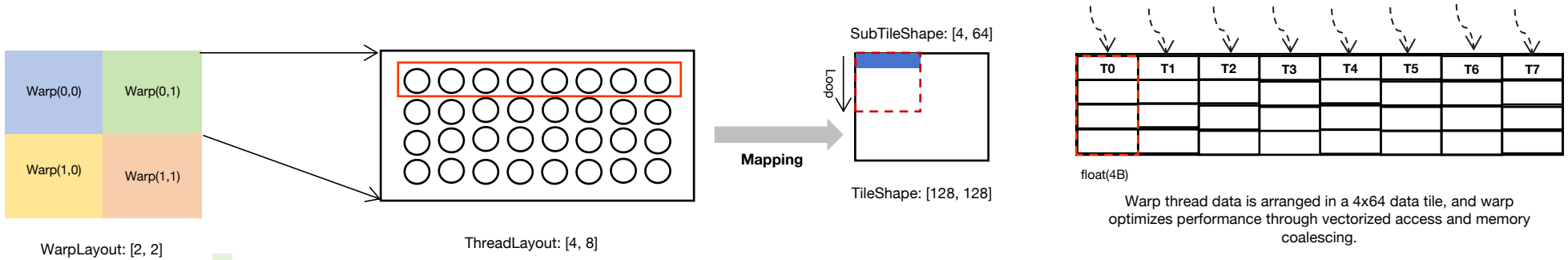
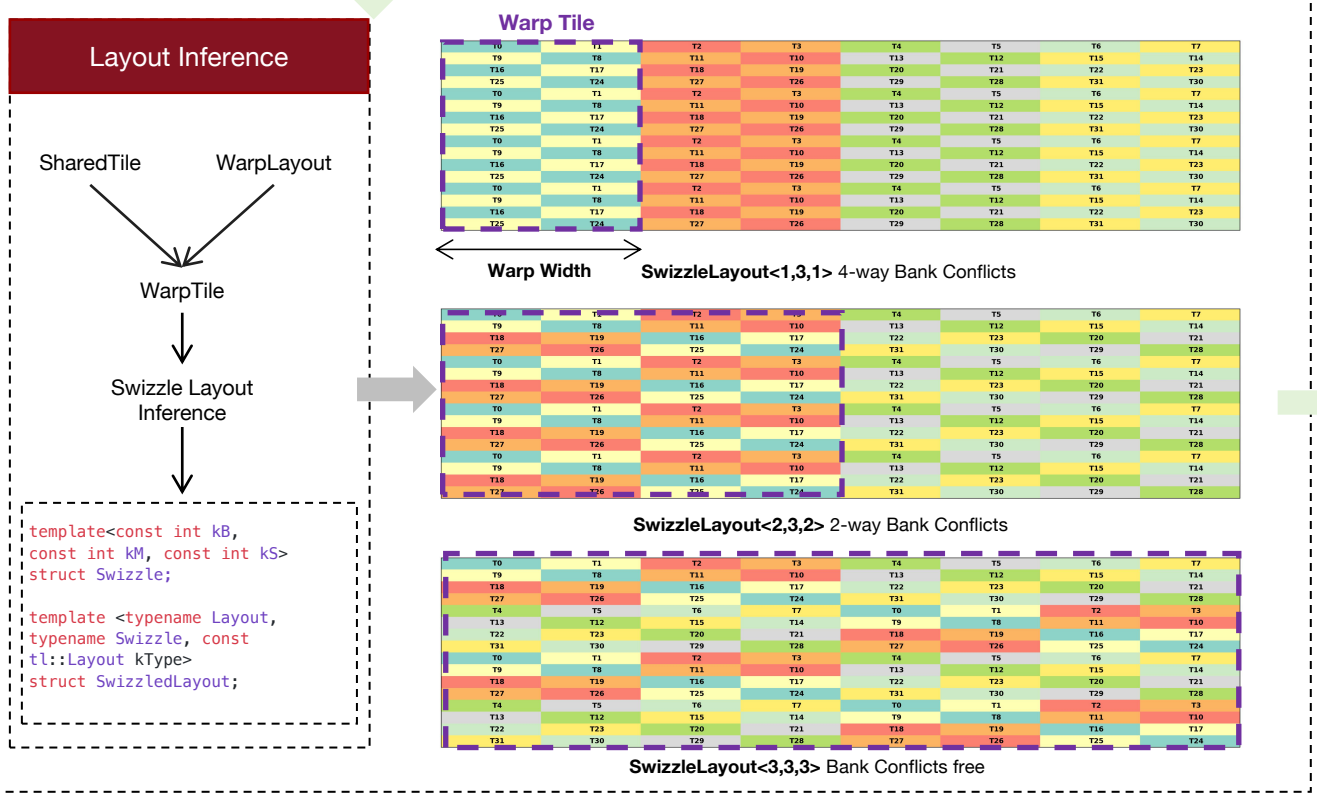
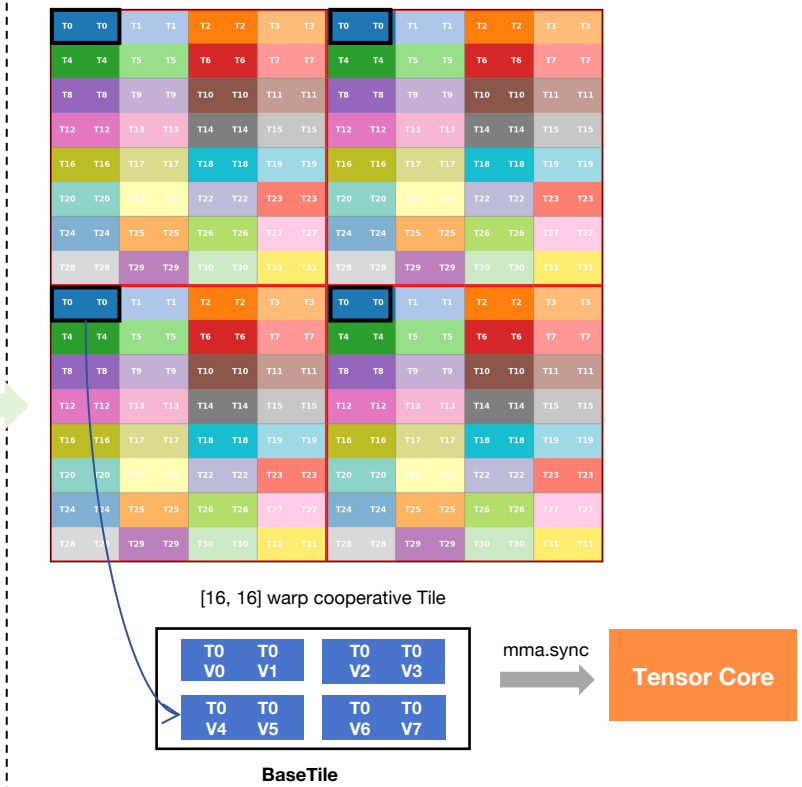



**(a) Warp-Level Layout (e.g., a 2 × 2 arrangement of warps with 4 × 8 threads each), establishing the (wi, wj) context.**




**(b) Shared Memory Swizzling is applied to buffers like sA and sB, transforming the logical layout to a physical one that avoids bank conflicts, thereby maximizing bandwidth.**




**(c) Thread-level layout within a single warp, AffineIR uses 16x16 as the base layout and extends it to multiple warps and repeats it in the MNK dimension.**