Inference Latency vs. Number of Servers (TinyBERT-6l)