



Fiddler

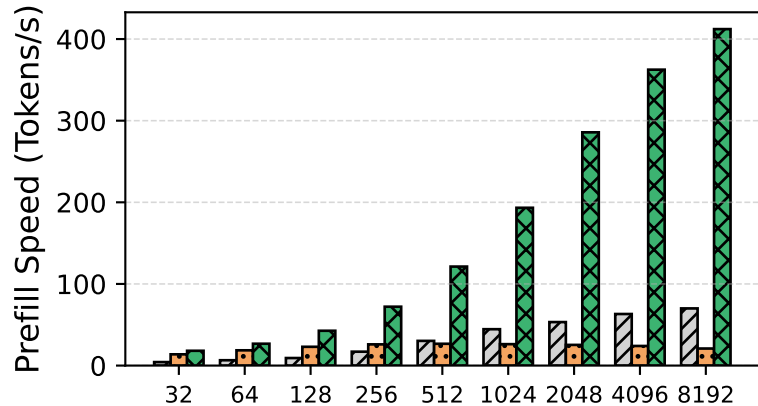


Llama.cpp

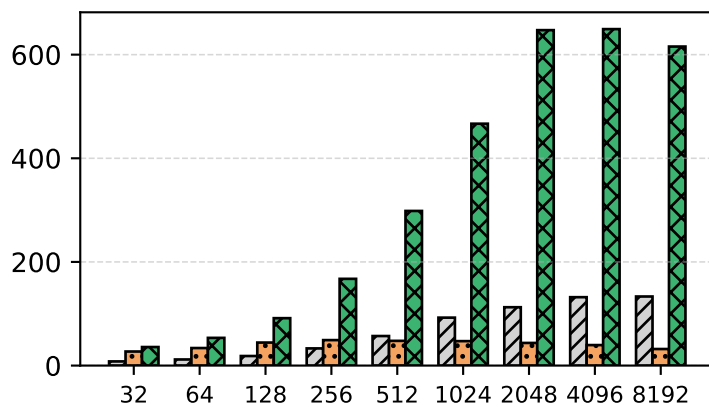


KTransformers

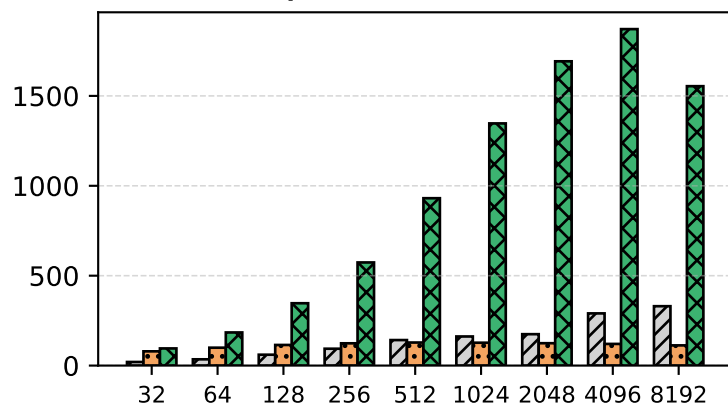
DS-3 BF16/FP16



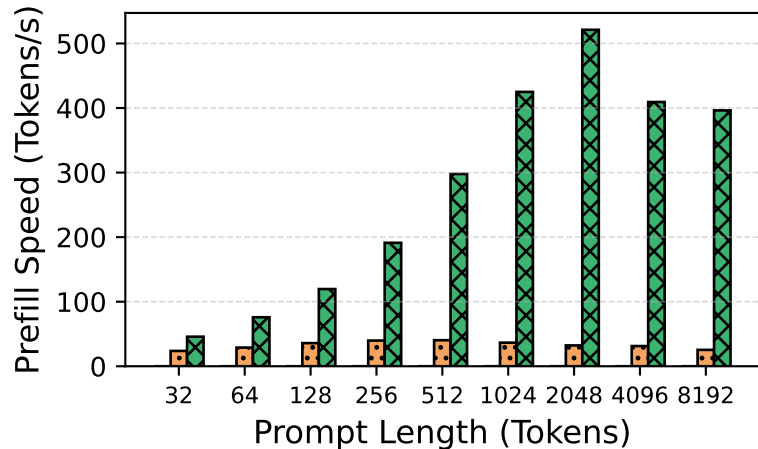
DS-2 BF16/FP16



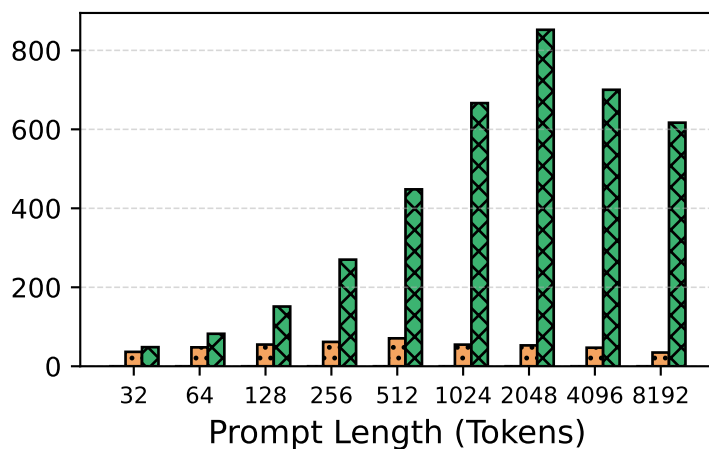
QW-2 BF16/FP16



DS-3 Int4



DS-2 Int8



QW-2 Int8

