

Inference Latency vs. Number of Servers (ViT-small)

