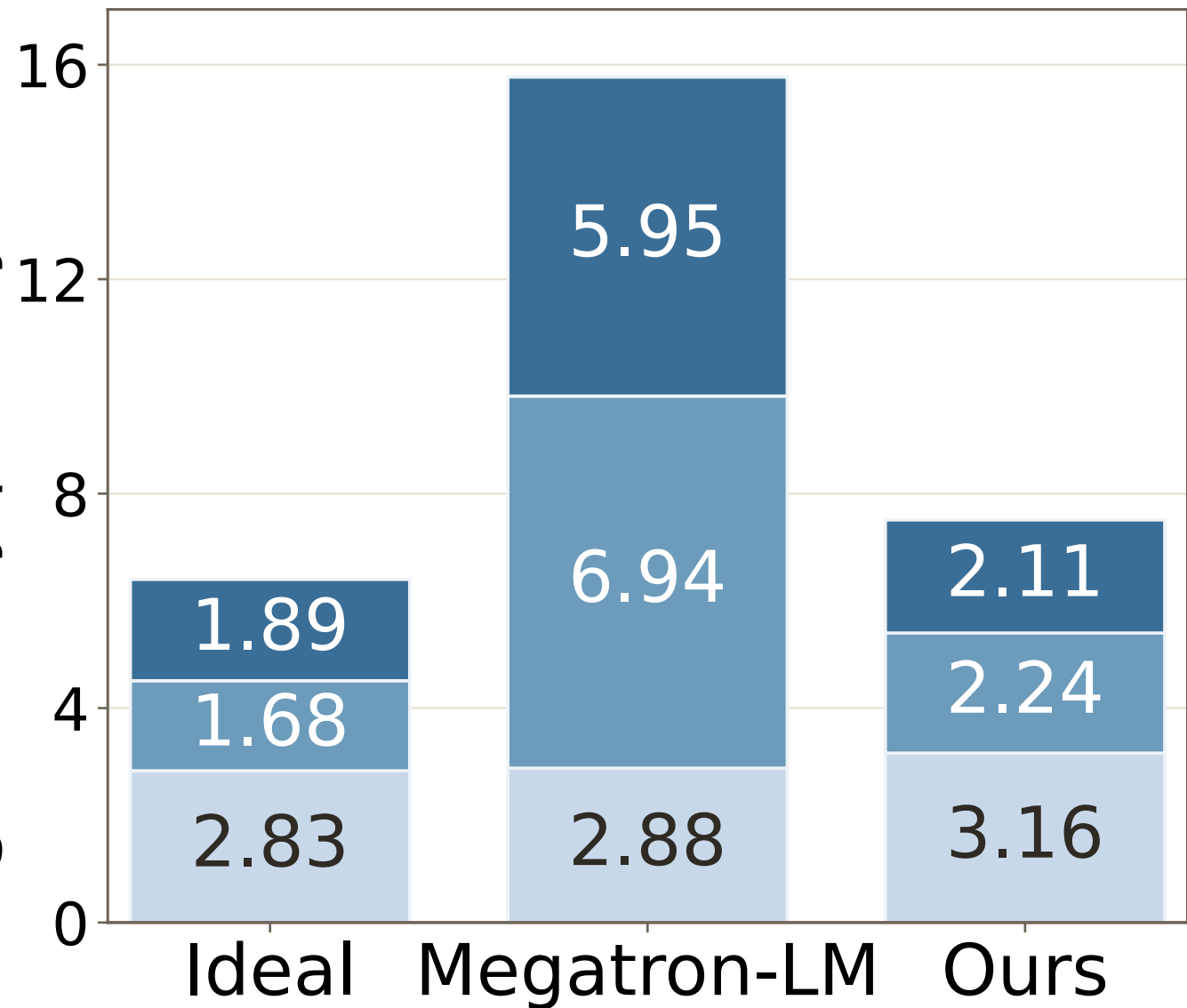


Avg Latency per Layer (ms)

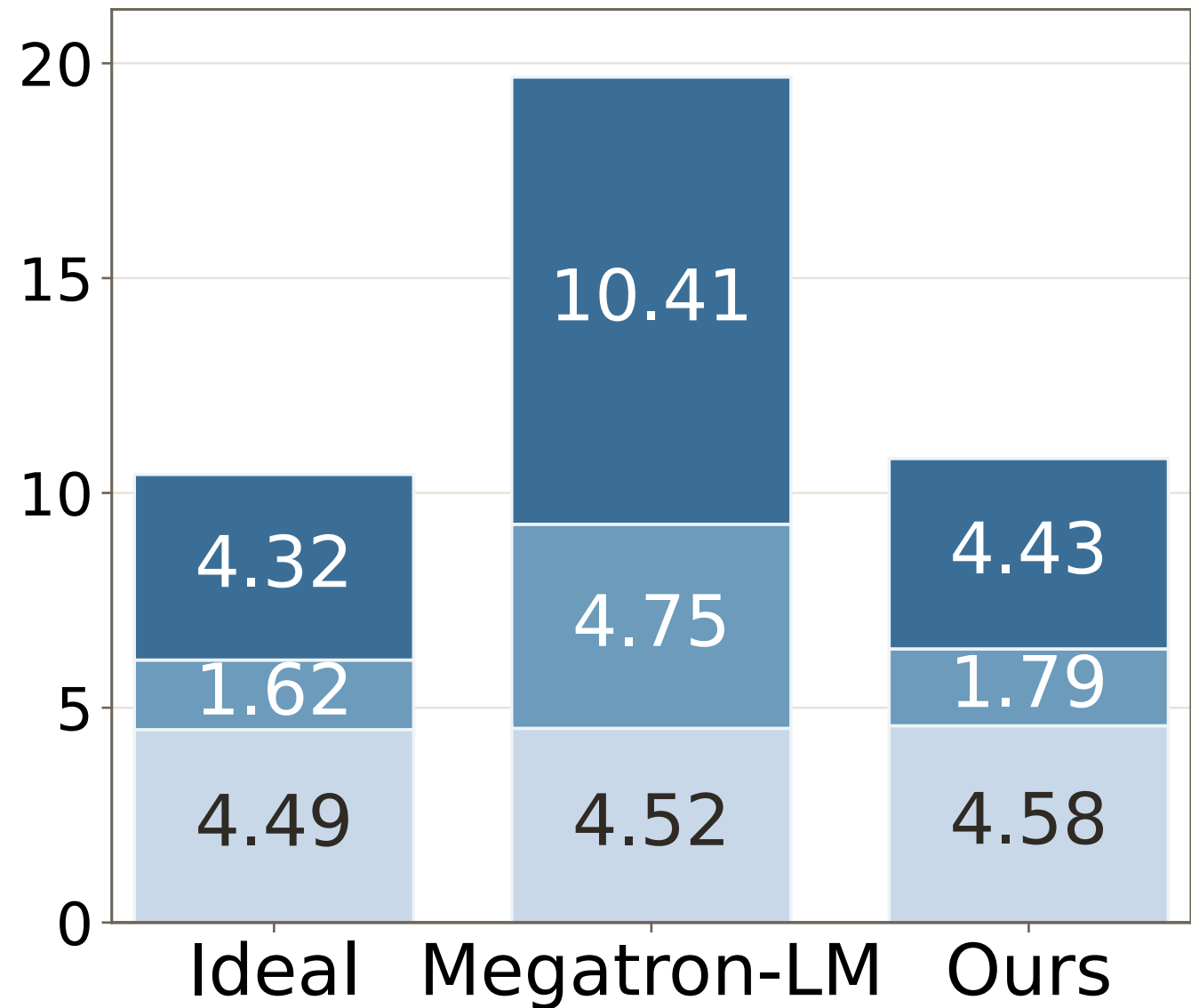
MoE Compute

Token All-to-All

Others



(a) Forward



(b) Backward