Inference Latency vs. Number of Servers (ALBERT-base)

Legend: OCC, DINA, MEIDA, Ours

Y-axis: Inference Latency (ms)

X-axis: Number of Servers

Lower is better