Inference Latency Distribution