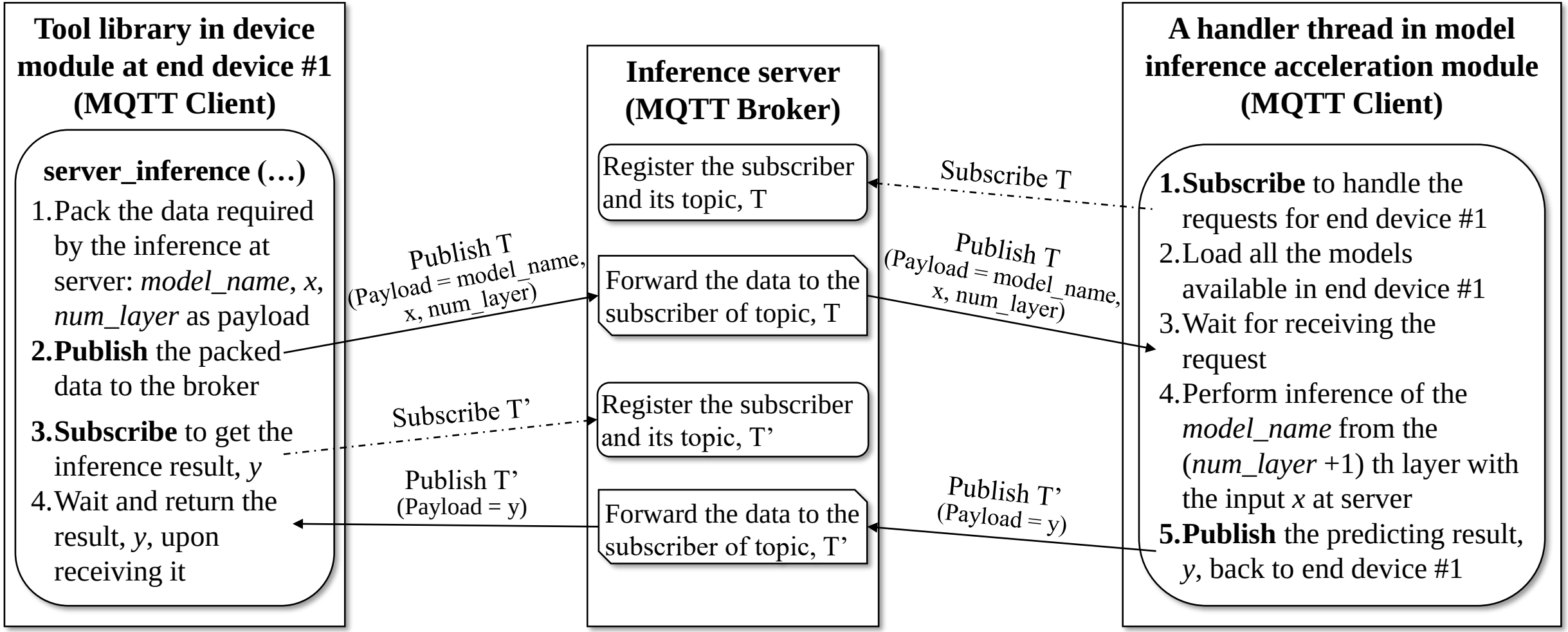
Tool library in device module at end device #1 (MQTT Client)
server_inference (…)
1. Pack the data required by the inference at server: model_name, x, num_layer as payload
2. Publish the packed data to the broker
3. Subscribe to get the inference result, y
4. Wait and return the result, y, upon receiving it
Inference server (MQTT Broker)
Register the subscriber and its topic, T
Forward the data to the subscriber of topic, T
Register the subscriber and its topic, T'
Forward the data to the subscriber of topic, T'
A handler thread in model inference acceleration module (MQTT Client)
1. Subscribe to handle the requests for end device #1
2. Load all the models available in end device #1
3. Wait for receiving the request
4. Perform inference of the model_name from the (num_layer +1) th layer with the input x at server
5. Publish the predicting result, y, back to end device #1
Publish T (Payload = model_name, x, num_layer)
Subscribe T
Publish T (Payload = model_name, x, num_layer)
Subscribe T'
Publish T' (Payload = y)
Publish T' (Payload = y)