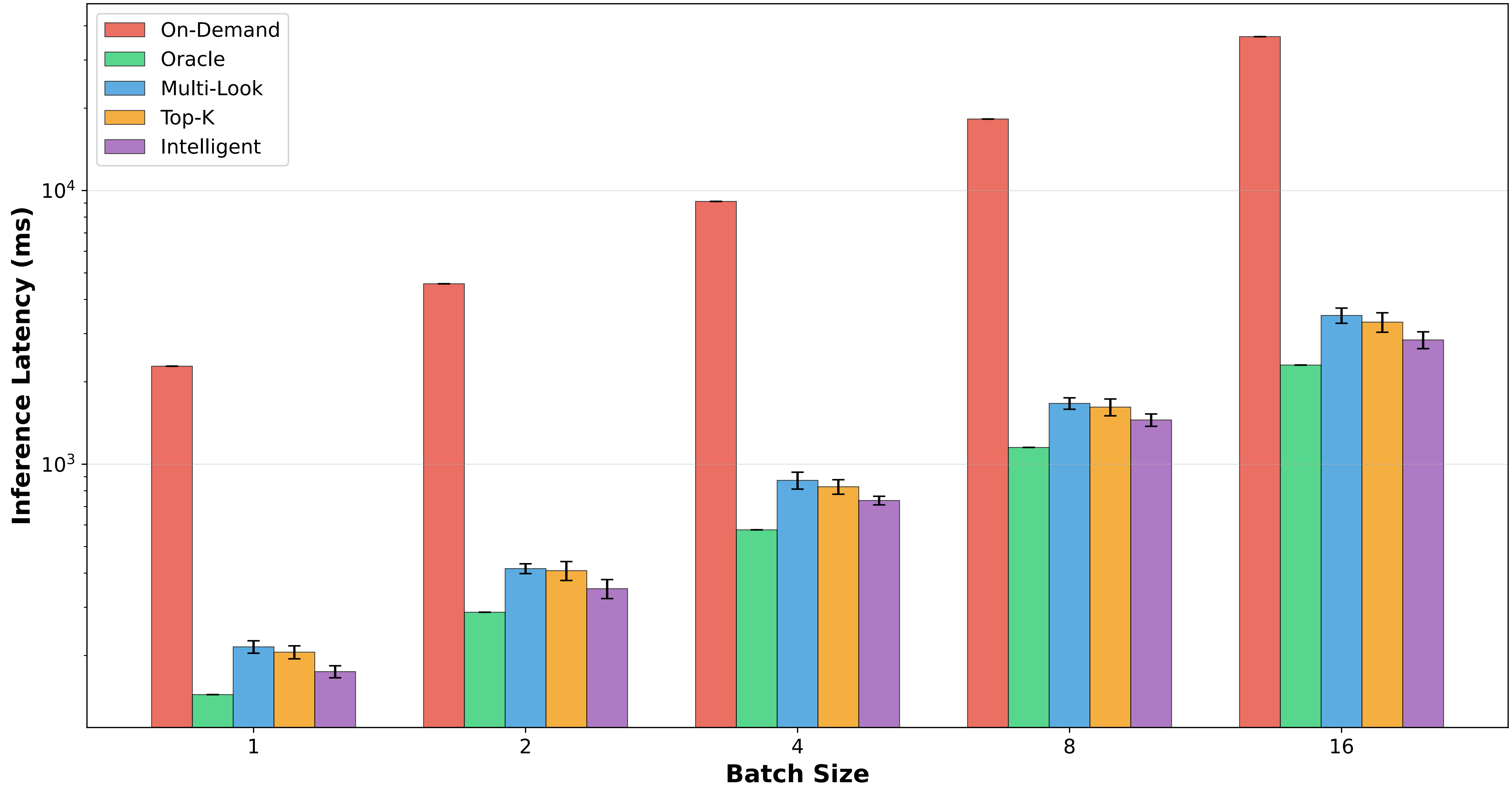
Inference Latency vs Batch Size


On-Demand
Oracle
Multi-Look
Top-K
Intelligent
Inference Latency (ms)
10^4
10^3
Batch Size