C: {Dense,Dense,{0,1}}
B: {Dense,Dense,{0,1}}
A: {Dense,Compressed,{0,1}}
stmt: C(i,k) = A(i,k) * B(j,k)

数据格式和表达式

```
scalarPromote(stmt.reorder({i, k, j}).fuse(i, k, io)
.split(io, ko, ki, 8).reorder({ko, ki, j})
.pos(j, jpos, A(i, j)).reorder({ko, ki, jpos})
.split(jpos, jpos0, jpos1, 32)
.reorder({ko, ki, jpos1, jpos0})
.parallelize(ko, GPUBlock, IgnoreRaces)
.parallelize(ki, GPUWarp, Atomics)
.parallelize(jpos1, GPUThread, ParallelReduction)))
```

调度语言

```
...
int32_t ko = blockIdx.x;
int32_t jpos1 = threadIdx.x % 32;
int32_t ki = threadIdx.x / 32;
...
atomicAddWarp<float>(C_vals, kC, tjpos1C_val);
...
```

底层代码