

Inference Latency vs. Number of Servers (TinyBERT-4l)

