

# Inference Latency vs. Number of Servers (BERT-large)

