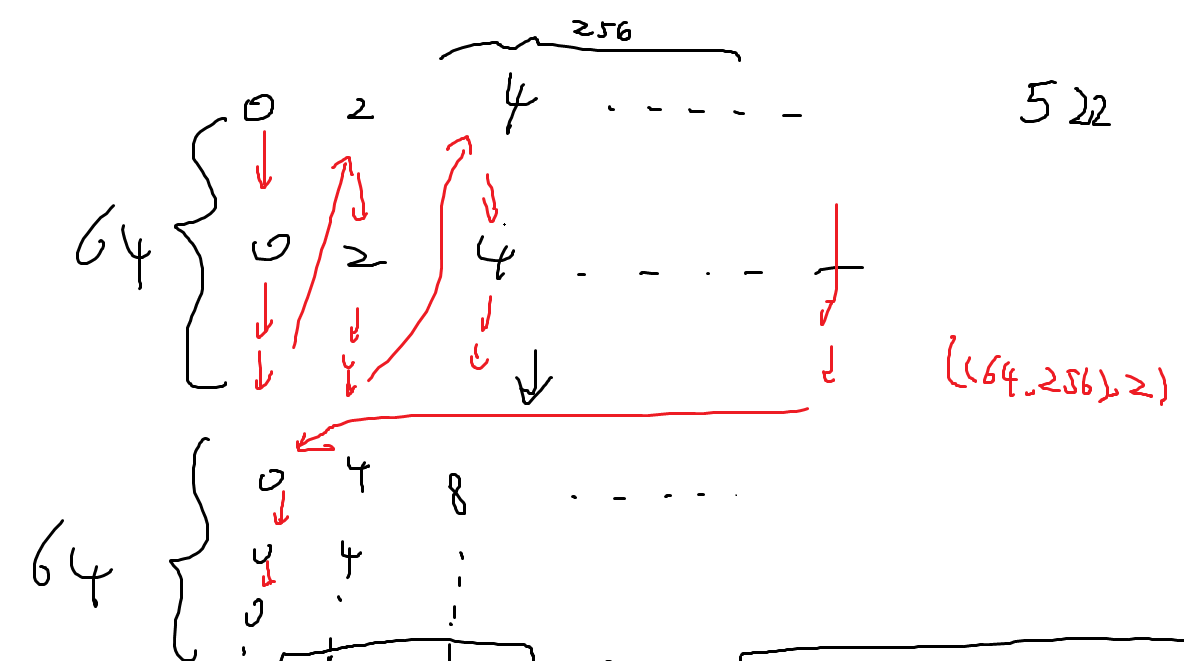
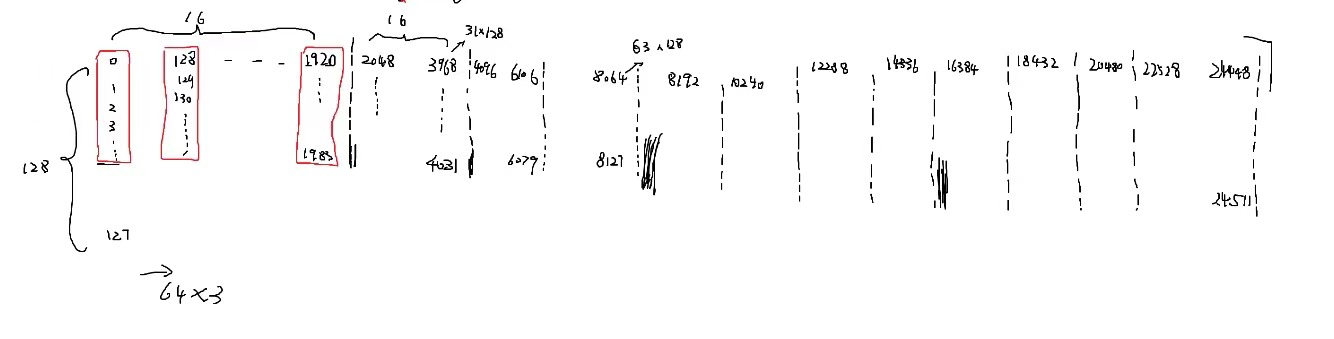
**问题描述：**

GEMM0的结果存储到共享内存的方式如下：

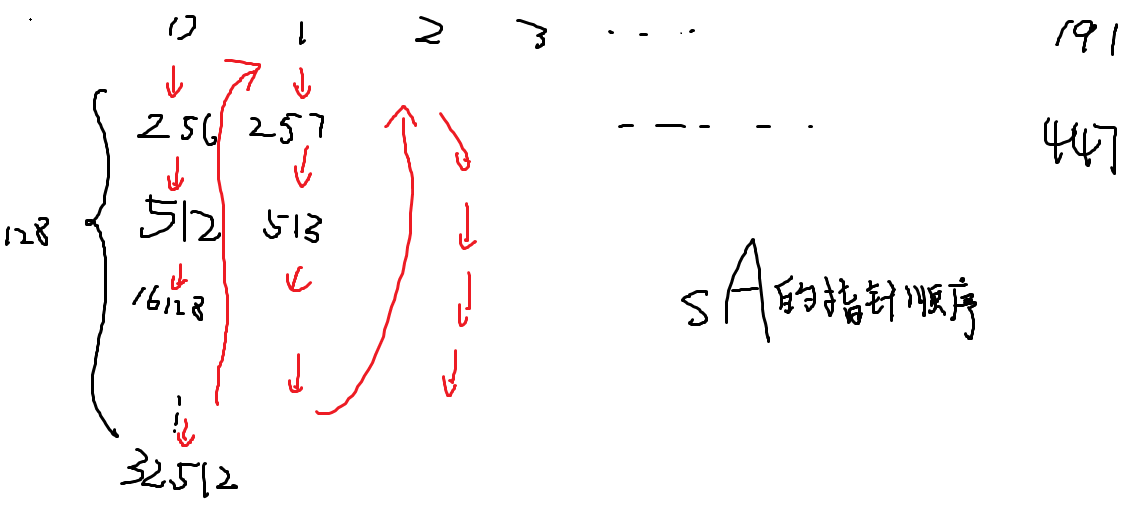
( (\_64,\_256),\_2),\_1,\_1,\_1):( (\_1,\_64),\_16384),\_0,\_0,\_0) bSG\_sD



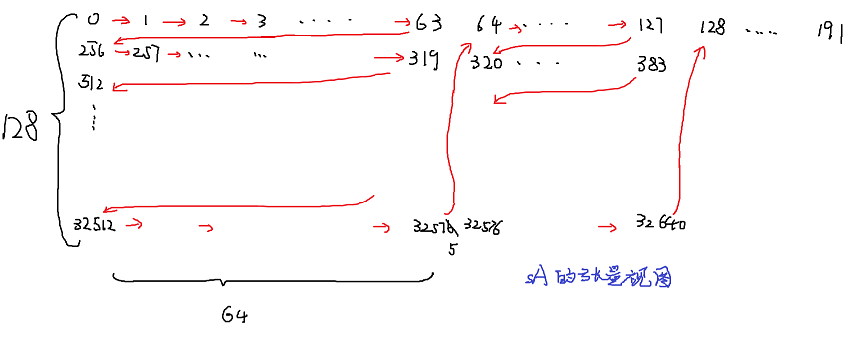
这里的箭头顺序是在内存中排布的顺序。



但是GEMM1的读取sA，在内存中排布如下：



就是说，sA原先是不希望打断上下两端的。然后尺寸变换方式如下：



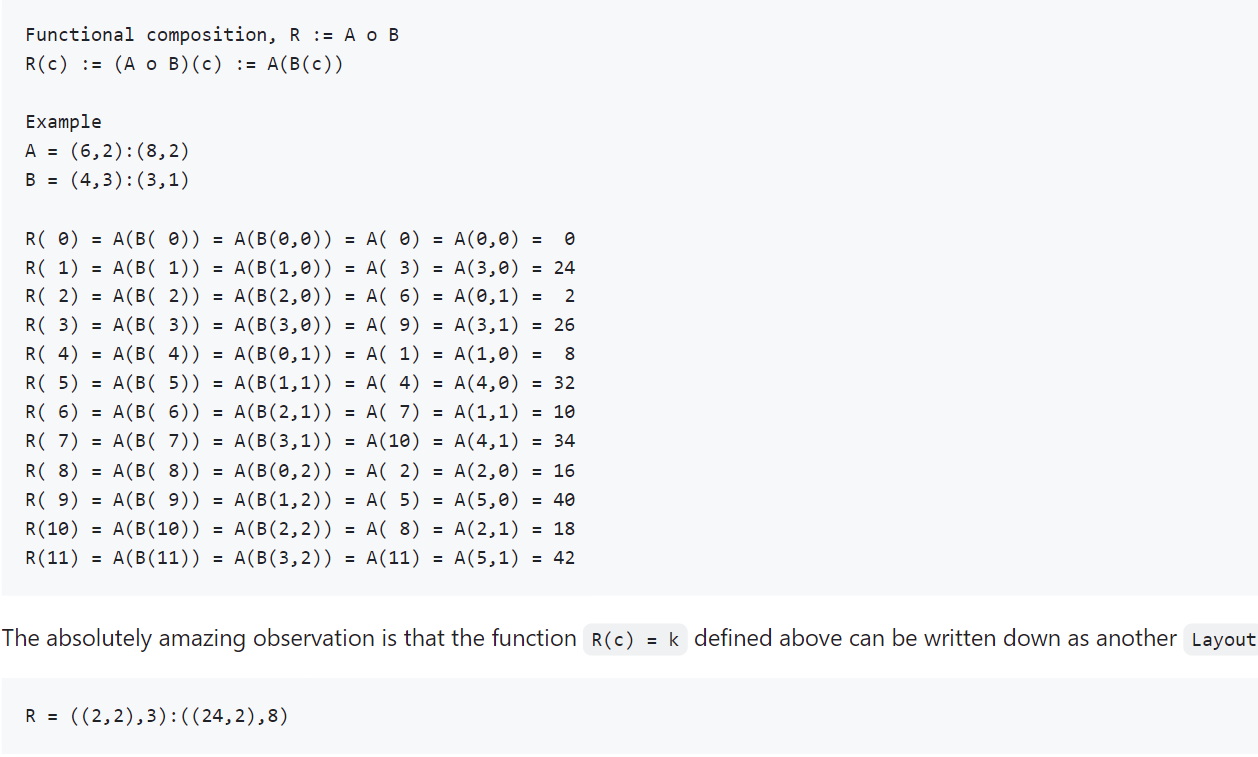
sA：((\_64,\_2),\_64,\_3):((\_1,\_4096),\_64,\_8192)---->((\_64,\_16),\_1,\_4,\_3):((\_1,\_64),\_0,\_1024,\_8192) tCsA----->GMMA::DescriptorIterator o (\_1,\_1,\_4,\_3):(\_0,\_128,\_1024) tCrA

上面是原先的数据变换方式。我们的GEMM0从哪里插入为好呢？最好是直接变成sA，但是怎么做呢？（没想明白。。不可能直接成为一样的shape&stride，因为初始指针排布都不一样）另一种可能是，保证最后落到每个线程的小张量正确即可，这样就是直接插到tCrA，但是这中间的变换都要我手动来写，就很复杂。

**备注：**

注意这里我们都统一为column major。以及sA看起来的宽度是192，GEMM0-result的宽度是256，不过这个宽度可以看做multi-stage的宽度。我们只需要把GEMM1的stage数变为4即可。

我看到对张量有composition方法。首先有指针A，然后进行layout L0，然后再进行layout L1，这两者可以等价于一个layout L3。举例如下：



似乎和我的需求很相关。。不过我没想明白怎么用。

**我的思路：**

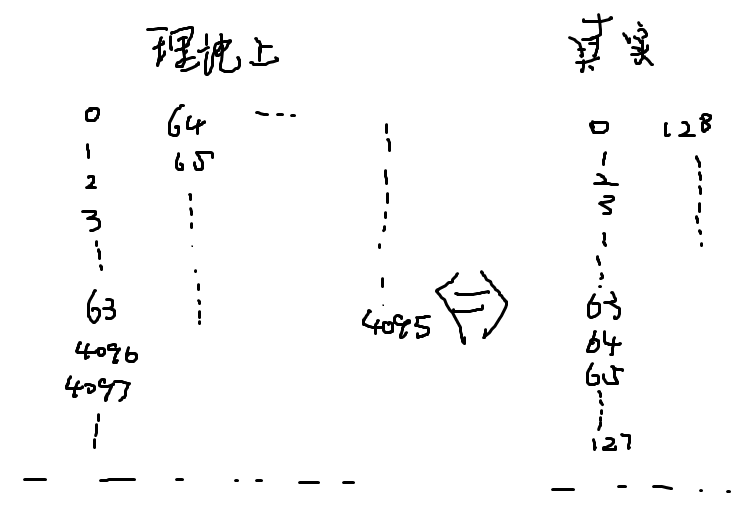
能不能，我从bSG\_gD的内存排布，实现sA的张量视图？然后composition到后续的tCsA？

好吧 我实在是搞不定这个方案。。。我想不明白该怎么操作。。。

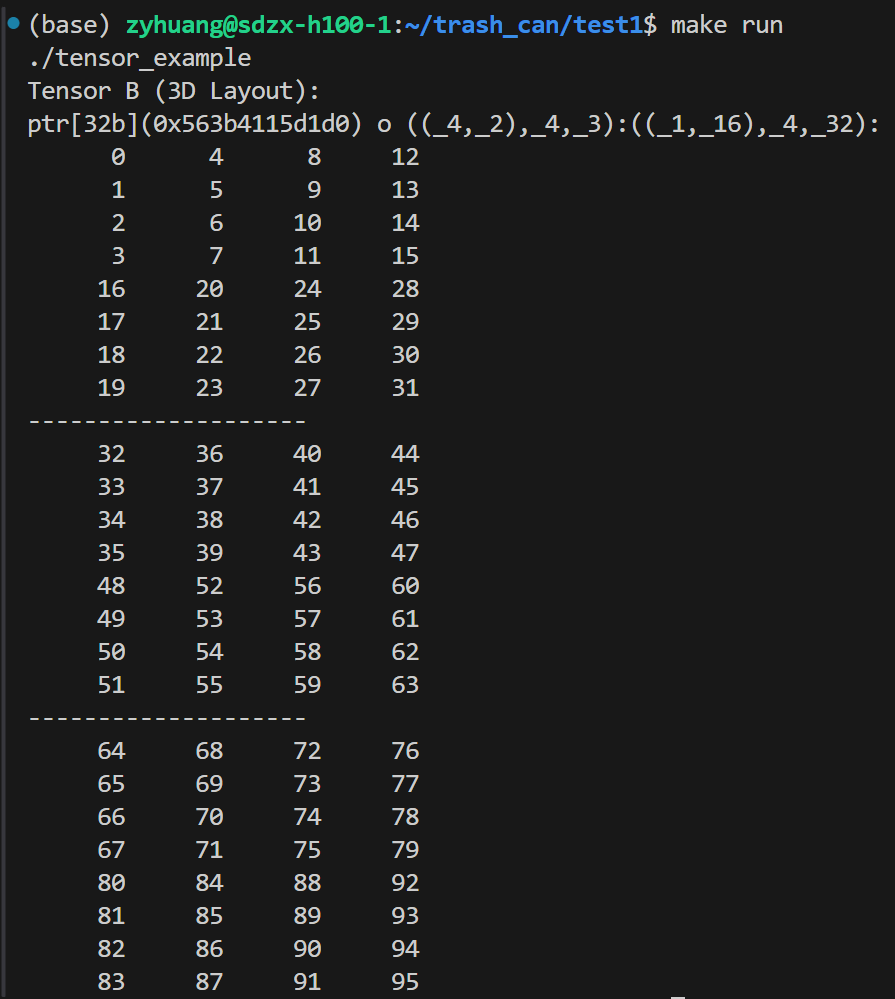
**遇到的问题：**

不过下面这俩问题并不卡住我的思路

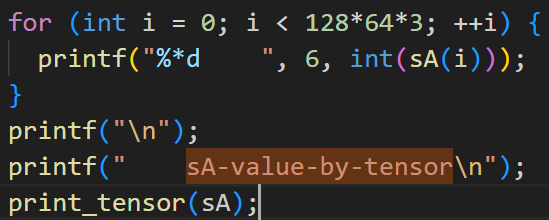
1. 按理说((\_64,\_2),\_64,\_3):((\_1,\_4096),\_64,\_8192) sA这样的会在行中间断一下。unit test也验证了。但是真实打印出来的结果却不是这样，是连续的。为什么呢？只给定shape和stride就能完全确定layout吗？难道还有其他的参量？

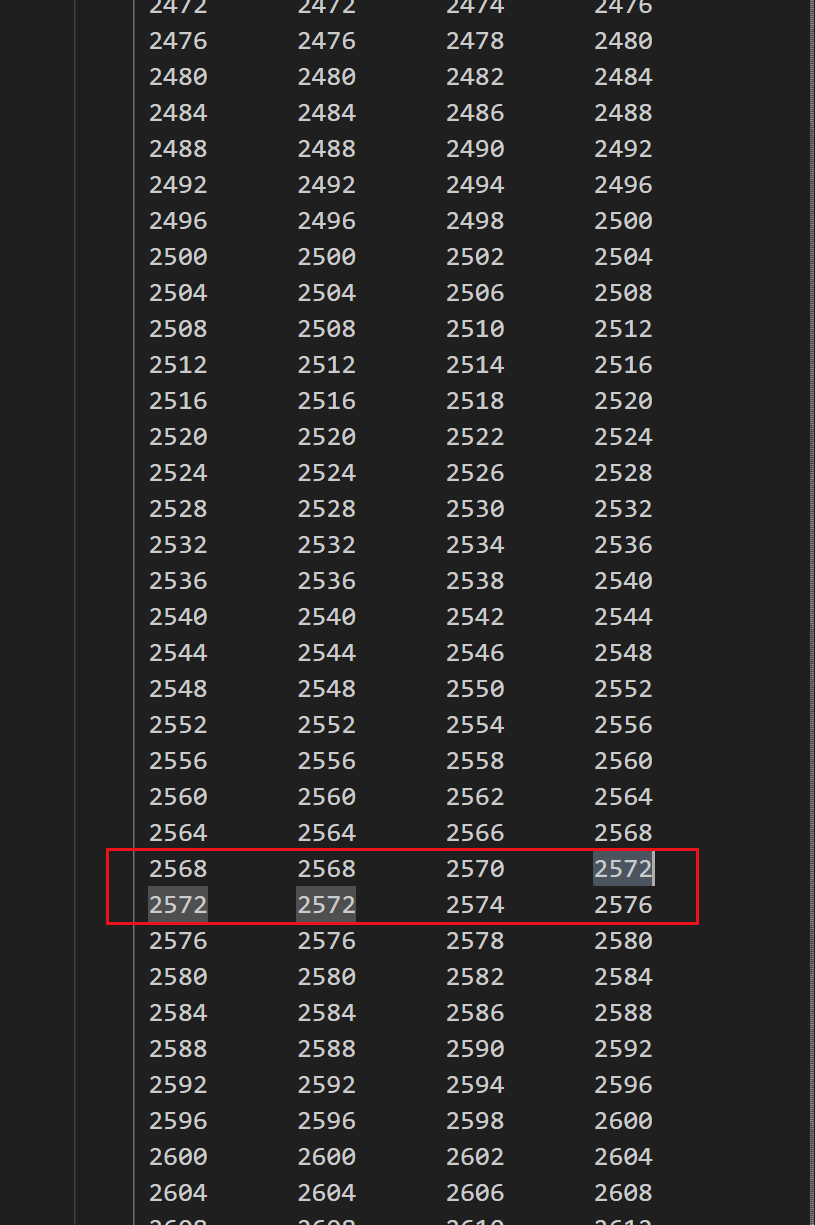


这个sA看起来更像(128, 64, 3)，而不是(64, 2), 64,3.。。。。这让我后续代码的书写很没底啊。。



1. 我从half转为float，然后打印，为什么很多值都被弄成一样的了？就是说，初始化的时候，我搞成32768，打印出来的时候，很多32765啦，32766啦都变成了32768，出现一大片32768，然后再跳跃到下一个值。但是哪怕是理论上，我这个位置的SMEM都不应该重复啊。。





在global打印（但是也是一样的打印方式rintf("%\*d ", 6, int(tensor(i, j)));）都是正常的。这里是SMEM的sA，按指针顺序打印出来。说不定是这里的打印方式有什么问题吗。。。

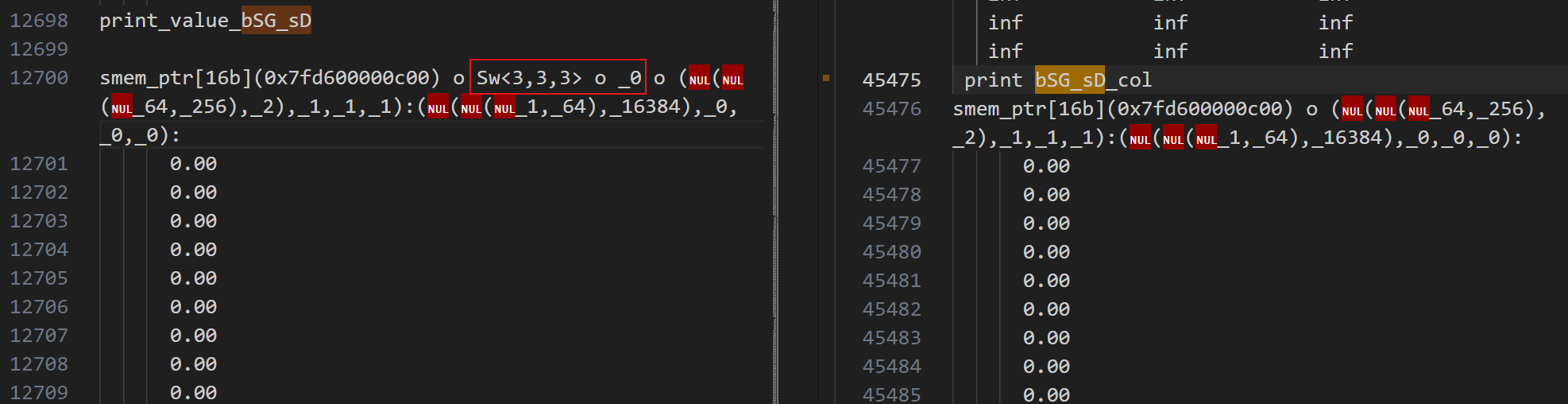
======================>为了避开这个问题，我现在的命令是：

make -j && ./49\_collective\_builder --m=128 --n=256 --k=64 --row\_end=128 --col\_end=64 > results.txt 2>&1

因为注意到重复往往都发生在较后的位置，说不定是多次SMEM读取，彼此相互干扰了。所以干脆只读一个stage的。现在就是，后面都是全0，就暂时看不到重复了。

=====>不行。还是有重复。经过检查发现half能够正确表示的整数范围只有-2048~+2048.。。。。。不过这不影响正确性。重复就重复吧。不管了。

以及，以后可以使用float2half，或者int2half之类的intrinsic函数，可能更准确一点？



这里红框里的是什么？为什么左右不一样？Chatgpt说可能是某种swizzle。嗯很有可能。从tensor core存出来是很乱的，然后自动做swizzle，以防bank conflict。那我如何从这个“swizzle”版本上再叠加后续的排序呢？也就是说，我不能利用指针来创建，我只能用composition，是吗？好吧。。。。？

详细来说，当前到底是什么layout？（因为swizzle，我其实搞不清楚），然后怎么composite到目标layout？（最后怎么验证是否一样？）

