Matrix optimization on GPU

JX-Ma

2024/8/3

Tiling

The matrix tiling is a solution that transform big problem into small ones. In CPU, we hope to keep the locality of computation,so we need to use adequately cache of L1 and L2. In GPU, we also need to allocate a major problem reasonably among different threads on this basis .Afterwards, determine the block size based on some hardware information.

Thread level optimization

There are not many optimizations that can be done for a thread, because GPUs are scheduled with one warp (i.e. 32 threads), so many single thread based optimizations, such as memory access optimization, cannot be directly applied to GPUs. One of the few optimization methods worth mentioning is whether a single thread should use vector inner product or vector outer product and double buffer during computation. But in essence, vector outer product cannot strictly be considered an optimization, because the compiler can help with this step during the compilation stage.

Warp level optimization

Based on some hardware parameters of the GPU, such as Global Memory and Shared Memory, optimization at the warp level has not been thoroughly examined here.

summary

This material explains some methods for optimizing matrices on GPUs and provides relevant code. And this article also provides the code for the optimization process, but the code has not been implemented in our own environment yet.