Todo：Solve the memory alignment problem.Fix the problem caused by optimization mechanism of FlashAttention for discontinuous memory distribution.Benchmark the memory optimization within the envelope under discontinuous memory distribution.

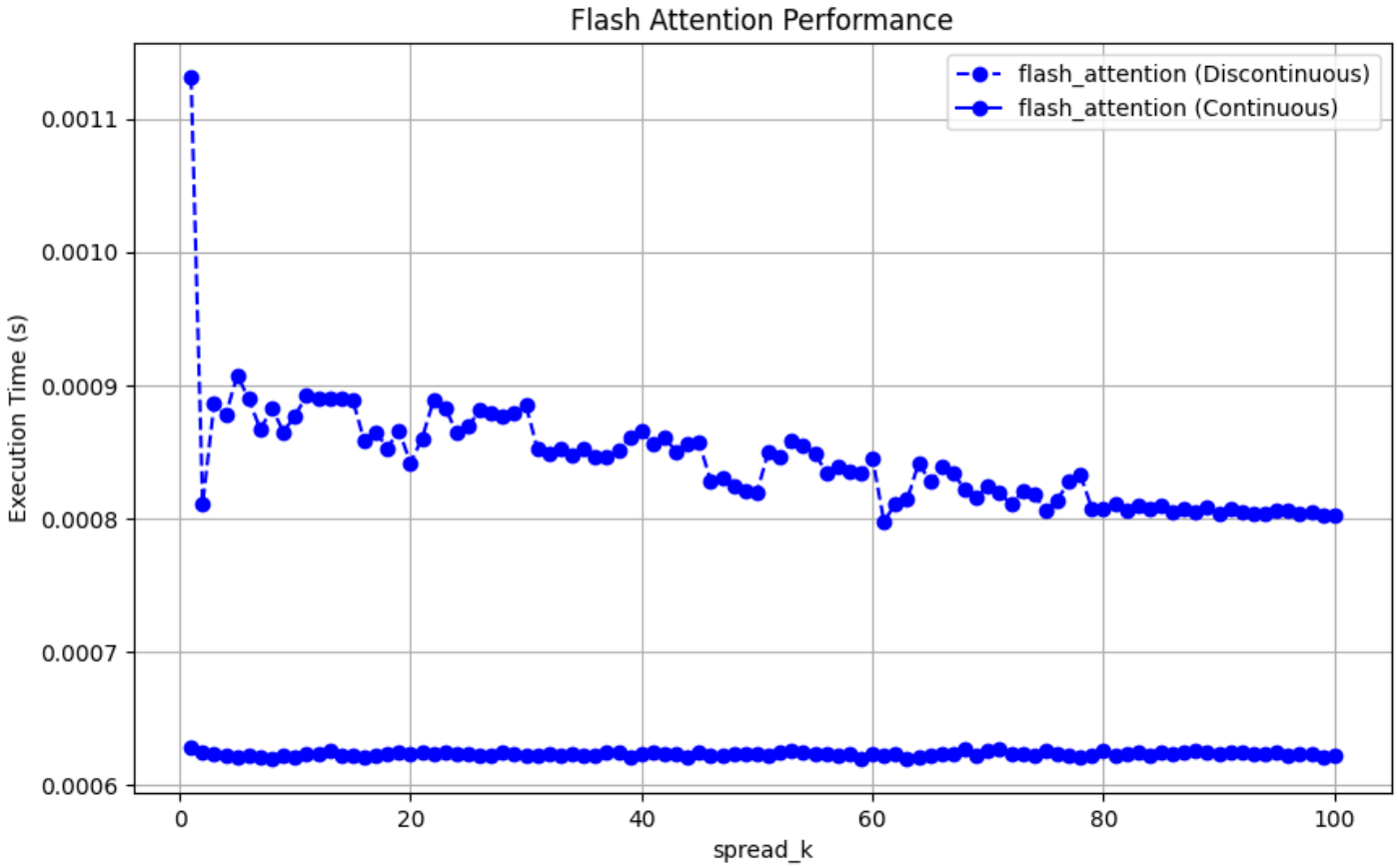
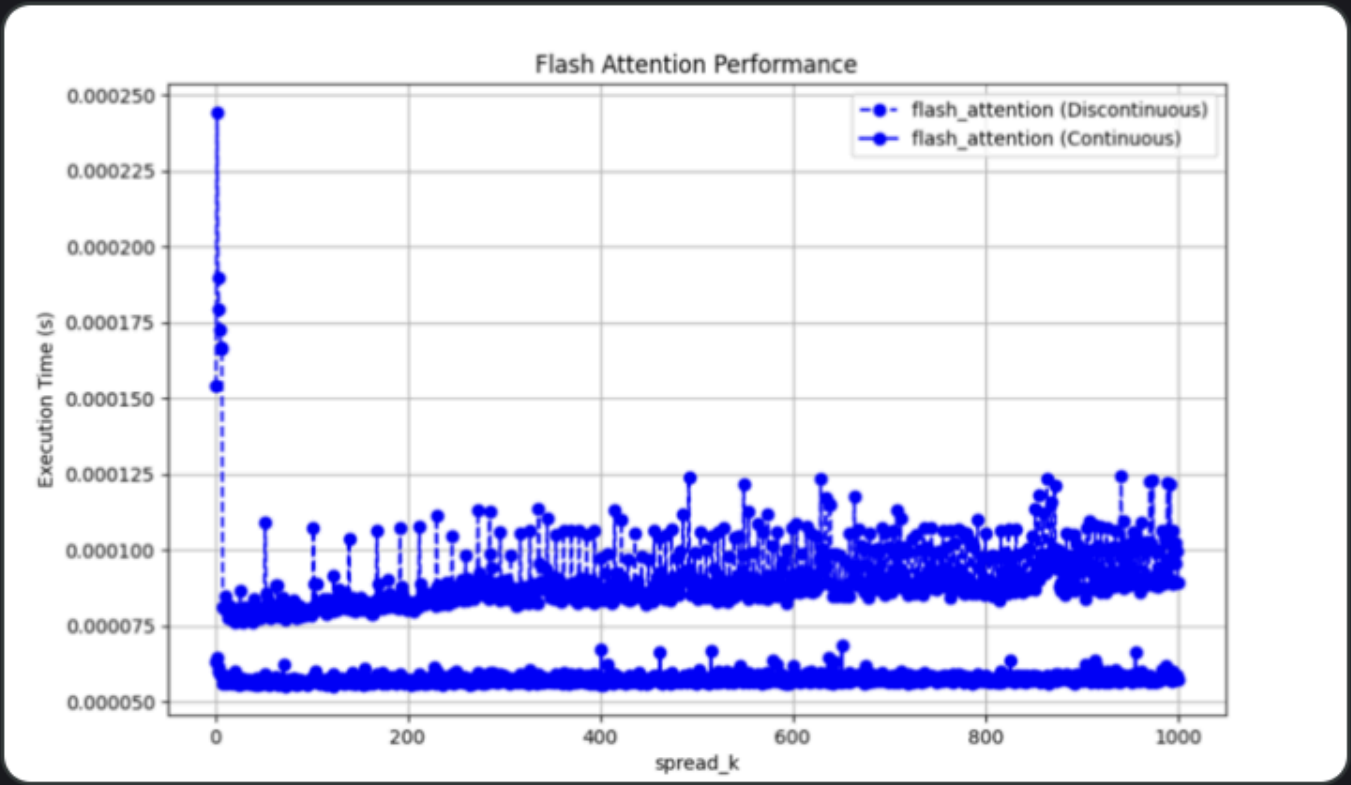


image.png



1

:42 image.png