Inference Latency vs. Network Bandwidth (BERT-large)