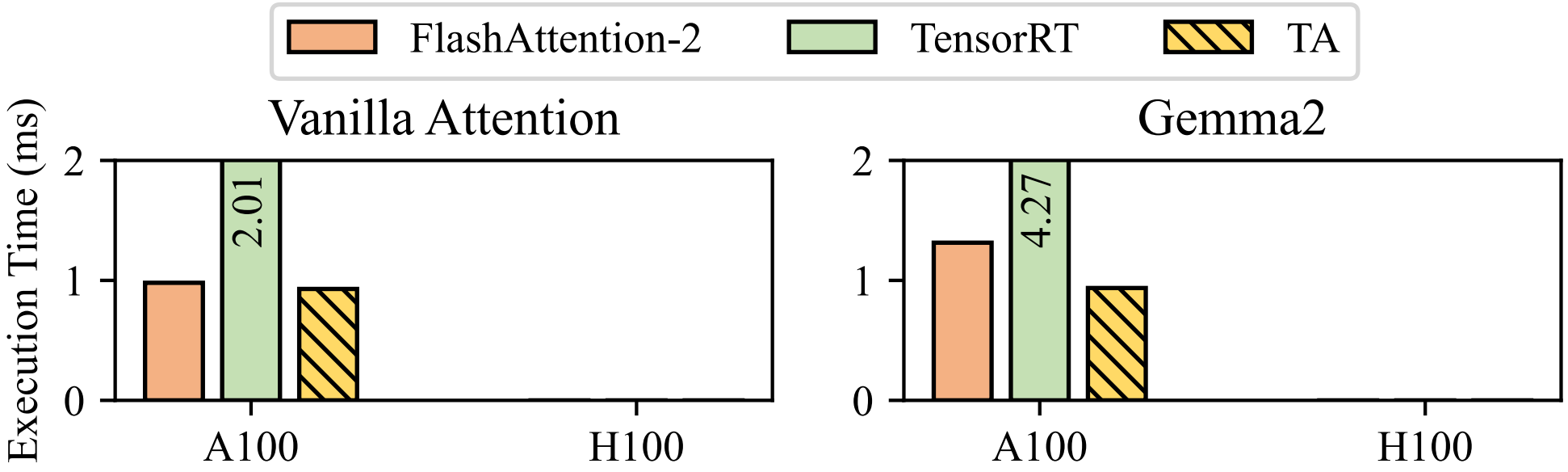
FlashAttention-2
TensorRT
TA
Vanilla Attention
Gemma2
Execution Time (ms)
2
1
0
2.01
4.27
A100
H100