

Inference Latency vs. Network Bandwidth (TinyBERT-4I)

