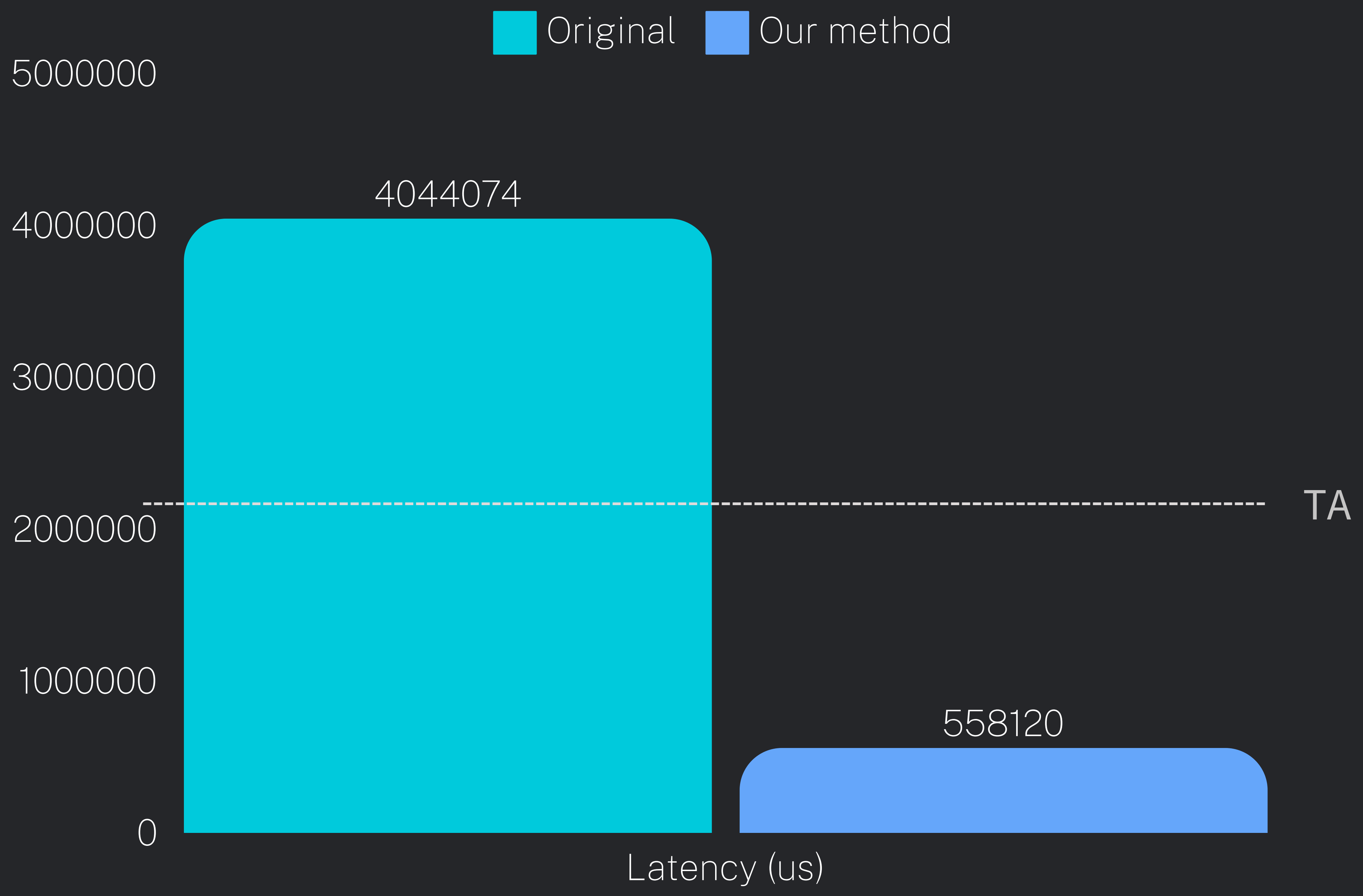
AAML FINAL PROJECT

GROUP 12

312832011	曾郁涵
312832003	林子泰
312833015	温峻揚
311511022	邱政岡

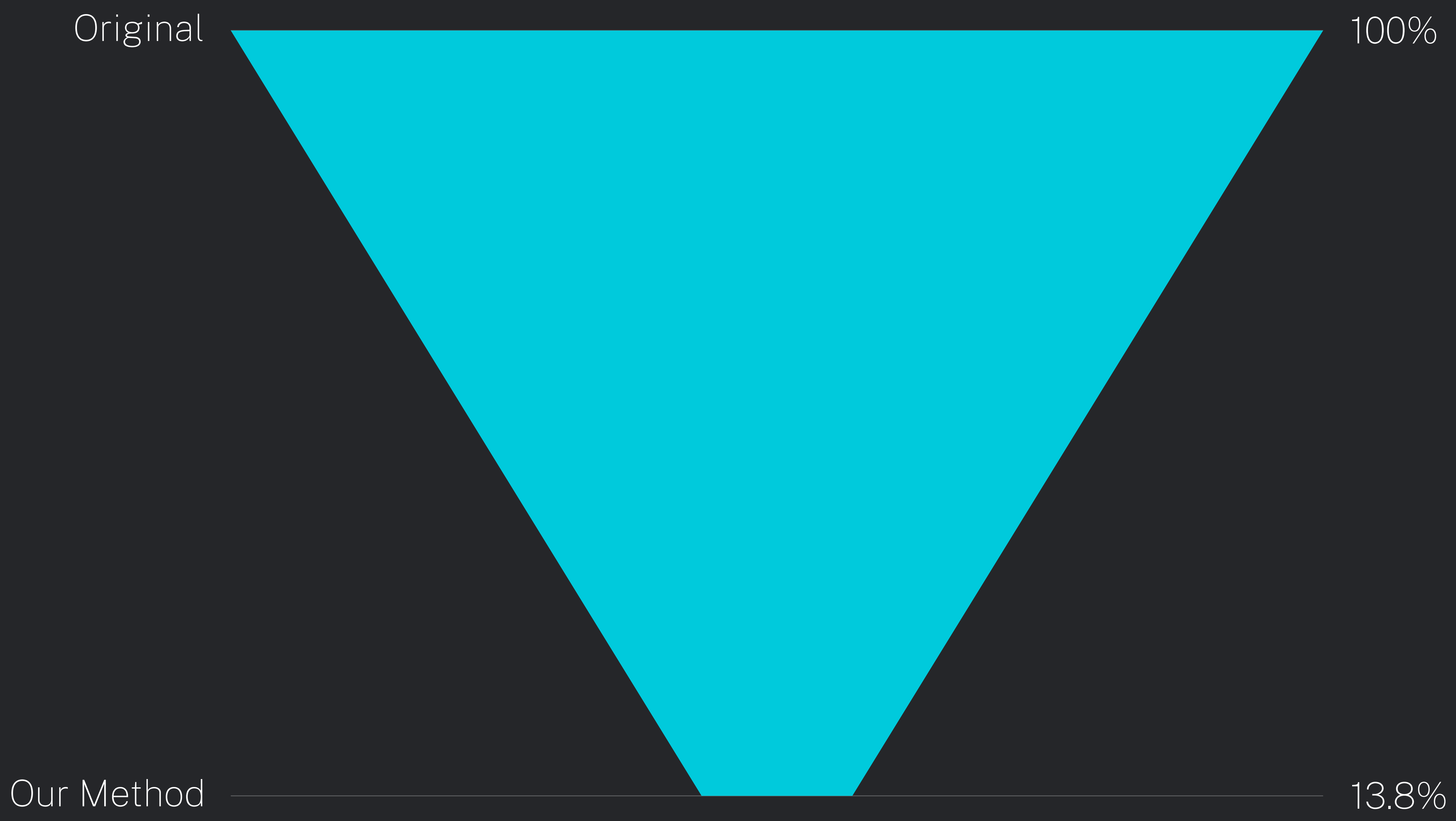


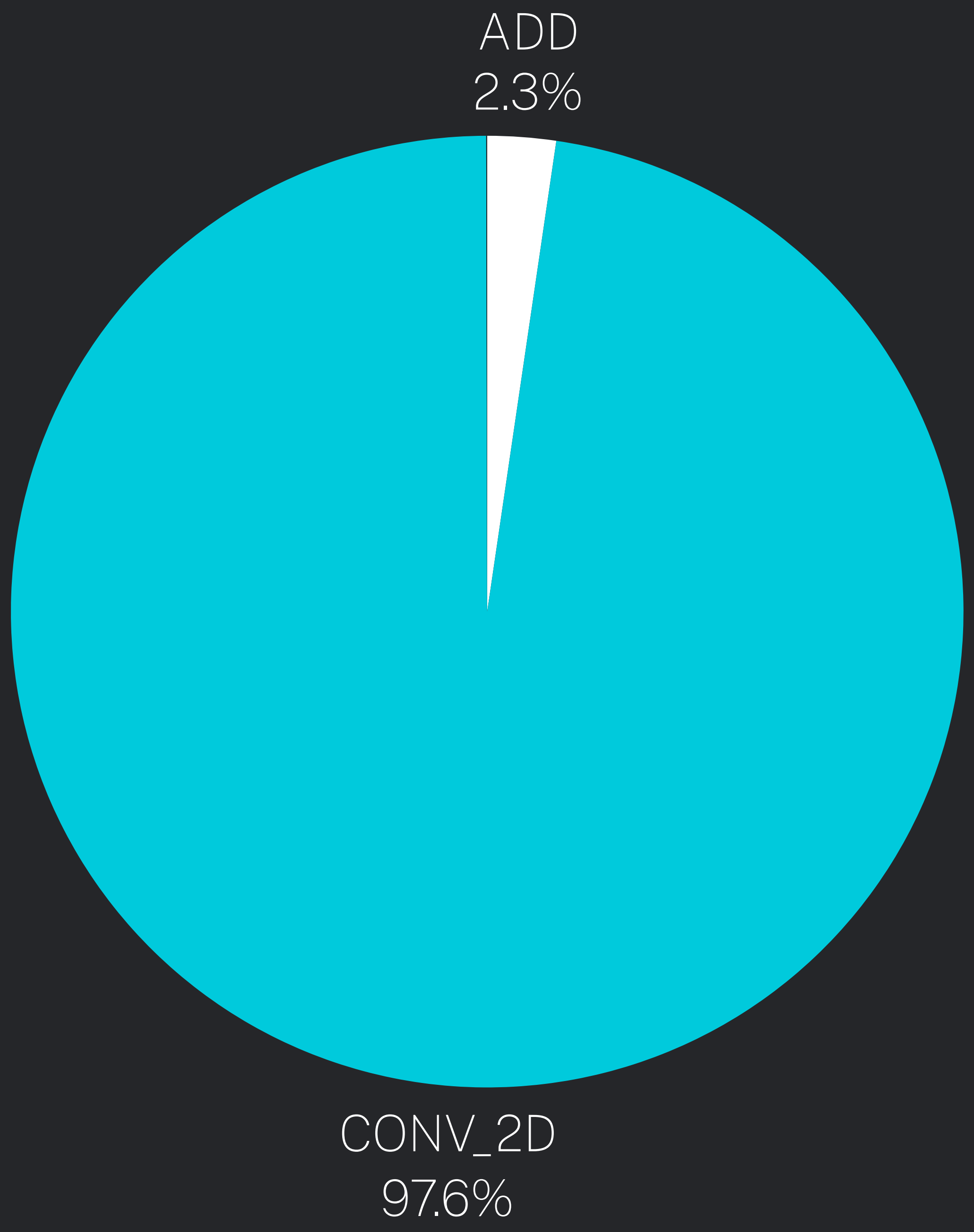
Performance

Acc : 87 %

Speedup in total: x7.25

Performance





Execution Time of Layers in Ticks

in MLPerf Tiny

Conv2D Layer Speedup

Method

1

SIMD

2

Software Optimization

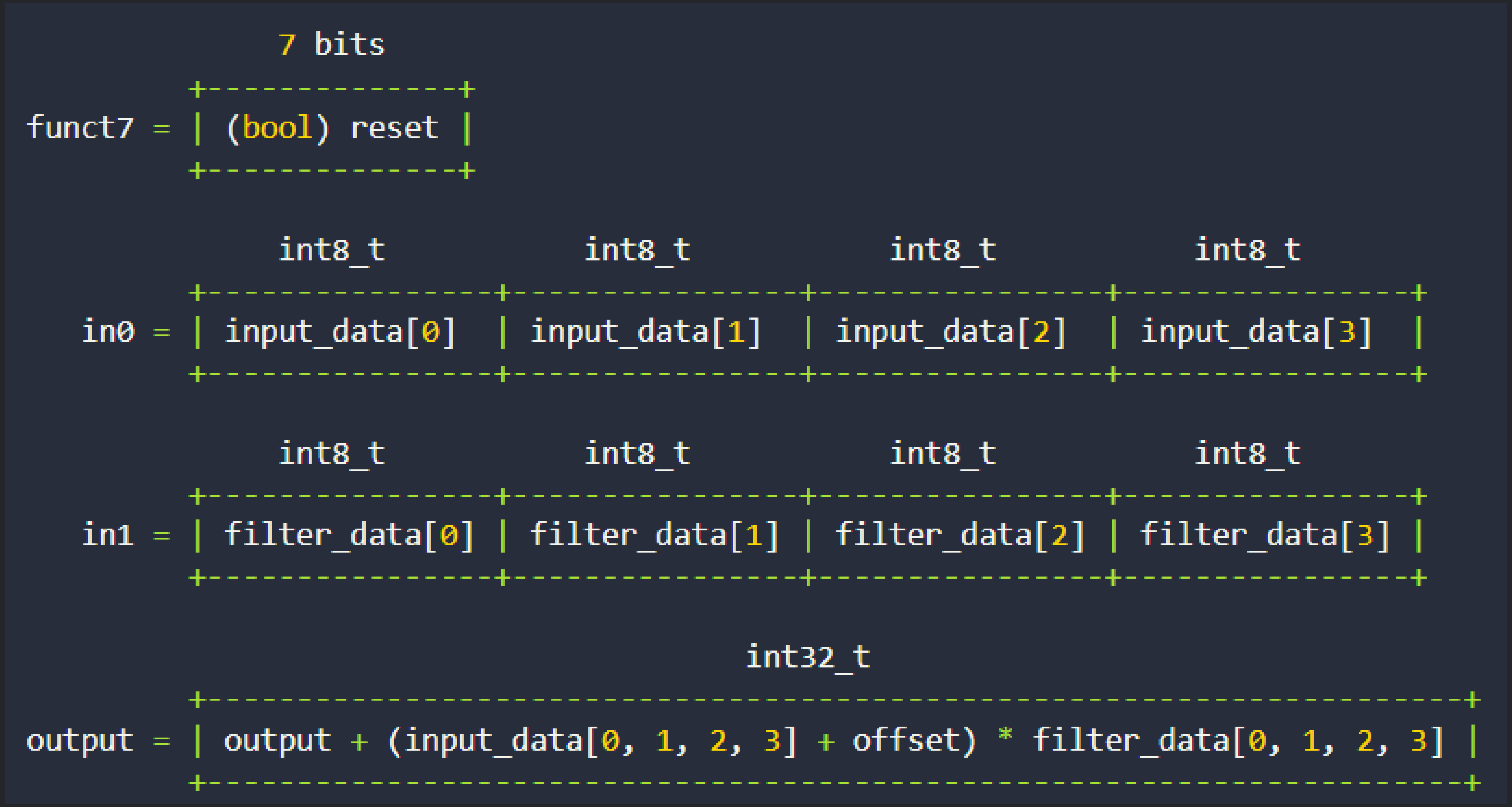
3

Postprocess on FPGA

4

Systolic Array with im2col

SIMD



```
              7 bits
         +--------------+
funct7 = | (bool) reset |
         +--------------+

              int8_t           int8_t           int8_t           int8_t
         +----------------+----------------+----------------+----------------+
   in0 = | input_data[0]  | input_data[1]  | input_data[2]  | input_data[3]  |
         +----------------+----------------+----------------+----------------+

              int8_t           int8_t           int8_t           int8_t
         +----------------+----------------+----------------+----------------+
   in1 = | filter_data[0] | filter_data[1] | filter_data[2] | filter_data[3] |
         +----------------+----------------+----------------+----------------+

                                        int32_t
         +----------------------------------------------------------------------+
output = | output + (input_data[0, 1, 2, 3] + offset) * filter_data[0, 1, 2, 3] |
         +----------------------------------------------------------------------+
```

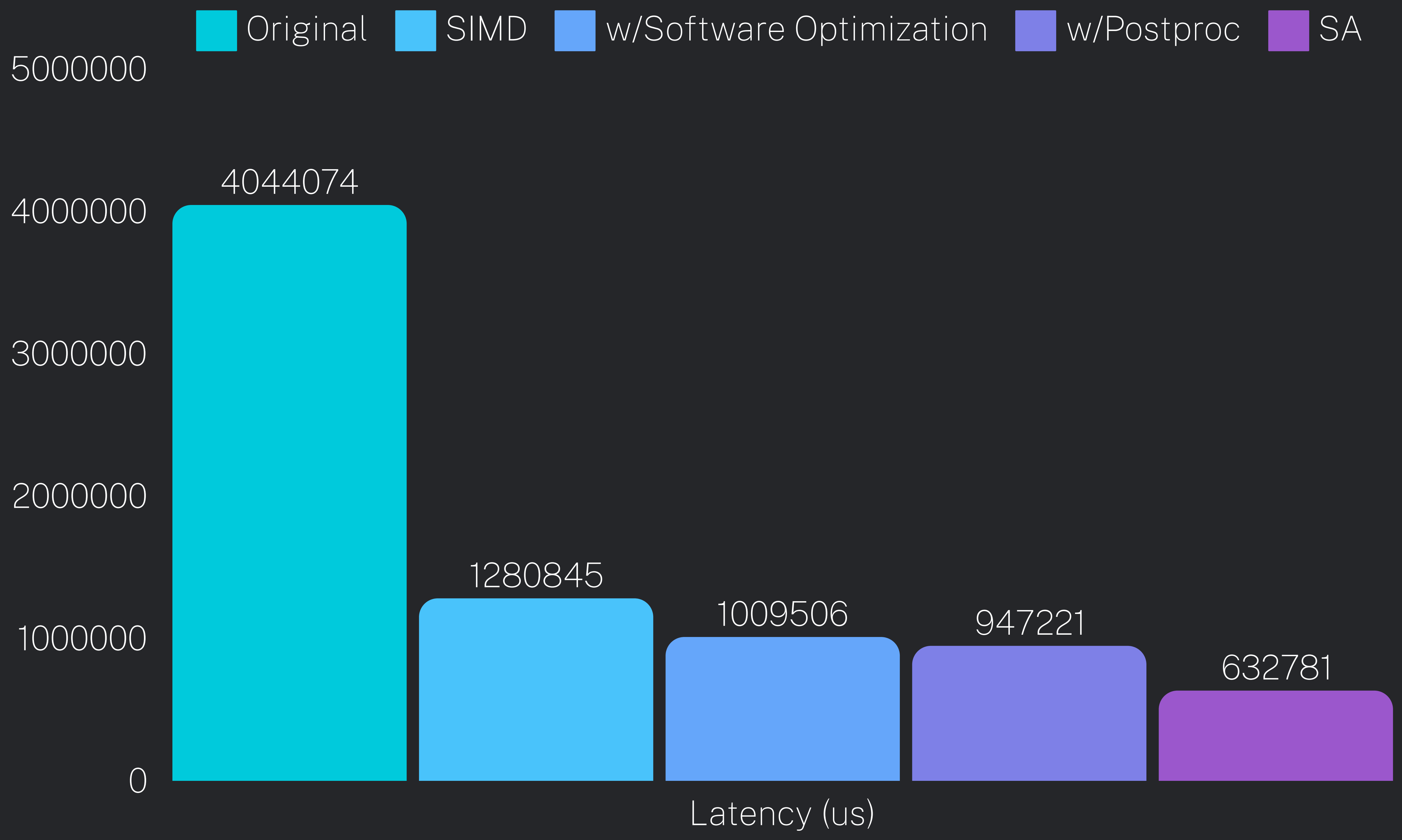
Software Optimization

1. Hardcode the constant parameters
2. Loop Unrolling
3. Minimize invocation of *Offset* func
4. Remove redundant computations

Postprocess on FPGA

MultiplyByQuantizedMultiplier

Systolic Array with im2col

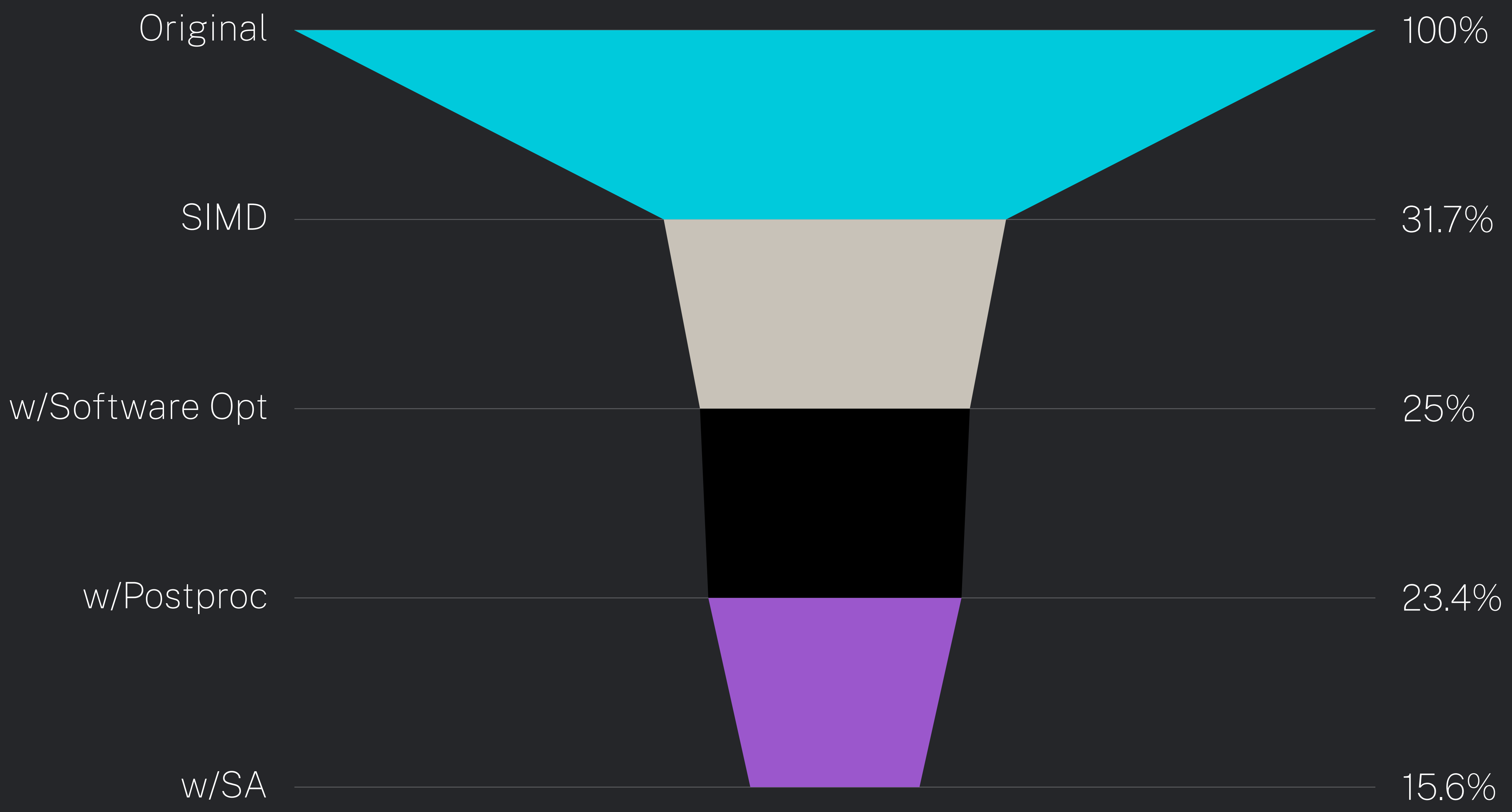


Performance

CONV_2D Layer

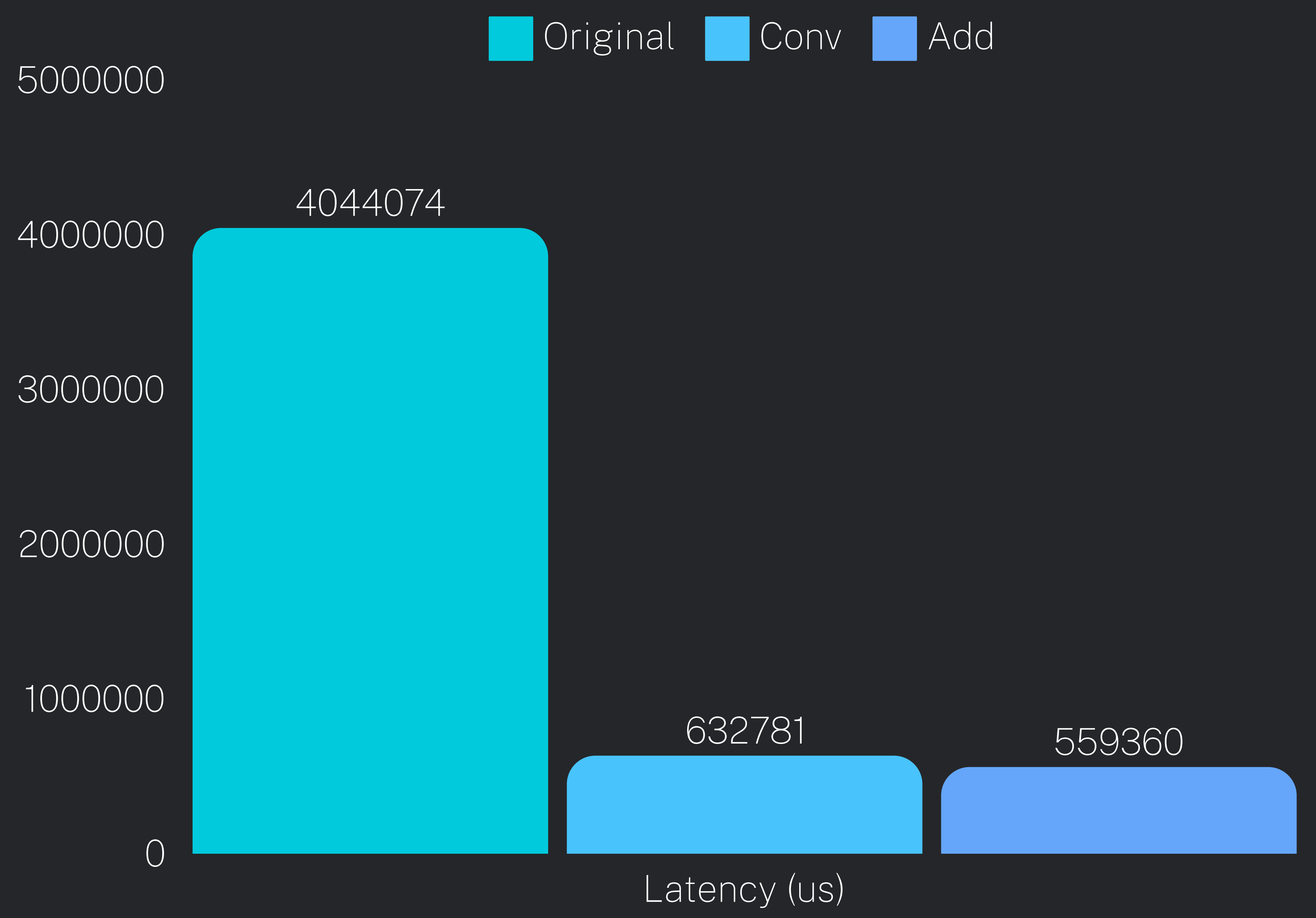
Speedup : x6.39

Performance



ADD Layer Speedup

SIMD & Postprocess on FPGA

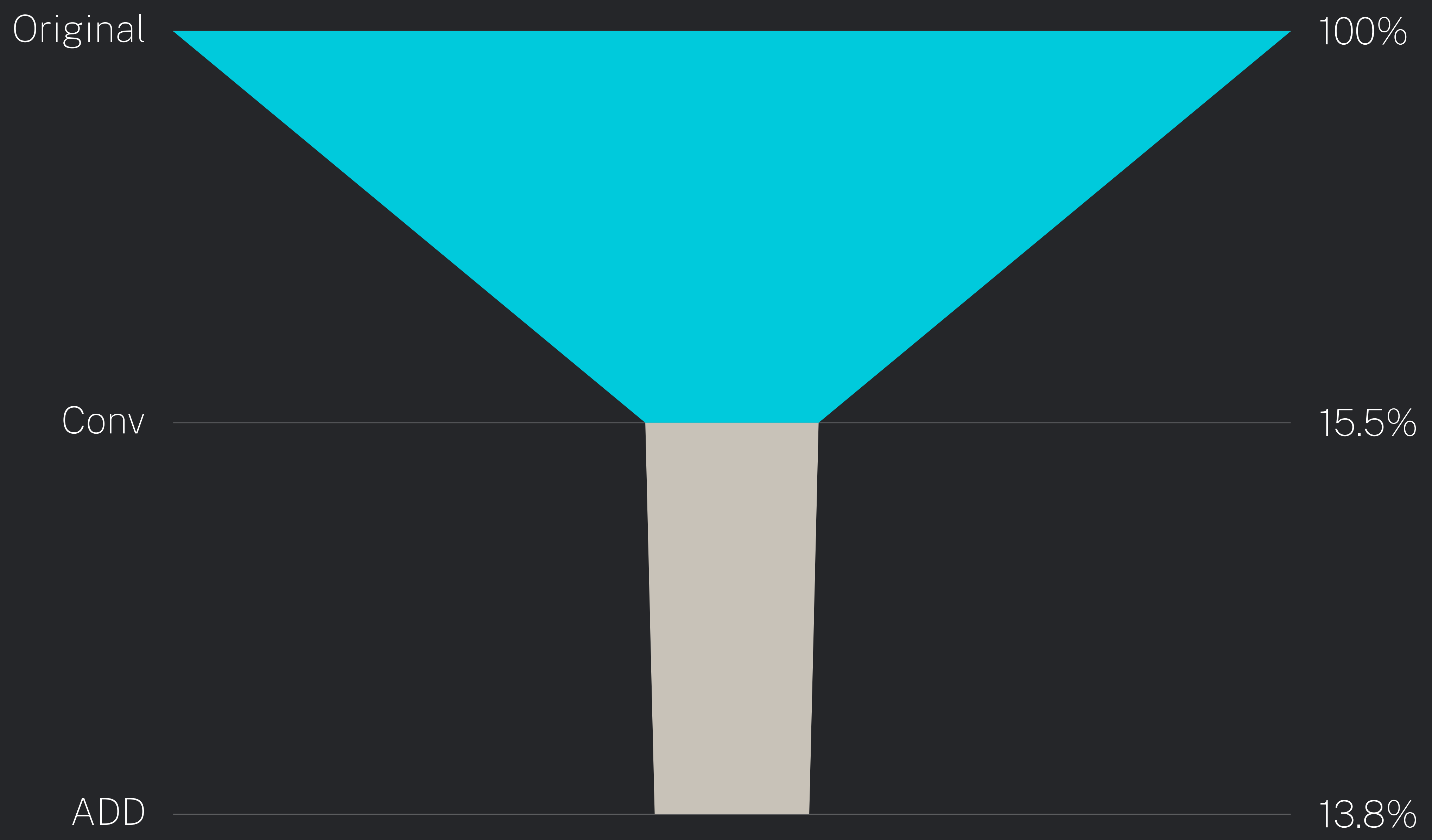


Performance

ADD Layer

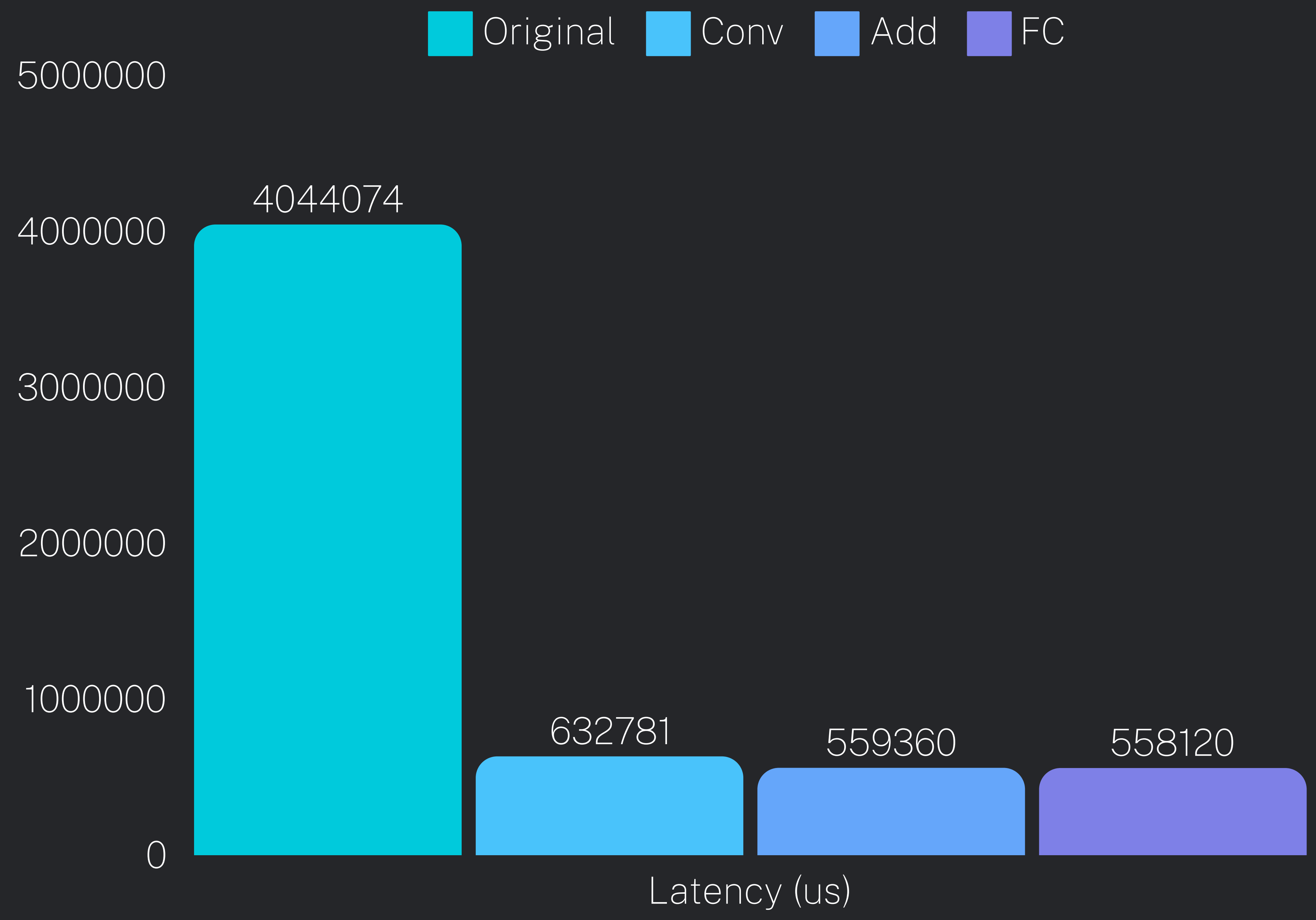
Speedup : x1.13

Performance



Fully Connected Layer Speedup

(Same as Conv2D)

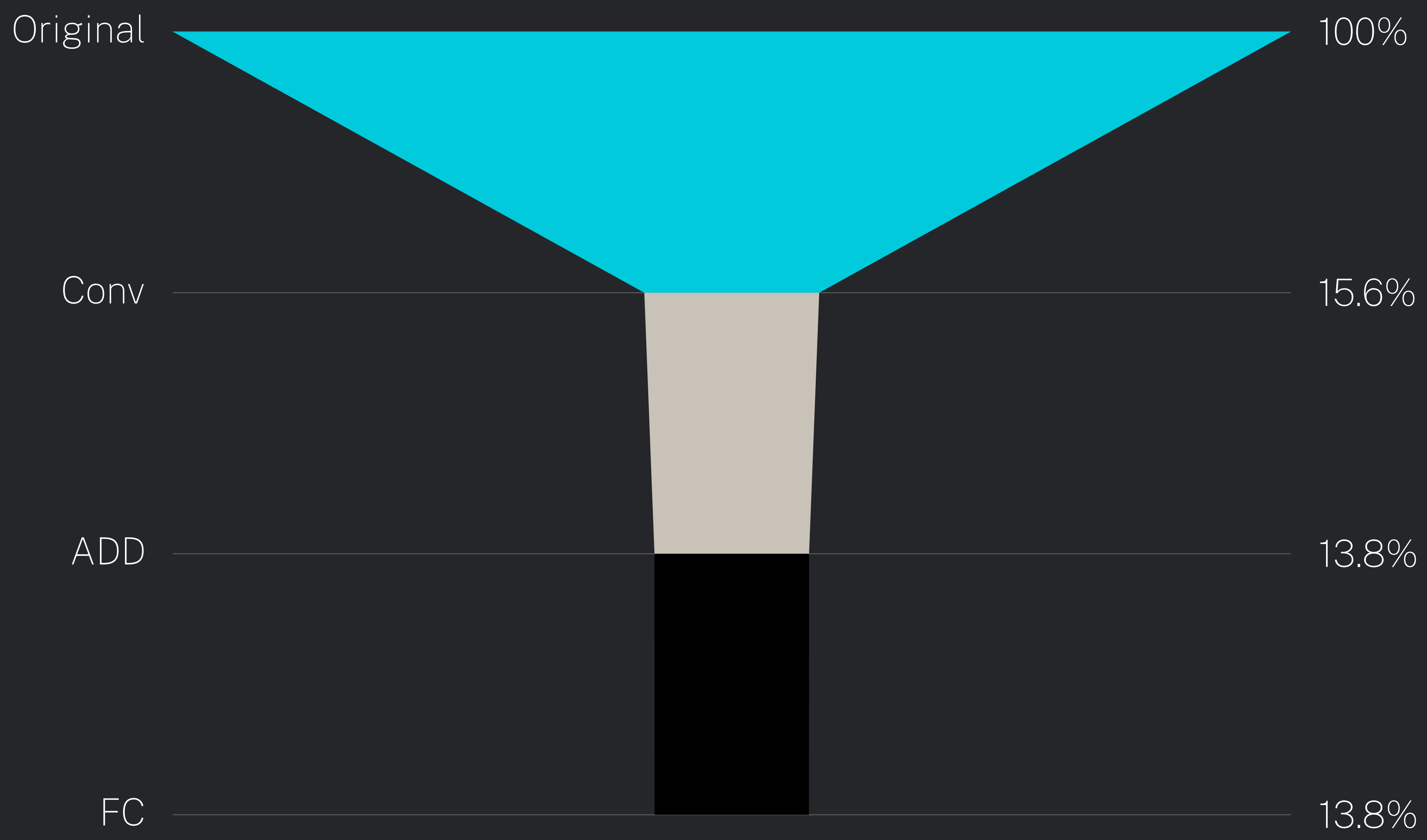


Performance

FULLY_CONNECTED Layer

Speedup : x1.002

Performance



Thank you!