Inference latency: Baseline vs INT4