Inference Latency vs. Number of Servers (ALBERT)