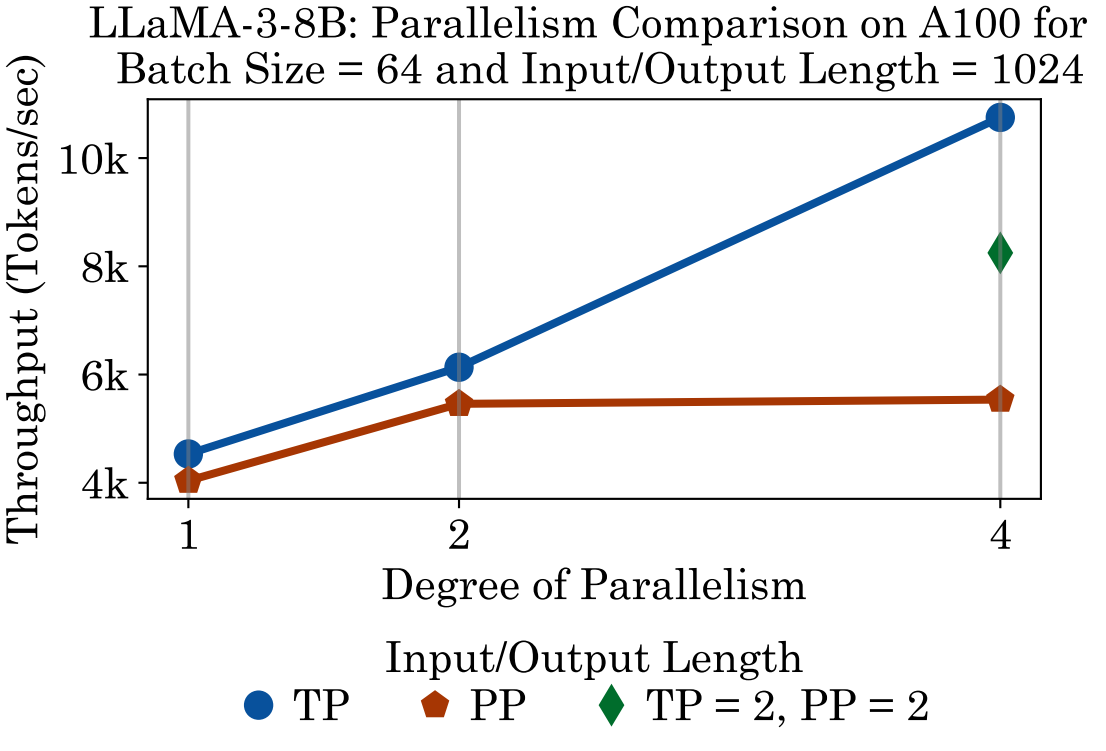
LLaMA-3-8B: Parallelism Comparison on A100 for
Batch Size = 64 and Input/Output Length = 1024
Throughput (Tokens/sec)
10k
8k
6k
4k
1
2
4
Degree of Parallelism
Input/Output Length
TP
PP
TP = 2, PP = 2