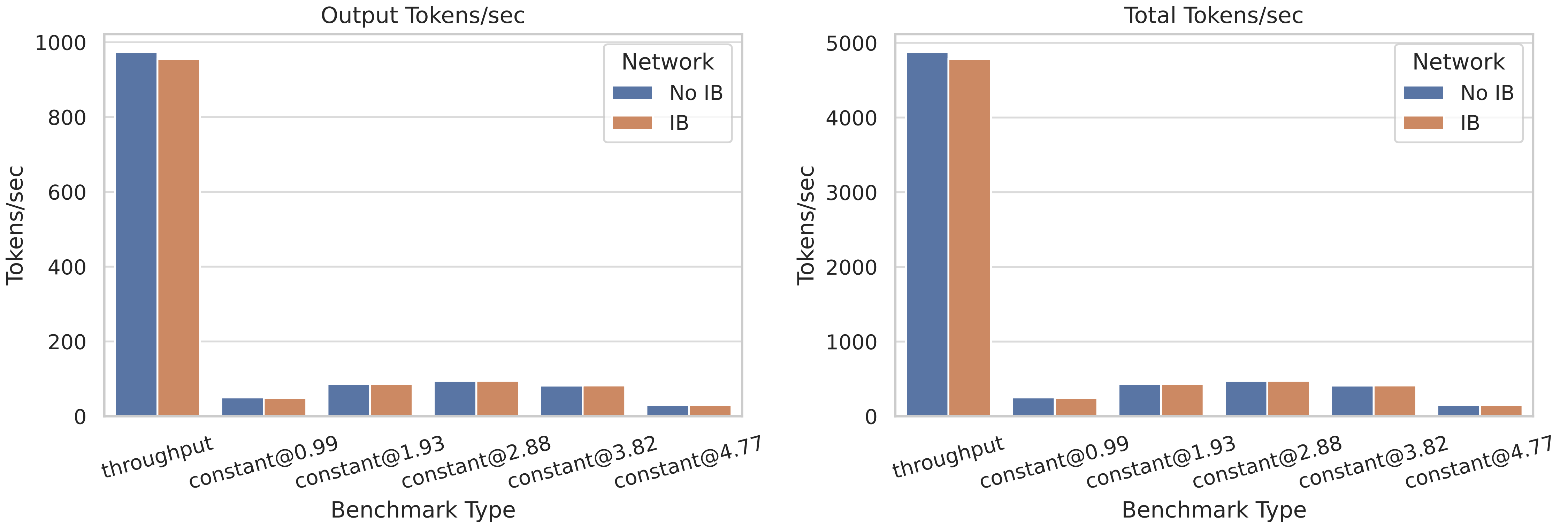
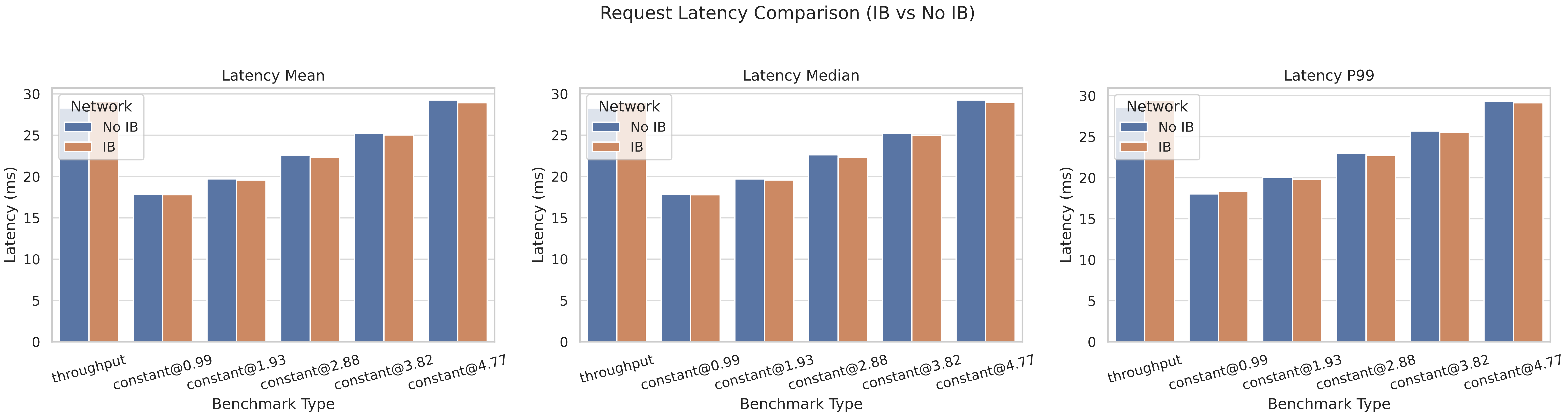
Token Throughput Comparison (IB vs No IB)

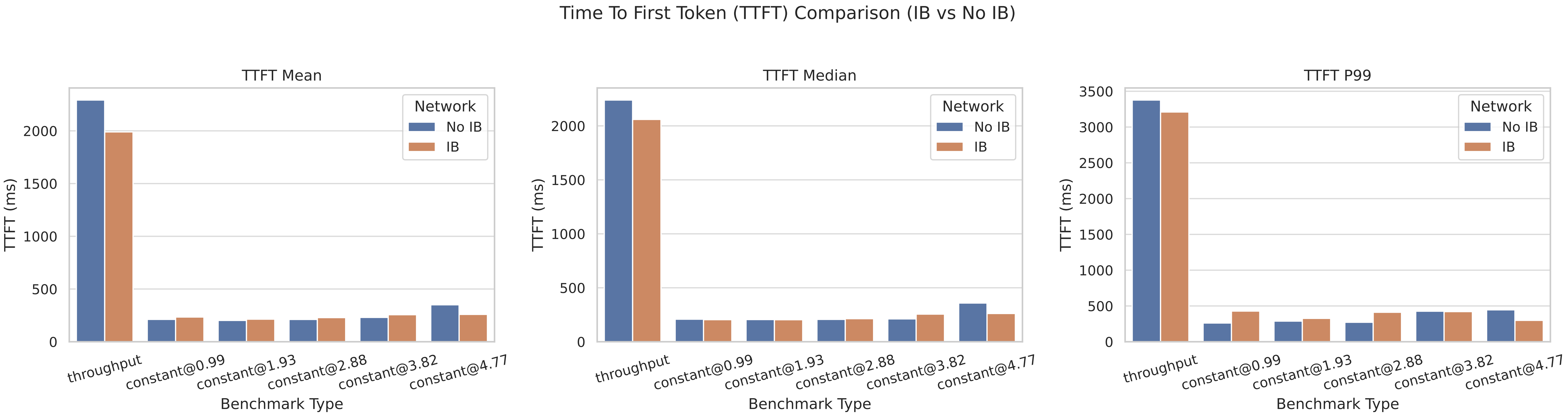


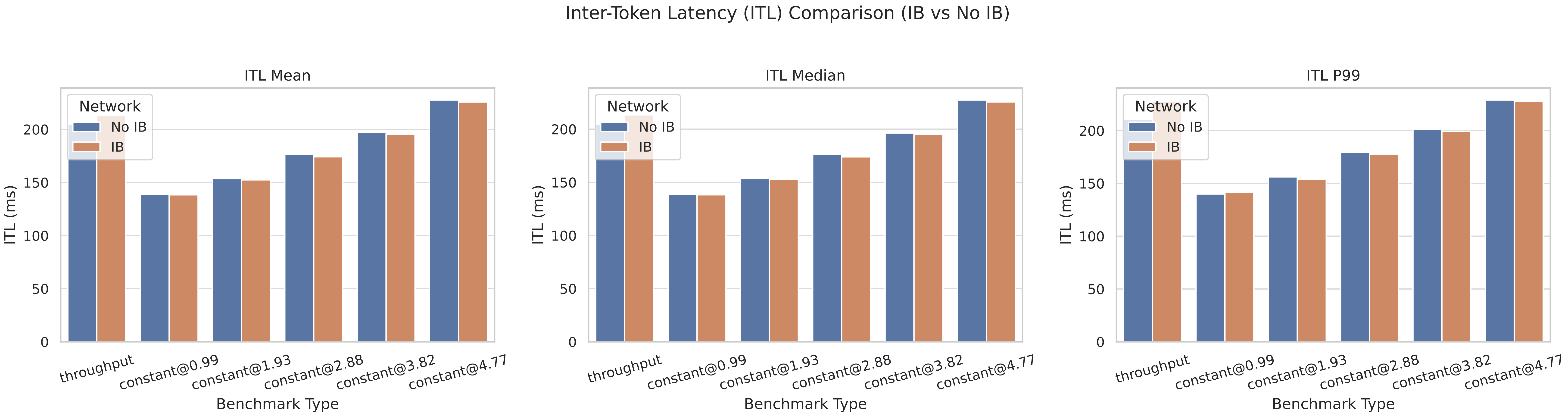
Output Tokens/sec

Total Tokens/sec

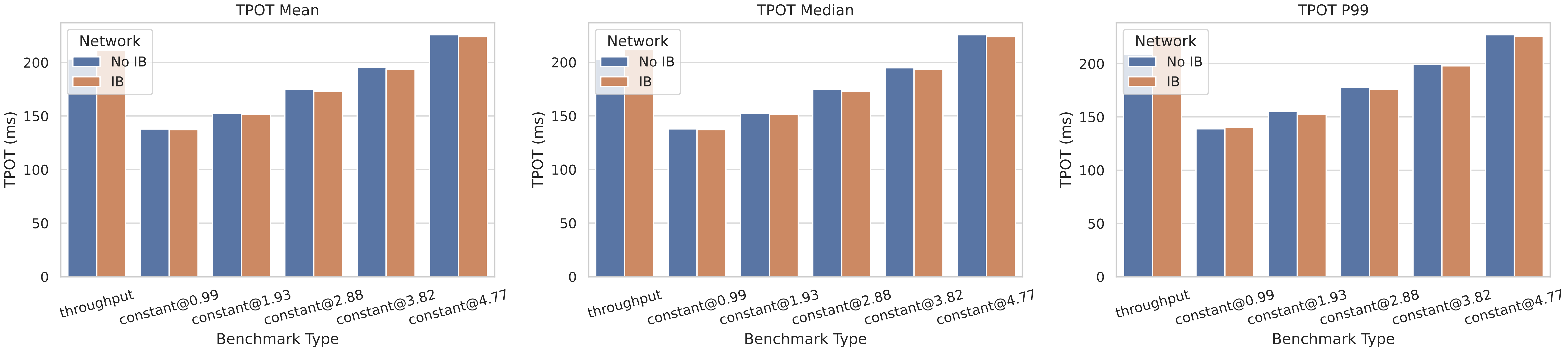


Request Latency Comparison (IB vs No IB)
Latency Mean
Latency Median
Latency P99
Network
No IB
IB
Latency (ms)
0
5
10
15
20
25
30
throughput
constant@0.99
constant@1.93
constant@2.88
constant@3.82
constant@4.77
Benchmark Type


Time To First Token (TTFT) Comparison (IB vs No IB)
TTFT Mean
TTFT Median
TTFT P99
Network
No IB
IB
TTFT (ms)
0
500
1000
1500
2000
2500
3000
3500
throughput
constant@0.99
constant@1.93
constant@2.88
constant@3.82
constant@4.77
Benchmark Type


Inter-Token Latency (ITL) Comparison (IB vs No IB)
ITL Mean
ITL Median
ITL P99
Network
No IB
IB
ITL (ms)
0
50
100
150
200
throughput
constant@0.99
constant@1.93
constant@2.88
constant@3.82
constant@4.77
Benchmark Type

Time Per Output Token (TPOT) Comparison (IB vs No IB)



TPOT Mean

TPOT Median

TPOT P99