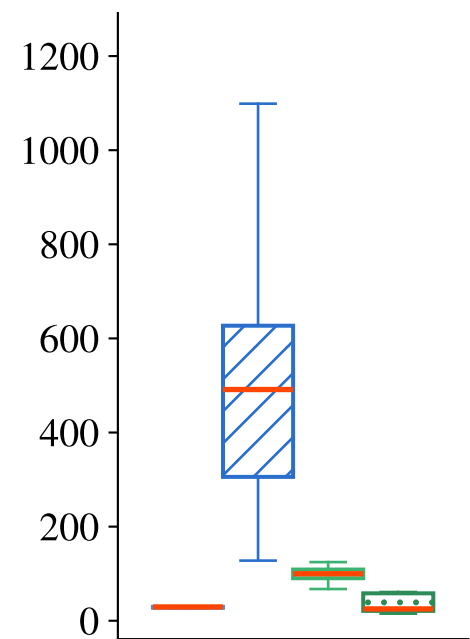
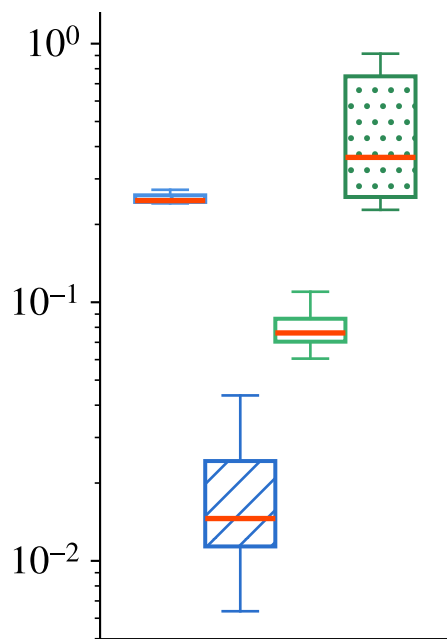


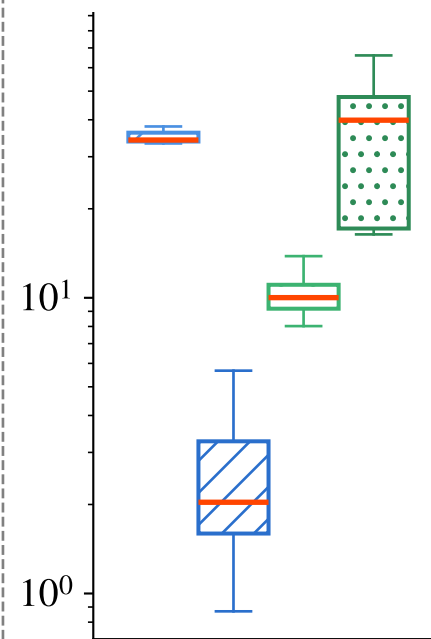
# Throughput (Req/s)



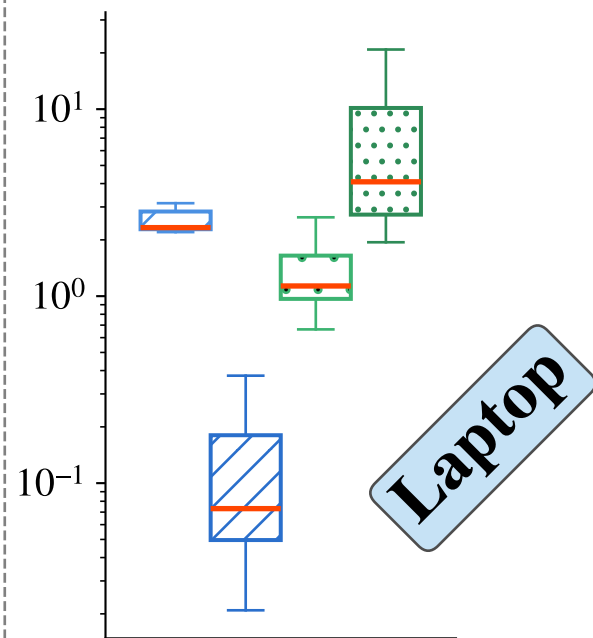
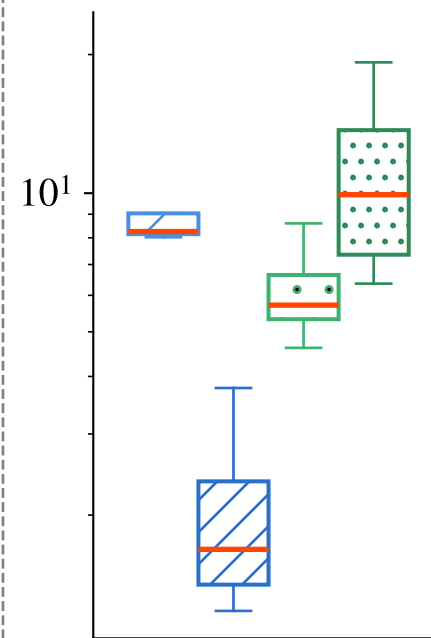
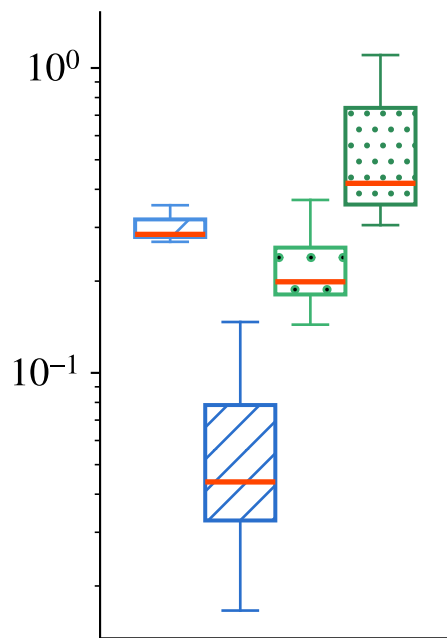
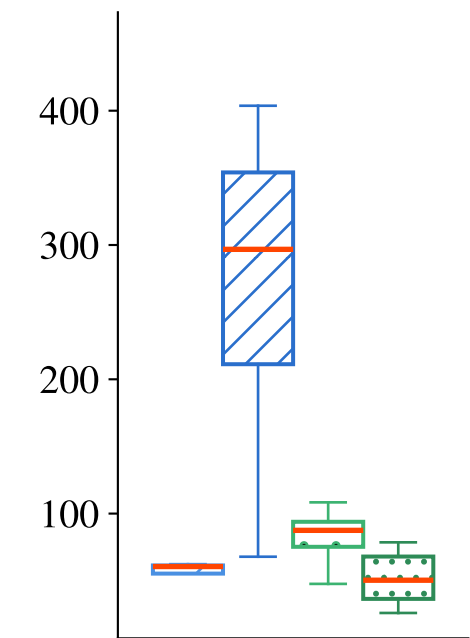
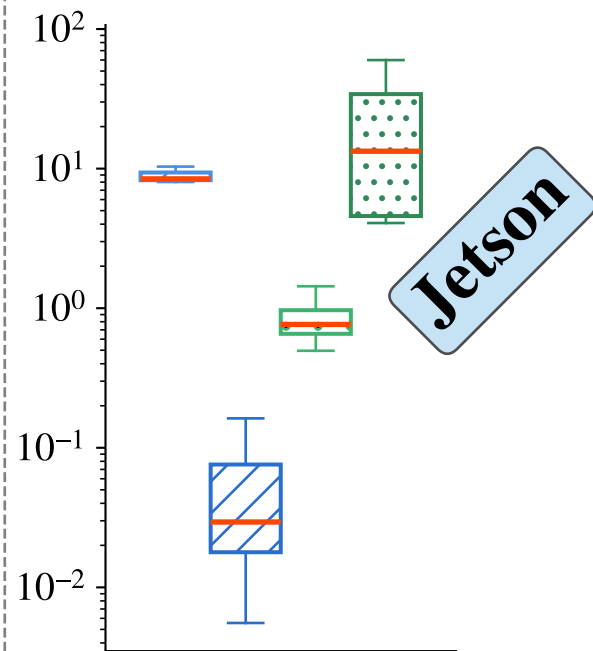
# Energy (J/Req)



# Latency (ms/Req)



# Energy $\times$ Delay (J $\cdot$ ms)



$R_{FT}$   
 $R_S$   
 $R_H$   
 $R_{LLM}$

$R_{FT}$   
 $R_S$   
 $R_H$   
 $R_{LLM}$

$R_{FT}$   
 $R_S$   
 $R_H$   
 $R_{LLM}$

$R_{FT}$   
 $R_S$   
 $R_H$   
 $R_{LLM}$