing
- No token dropping

- Up to 6.4× faster than DeepSpeed



[1] Gale, Trevor, et al., MLSys 2023