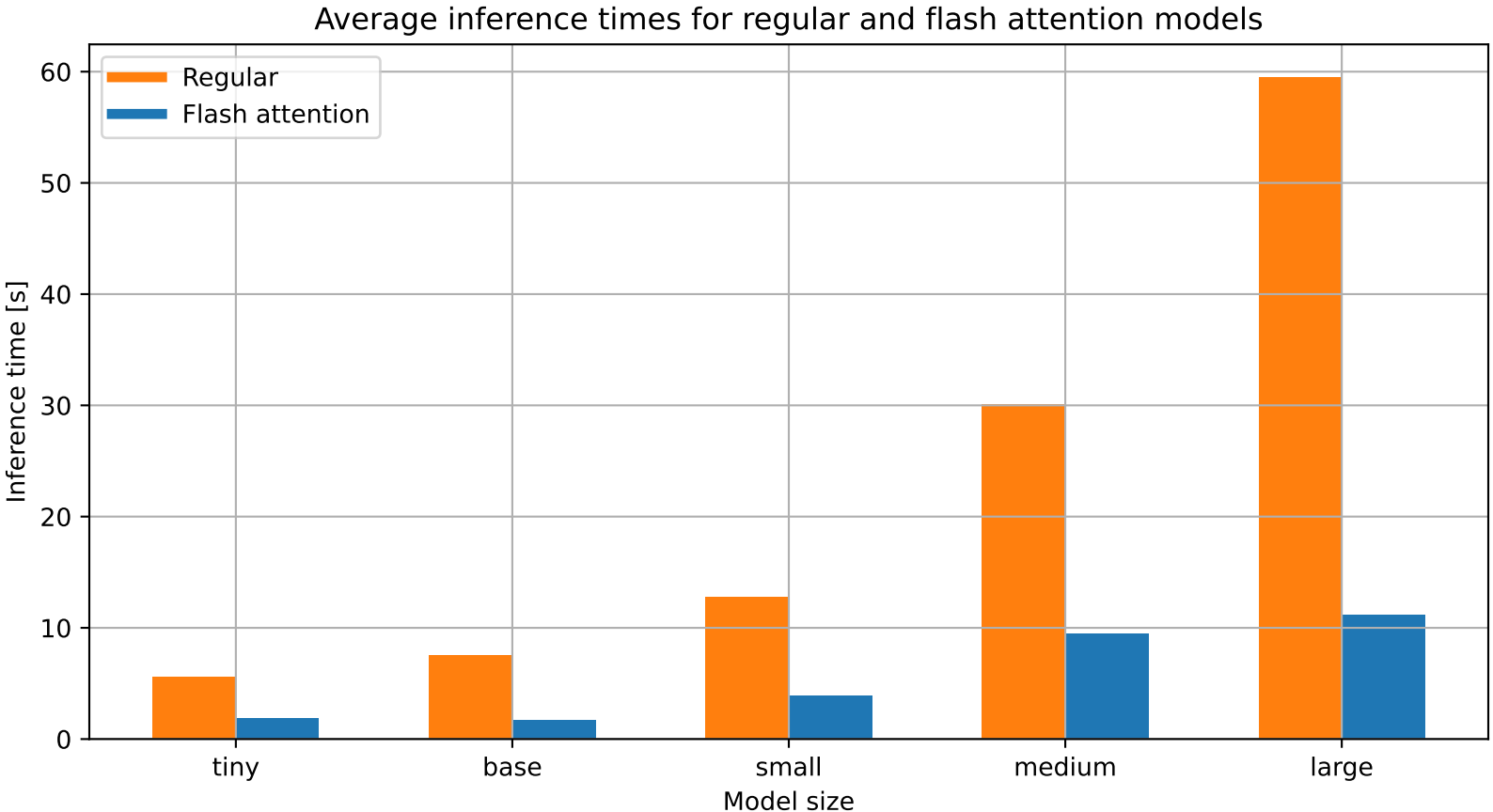
Average inference times for regular and flash attention models
Regular
Flash attention
Inference time [s]
0
10
20
30
40
50
60
tiny
base
small
medium
large
Model size