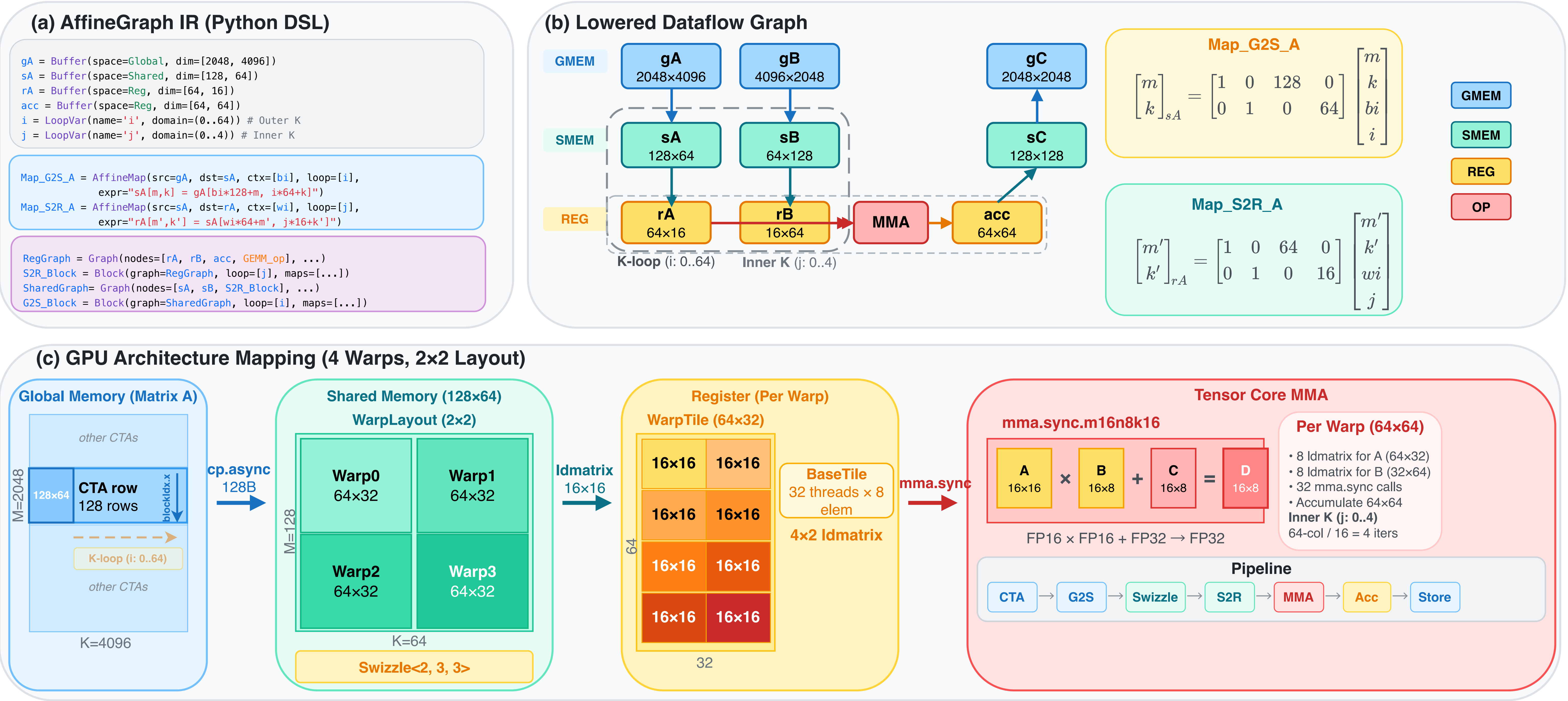

(a) AffineGraph IR (Python DSL)

```
gA = Buffer(space=Global, dim=[2048, 4096])
sA = Buffer(space=Shared, dim=[128, 64])
rA = Buffer(space=Reg, dim=[64, 16])
acc = Buffer(space=Reg, dim=[64, 64])
i = LoopVar(name='i', domain=(0..64)) # Outer K
j = LoopVar(name='j', domain=(0..4)) # Inner K
```

```
Map_G2S_A = AffineMap(src=gA, dst=sA, ctx=[bi], loop=[i],
              expr="sA[m,k] = gA[bi*128+m, i*64+k]")
Map_S2R_A = AffineMap(src=sA, dst=rA, ctx=[wi], loop=[j],
              expr="rA[m',k'] = sA[wi*64+m', j*16+k']")
```

```
RegGraph = Graph(nodes=[rA, rB, acc, GEMM_op], ...)
S2R_Block = Block(graph=RegGraph, loop=[j], maps=[...])
SharedGraph= Graph(nodes=[sA, sB, S2R_Block], ...)
G2S_Block = Block(graph=SharedGraph, loop=[i], maps=[...])
```

(b) Lowered Dataflow Graph

Map_G2S_A

$$\begin{bmatrix} m \\ k \end{bmatrix}_{sA} = \begin{bmatrix} 1 & 0 & 128 & 0 \\ 0 & 1 & 0 & 64 \end{bmatrix} \begin{bmatrix} m \\ k \\ bi \\ i \end{bmatrix}$$

Map_S2R_A

$$\begin{bmatrix} m' \\ k' \end{bmatrix}_{rA} = \begin{bmatrix} 1 & 0 & 64 & 0 \\ 0 & 1 & 0 & 16 \end{bmatrix} \begin{bmatrix} m' \\ k' \\ wi \\ j \end{bmatrix}$$

(c) GPU Architecture Mapping (4 Warps, 2×2 Layout)