Inference Latency vs. Number of Servers (InceptionV3)