Building a MLIR compiler for real-time AI on existing 5G infrastructure

Isaac Nudelman, Ankush Tyagi

2025 LLVM Developers Meeting

Connecting people requires low latency

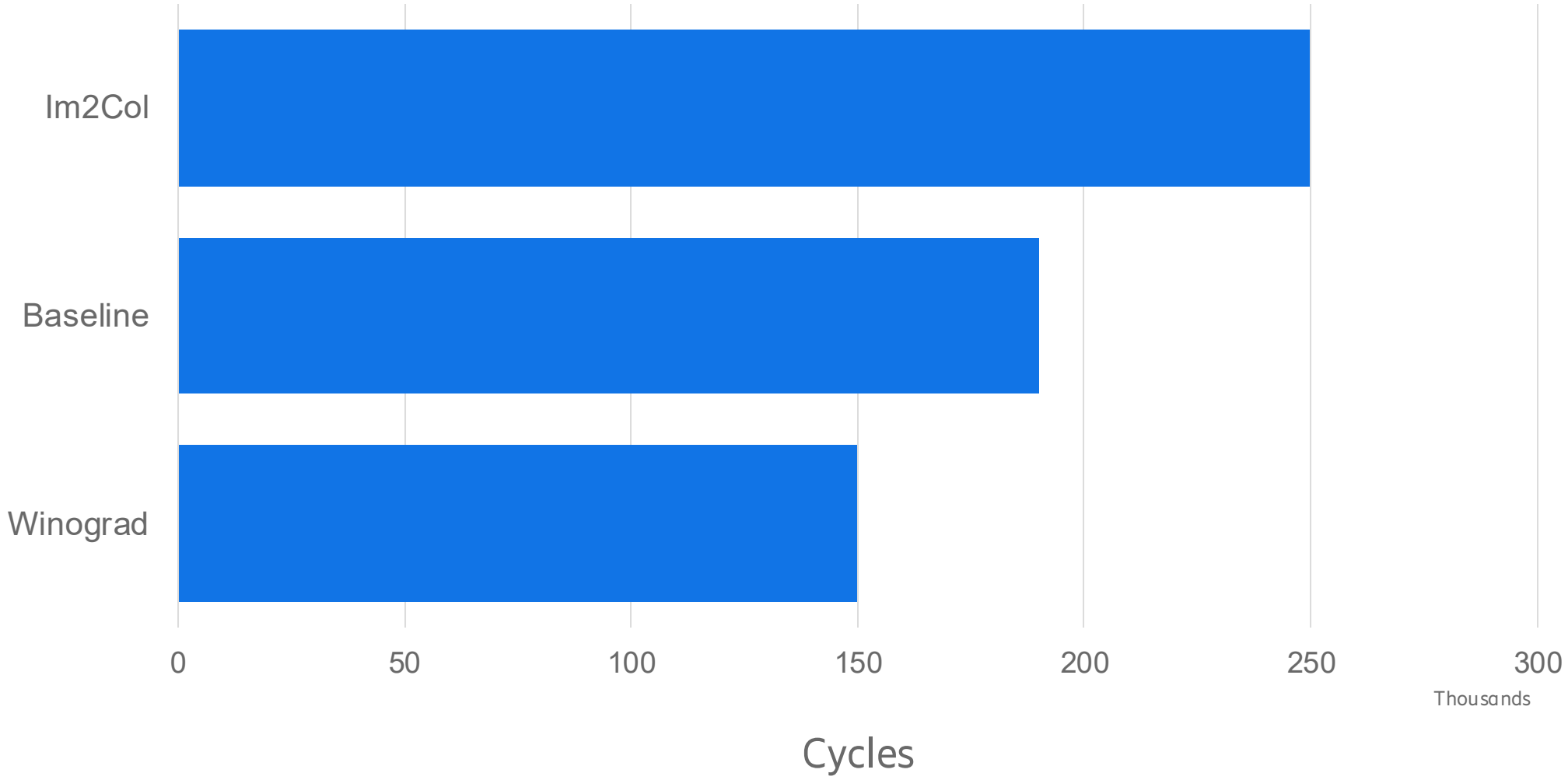
Processing ~ 25Gbps / antenna

Deadline as low as 100µs

Ericsson Many Core Architecture (EMCA)

Ericsson is upstream first

Picking optimization strategies for EMCA



Unplanned optimization is worse than nothing

ERICSSON

```
affine.for %arg2 = 0 to 25 {
  affine.for %arg3 = 0 to 14 {
    affine.for %arg4 = 0 to 3 {
      affine.for %arg5 = 0 to 3 {
        affine.for %arg6 = 0 to 3 {
          %10 = affine.apply #map1(%arg2, %arg5)
          %11 = affine.apply #map1(%arg3, %arg6)
          %12 = affine.load %alloca_1[%10, %11] : memref<27x16xf32>
          %13 = affine.load %5[%arg5, %arg6, %arg4] : memref<3x3x3xf32>
          %14 = affine.load %alloca[%arg2, %arg3, %arg4] : memref<25x14x3xf32>
          %15 = arith.mulf %12, %13 : f32
          %16 = arith.addf %14, %15 : f32
          affine.store %16, %alloca[%arg2, %arg3, %arg4] : memref<25x14x3xf32>
        }
      }
    }
  }
}
```

The biggest performance gains come from model architecture

DSP Cores

♥

Quantization

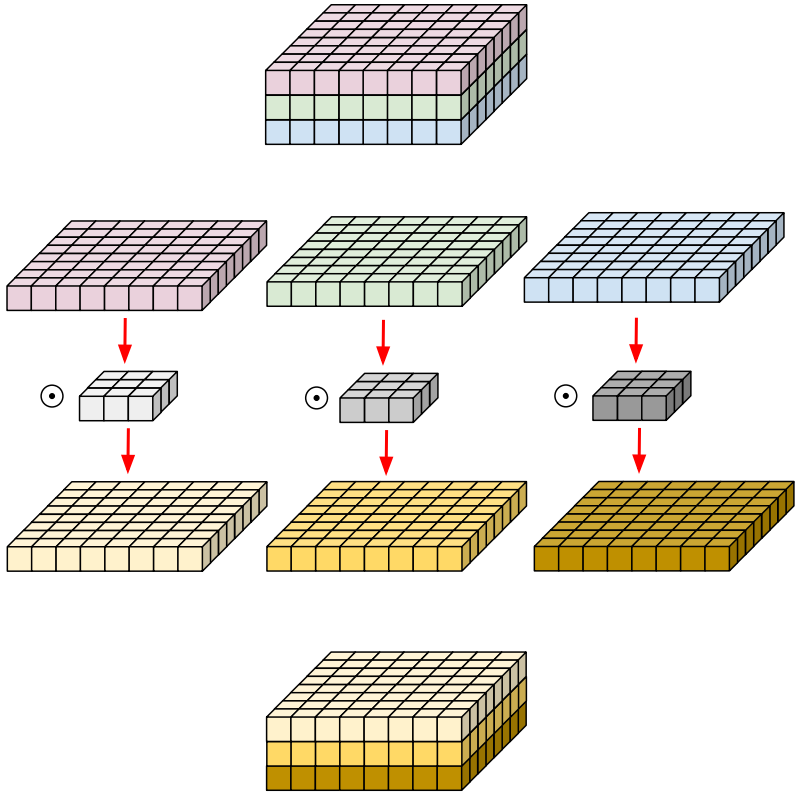


Image Credit: https://eli.thegreenplace.net/2018/depthwise-separable-convolutions-for-machine-learning/

Future Directions

ERICSSON

01100101	01101101	01100011	01100001

Unpacking required

00000000	00000000	00000000	01100001

00000000	00000000	00000000	01100101

Compute on in-memory representation

01100101	01101101	01100011	01100001

01100101	01011111	01011111	01011111


ERICSSON