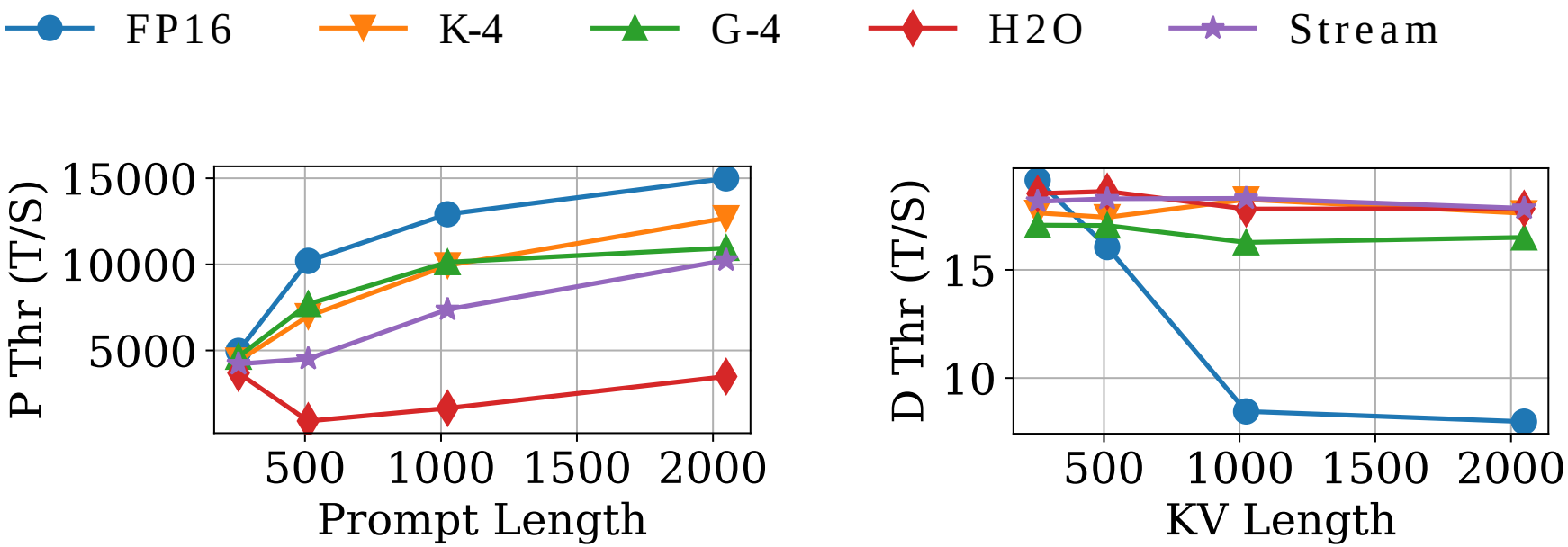

Prefill, Batch Size 1

Decoding, Batch Size 1

We run meta-llama/Llama-2-70b-hf on a GPU server of eight H800 GPUs with tensor parallel of 8.