Triton

高层次Kernel开发语言Triton

The aim of Triton is to provide an open-source environment to write fast code at higher productivity than CUDA, but also with higher flexibility than other existing DSLs.

https://github.com/openai/triton

https://triton-lang.org/main/index.html

Triton更像是一个面向AI加速器算子开发的领域开发语言，为了能够将用户使用Triton语言开发的kernel映射到具体硬件上的执行码，需要设计开发相应的Triton compiler来完成这层映射。所以当我们说Triton的时候，其实隐指了**Triton**语言**+Triton**编译器这两个事物的综合体。

Triton的核心设计思想**----Block-wise**编程，**Block**上面的归用户，**Block**内部的归**Triton compiler**自动化处理。相应地，Block内部的优化细节，也交由Triton compiler处理了。

优化Passes在Triton当前的实现里主要包括了NV GPU计算kernel优化的一些常见技巧，包括用于辅助向量化访存的coalescing、用于缓解计算访存差异的pipeline/prefetch，用于避免shared memory访问bank-conflict的swizzling。

Revision #1
Created 11 January 2025 09:46:28 by Colin
Updated 12 January 2025 06:35:20 by Colin