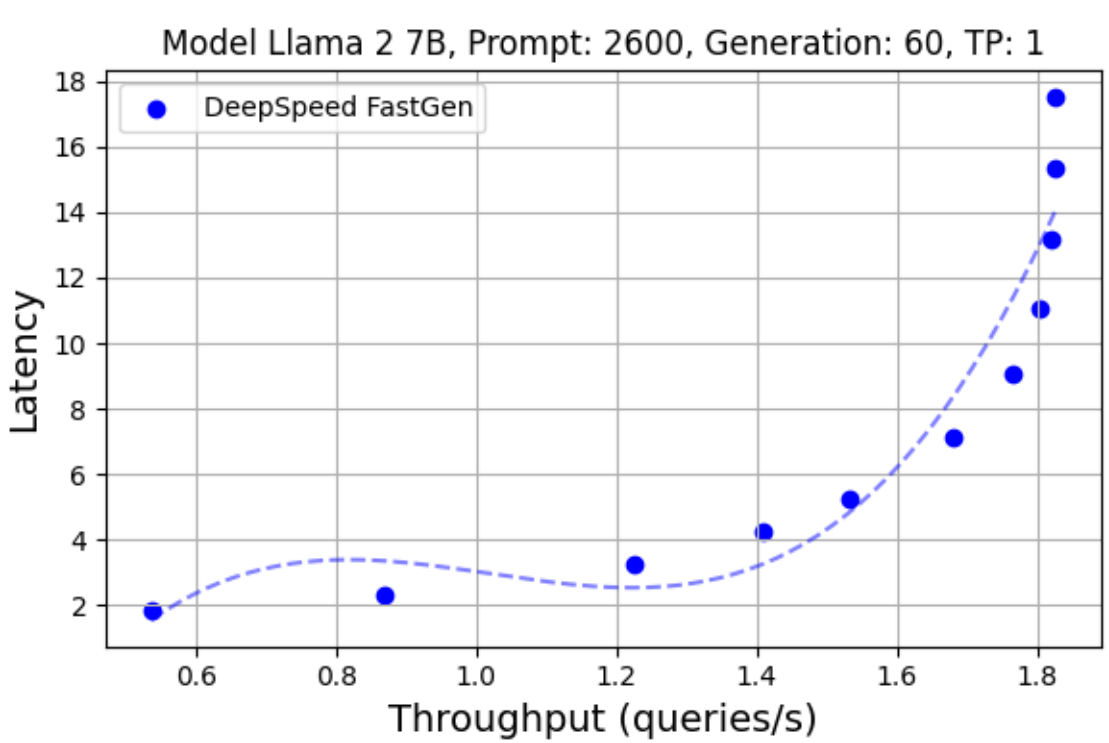
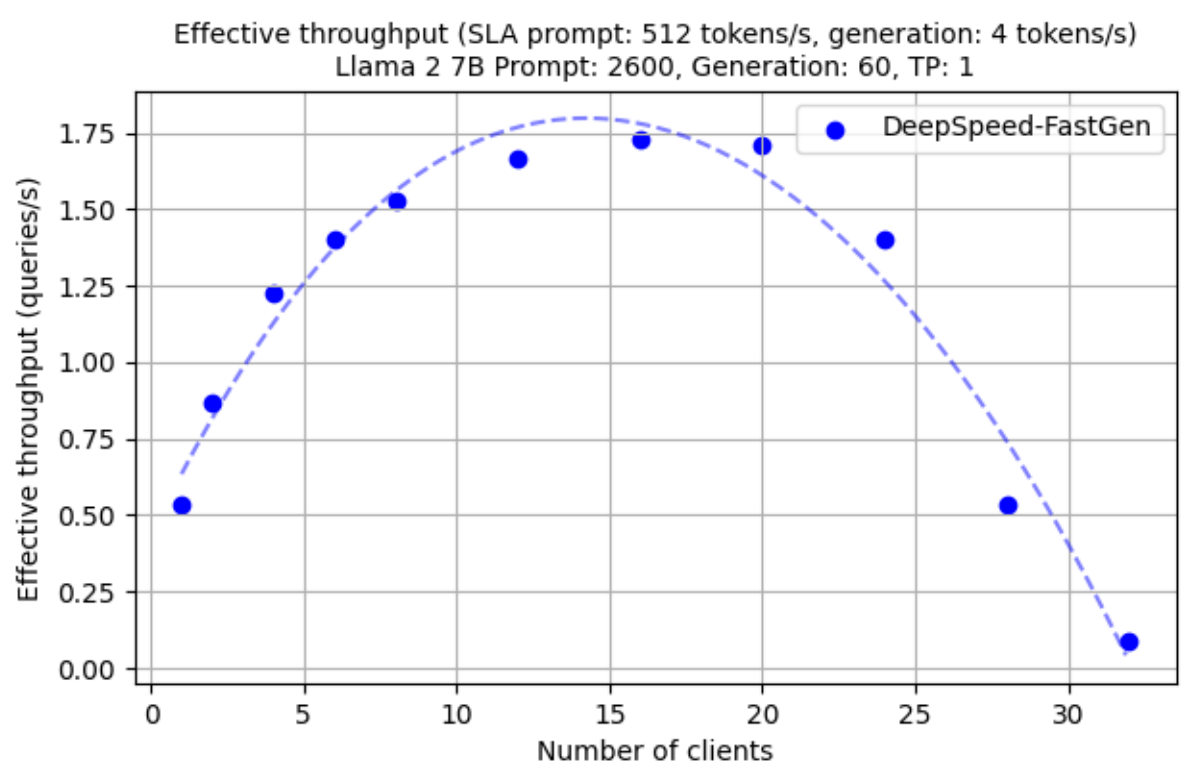

Model Llama 2 7B, Prompt: 2600, Generation: 60, TP: 1
DeepSpeed FastGen
Latency
Throughput (queries/s)
18
16
14
12
10
8
6
4
2
0.6
0.8
1.0
1.2
1.4
1.6
1.8


Effective throughput (SLA prompt: 512 tokens/s, generation: 4 tokens/s)
Llama 2 7B Prompt: 2600, Generation: 60, TP: 1
DeepSpeed-FastGen
Effective throughput (queries/s)
Number of clients
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
0
5
10
15
20
25
30