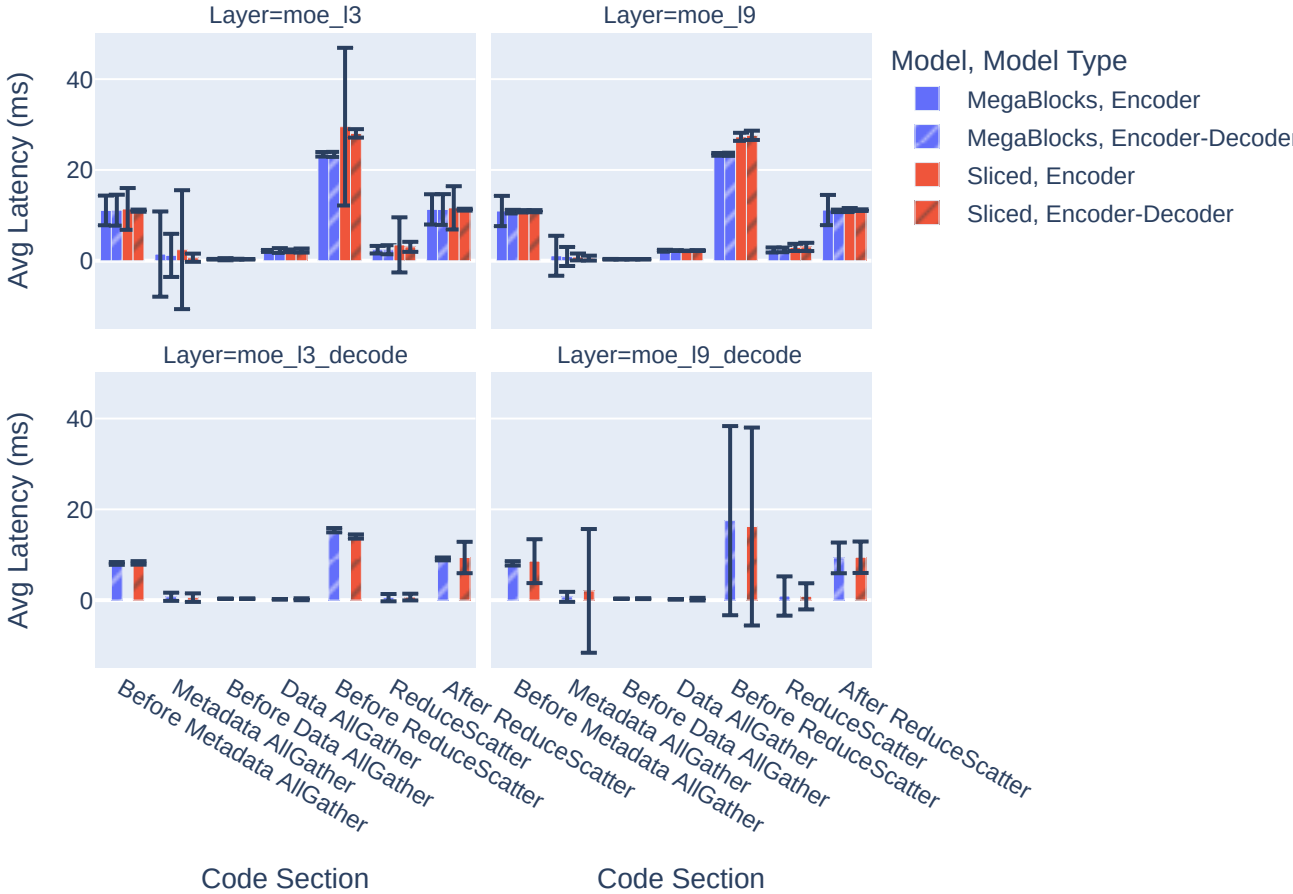
Layer=moe_l3
Layer=moe_l9
Layer=moe_l3_decode
Layer=moe_l9_decode
Avg Latency (ms)
40
20
0
Model, Model Type
MegaBlocks, Encoder
MegaBlocks, Encoder-Decoder
Sliced, Encoder
Sliced, Encoder-Decoder
Before Metadata AllGather
Metadata AllGather
Before Data AllGather
Data AllGather
Before ReduceScatter
ReduceScatter
After ReduceScatter
Code Section