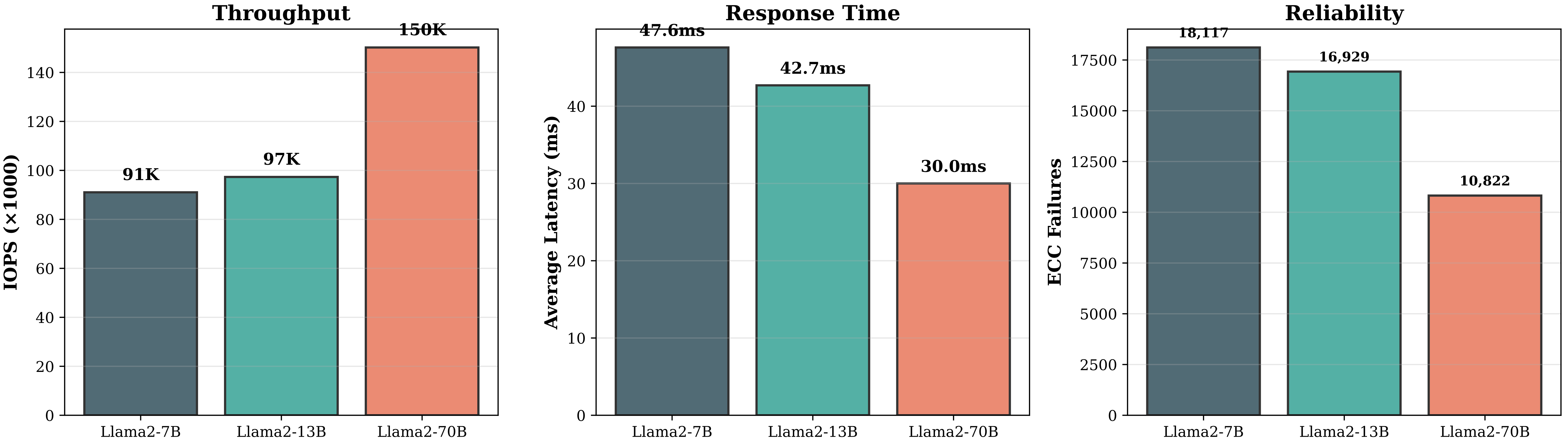
Baseline Performance Comparison (10K tokens)
Larger Models → Better Parallelism


Throughput
IOPS (×1000)
91K
97K
150K
Llama2-7B
Llama2-13B
Llama2-70B
Response Time
Average Latency (ms)
47.6ms
42.7ms
30.0ms
Llama2-7B
Llama2-13B
Llama2-70B
Reliability
ECC Failures
Llama2-7B
Llama2-13B
Llama2-70B