## System Architecture during Inference

