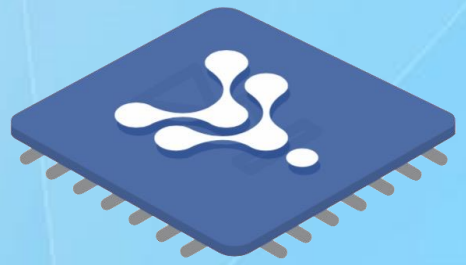
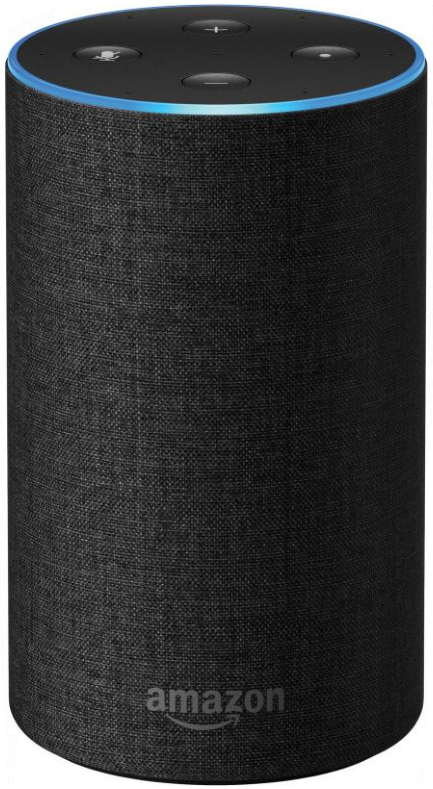


FWDNXT

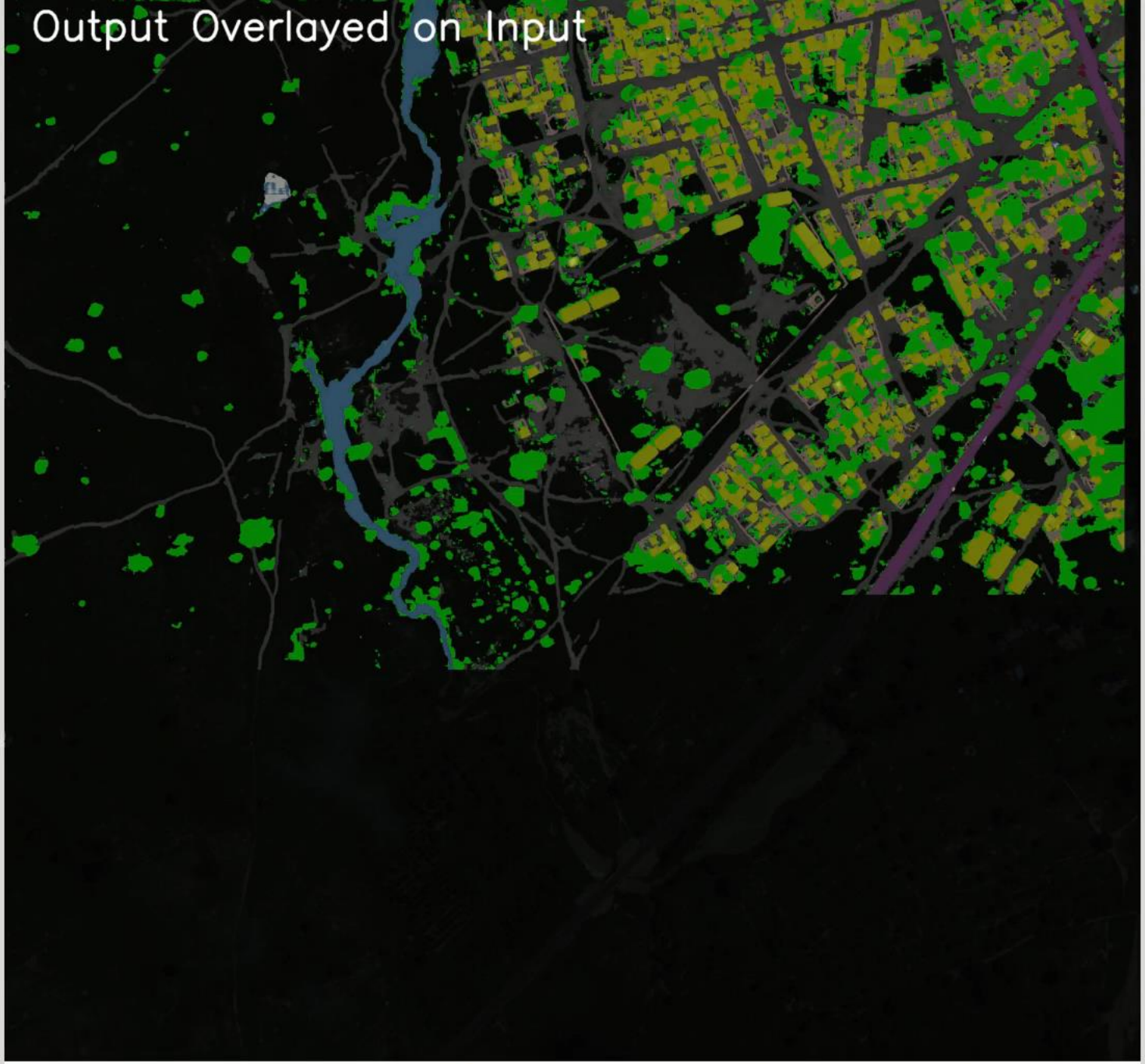
Snowflake deep neural network accelerator

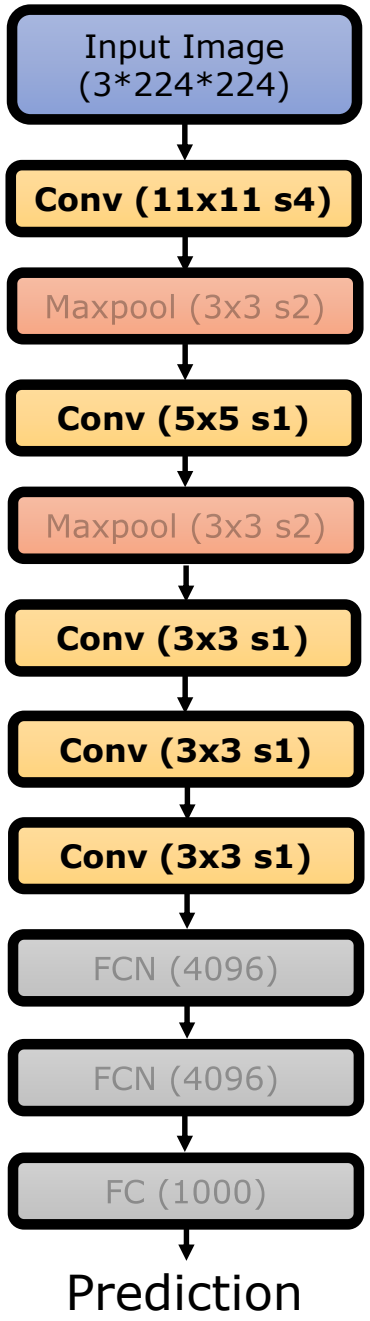
Deep Learning



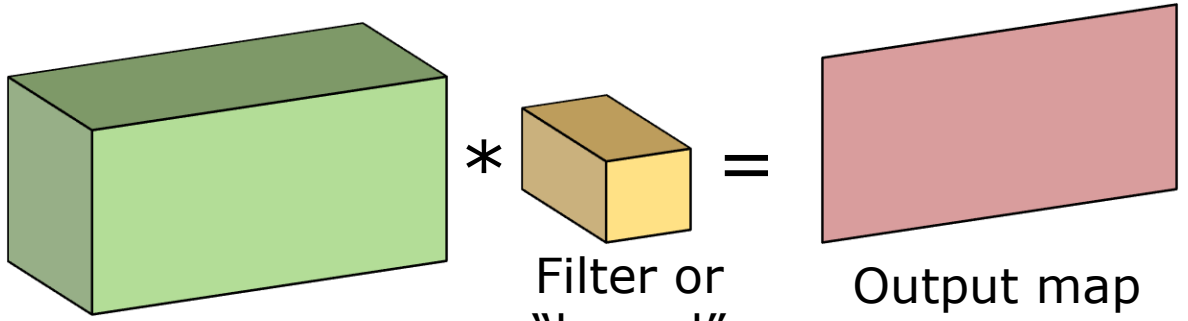

Person
Person
Euge

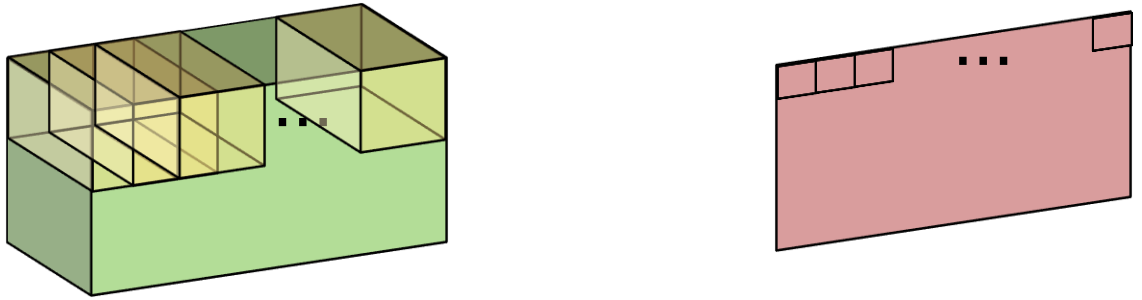
```
achang@FPGA3 ~/compiler $
achang@FPGA3 ~/compiler $ cd demos/
achang@FPGA3 ~/compiler/demos $ python3 test_build.py -l
HW PLAY 1 layer 1
Loading FPGA 0
Bringing up 2 hmc_phy_15g.v link(s) on HMC
Finished setting up the FPGAs
WARNING: overflows in node 0: 10.29 % (14402/139968)
datain[B]: 15094784.0, dataout[B]: 279936.0
Ops: 408969216.0000 [ops]
Time[ms] Expected: 4.2603 Measured: 4.6301
Bandwidth[GB/s] Expected: 3.3610 Measured: 3.0925
Eff Measured: 0.9201
SW run clus 0 card 0
output (H27, W27, P192)
Pass

done
>> Testing networks with test_hwsw
Network test failed for type 4
>> Testing networks with vertical padding optimization
>> Testing torchvision onnx networks
Traceback (most recent call last):
  File "test_build.py", line 81, in <module>
    img=Image.open('dog.jpg')
  File "/usr/local/lib/python3.5/dist-packages/PIL/Image.py", line 2543, in open
    fp = builtins.open(filename, "rb")
FileNotFoundError: [Errno 2] No such file or directory: 'dog.jpg'
achang@FPGA3 ~/compiler/demos $ cd ../
achang@FPGA3 ~/compiler $ ll
total 27908
-rwxrwxr-x 1 achang achang       90 Mar  6 10:18 allrun
drwxr-xr-x 3 achang achang     4096 Jul 19  2017 bin
-rw-rw-r-- 1 achang achang 19924038 Mar  6 12:25 bitfile.bit
drwxrwxr-x 2 achang achang     4096 Feb 26 23:07 bitfiles
drwxrwxr-x 5 achang achang     4096 Mar  6 10:19 bk
drwxrwxr-x 2 achang achang     4096 Mar  8 23:25 code_gen
-rw-rw-r-- 1 achang achang     2852 Jul 26  2017 compiler.cscope_file_list
-rw-rw-r-- 1 achang achang    28964 Mar  5 16:28 compiler.depend
-rw-rw-r-- 1 achang achang     9092 Mar  6 09:11 compiler.layout
drwxrwxr-x 3 achang achang     4096 Mar  8 23:16 demos
-rw-rw-r-- 1 achang achang    17379 Mar  7 16:36 instr_c0.txt
drwxrwxr-x 2 achang achang     4096 Mar  8 14:00 interface
-rwxrwxr-x 1 achang achang  1406800 Mar  8 22:15 libsnowflaked.so
-rwxrwxr-x 1 achang achang  1410928 Mar  8 23:32 libsnowflake.so
-rwxrwxr-x 1 achang achang   496320 Mar  5 16:07 loadfpga
-rw-rw-r-- 1 achang achang     2467 Mar  8 14:00 loadfpga.cpp
-rw-rw-r-- 1 achang achang     2057 Mar  6 10:18 Makefile
-rw-rw-r-- 1 achang achang     9214 Mar  1 13:51 Makefile.common
-rw-rw-r-- 1 achang achang      252 Mar  5 16:06 Makefile.load
-rw-rw-r-- 1 achang achang      787 Mar  8 23:24 Makefile.micron
drwxrwxr-x 2 achang achang     4096 Mar  6 10:18 micron
drwxrwxr-x 2 achang achang     4096 Mar  8 23:24 misc
drwxrwxr-x 8 achang achang     4096 Mar  8 17:53 models
drwxr-xr-x 3 achang achang     4096 Dec  3 20:55 obj
drwxrwxr-x 2 achang achang     4096 Mar  6 10:18 parse
drwxrwxr-x 2 achang achang     4096 Mar  8 23:24 partition
-rw-rw-r-- 1 achang achang      914 Oct  1 19:47 README.md
drwxrwxr-x 2 achang achang     4096 Mar  8 23:25 test
-rw-rw-r-- 1 achang achang    11385 Mar  8 09:33 test.csv
-rwxrwxr-x 1 achang achang  1095200 Mar  8 23:31 test_hwsw
-rwxrwxr-x 1 achang achang  4056384 Mar  8 22:15 test_hwswd
achang@FPGA3 ~/compiler $ ./test_hwsw -h 10 -l 0
HW PLAY 10 layer 0
Finished setting up the FPGAs
MENIO:
What art thou this the death,
```


Output Overlayed on Input
Building
Manmade
Road
Track
Trees
Crops/Unlabeled
Waterway
Pond
Truck
Car

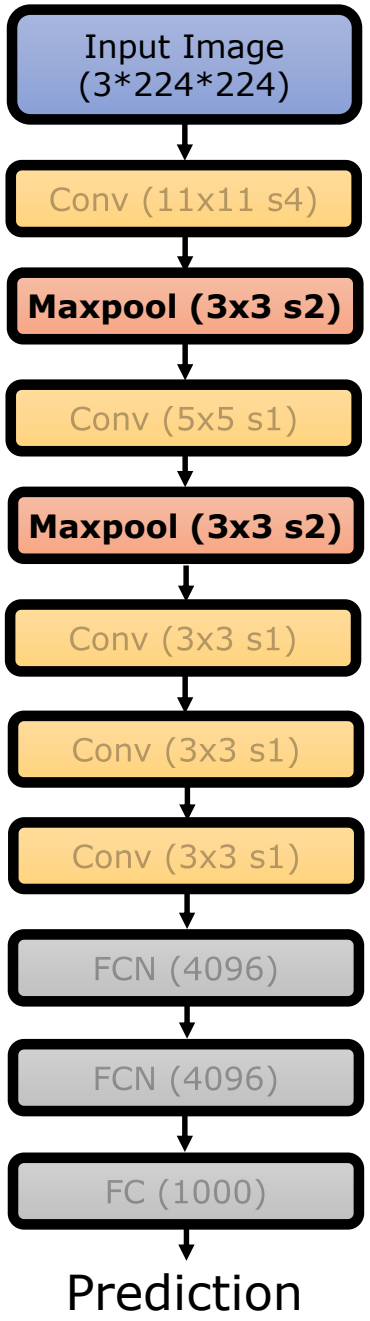


Convolutional Neural Networks

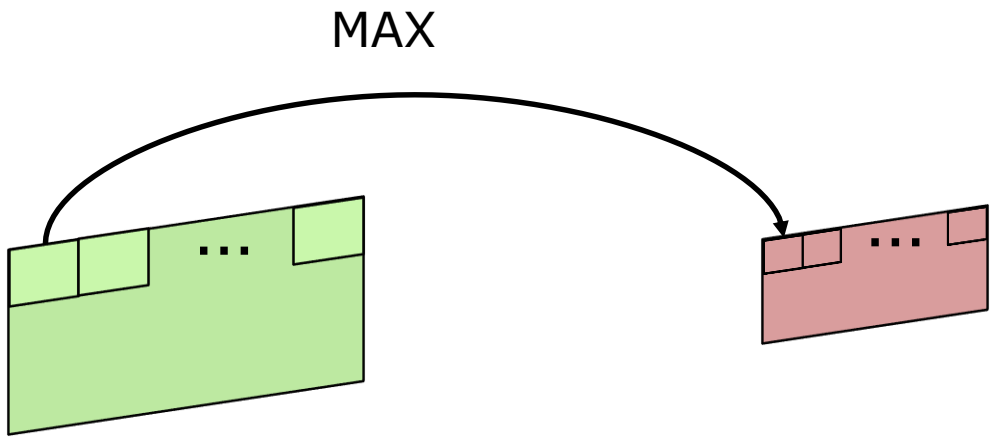




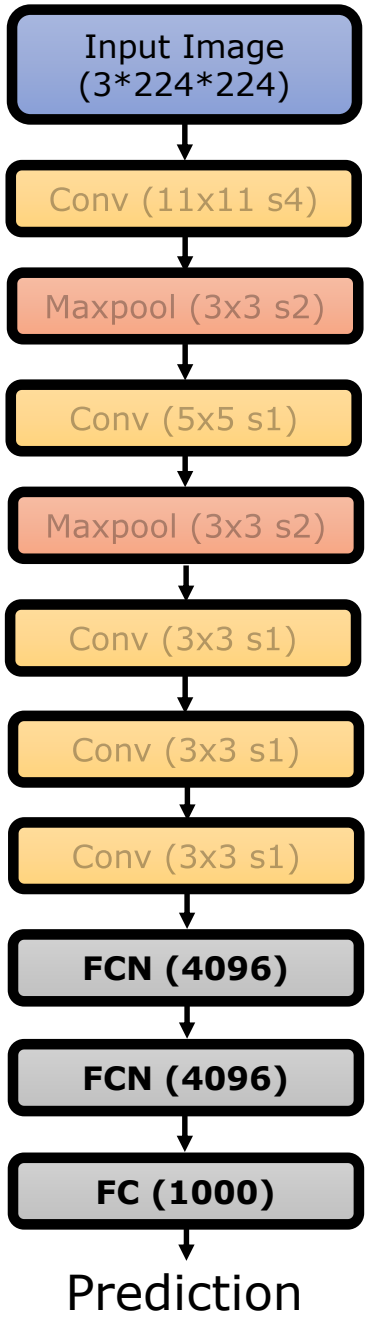
- Compute intensive
- Embarrassingly parallel
- Comprise > 95% of the workload
- Comprised of mult-acc (MAC) ops



Convolutional Neural Networks



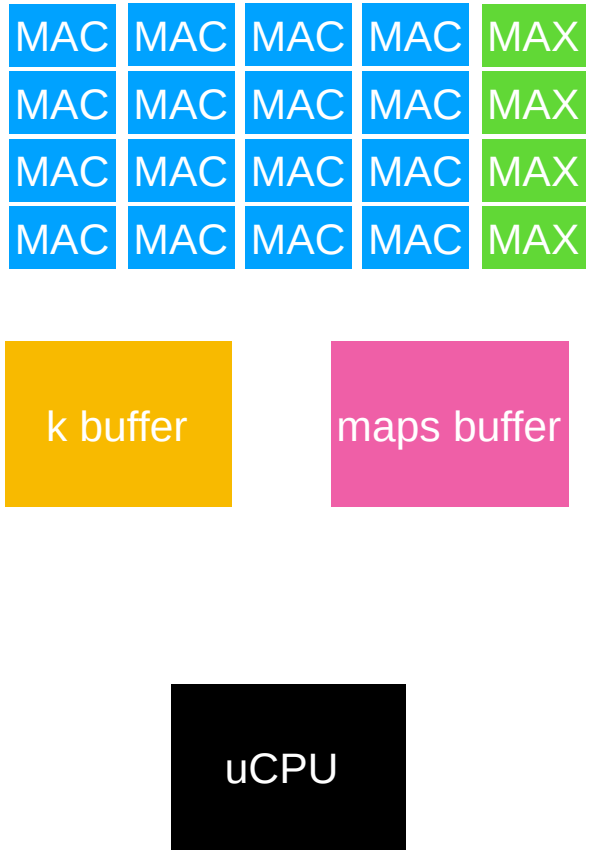
- Make up ~1% of the workload
- Lesser parallelism to exploit
- Comprised of comparisons



Convolutional Neural Networks

- Tens of MB of weights
- No weights reuse
- Bandwidth intensive
- Comprised of MACs

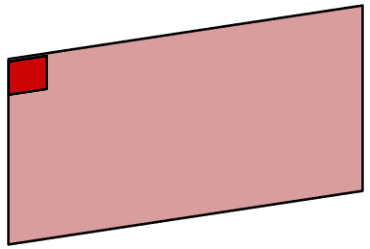
Accelerator Hardware



- **Functional units**
  - Multiply-accumulate (MAC)
  - Comparators (maxpool)
- **On-chip memory**
  - Buffers for maps and weights
- **Configuration logic**
  - Instruct on-chip memory to stream to MACs
  - Instruct MACs to write-back results

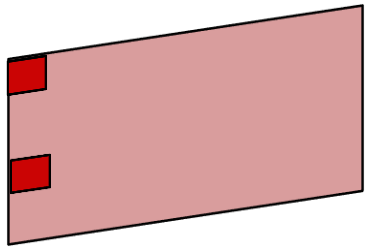
Types of Parallelism

Intra-map, intra-activation



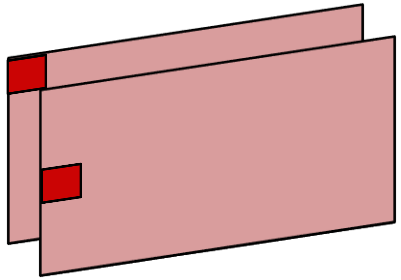
MACs share both input operands

Intra-map, inter-activation



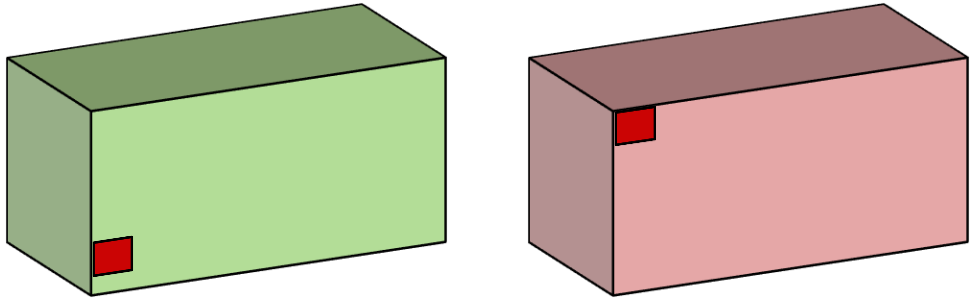
MACs share weights

Inter-map



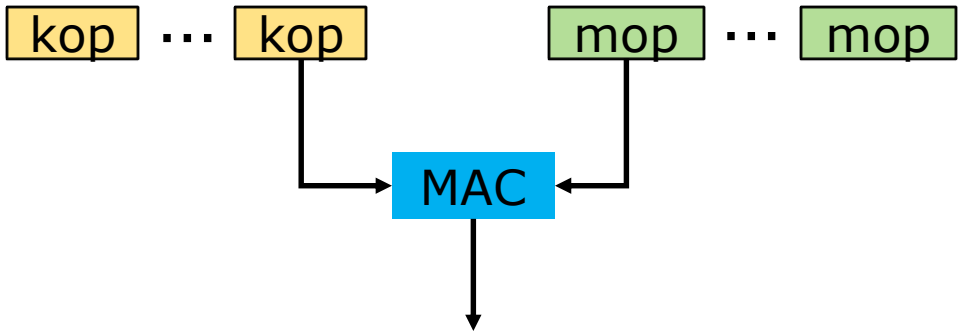
No data sharing

Inter-layer



No data sharing

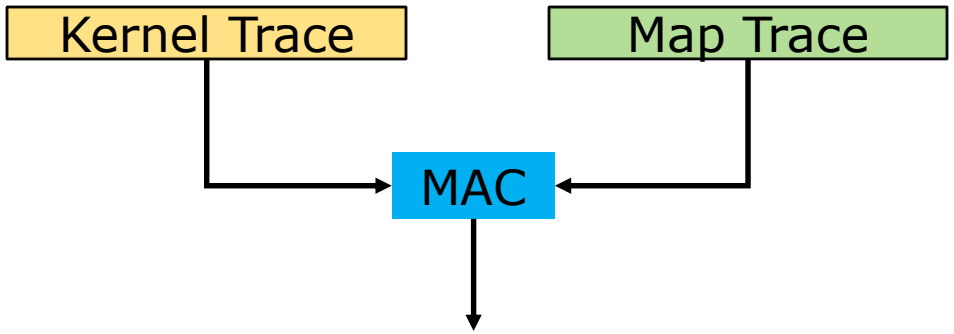
Traces



Require N instrs
1 instr per cycle

MAC R0, R1, R2

kop mop



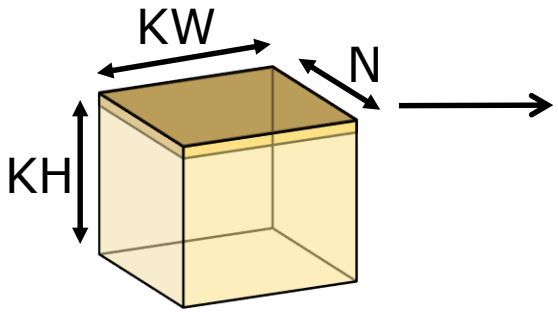
Require 1 instr
1 instr per trace
Require start addr, length

MAC R0, R1, R2, imm.

length

Kernel trace start addr. map trace start addr.

Data Organization

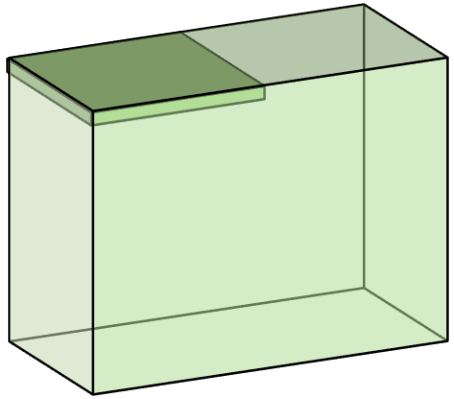


Kernel Trace

KW * N

Vec 1 | Vec 2 | … | Vec M

Subdivided into M vectors



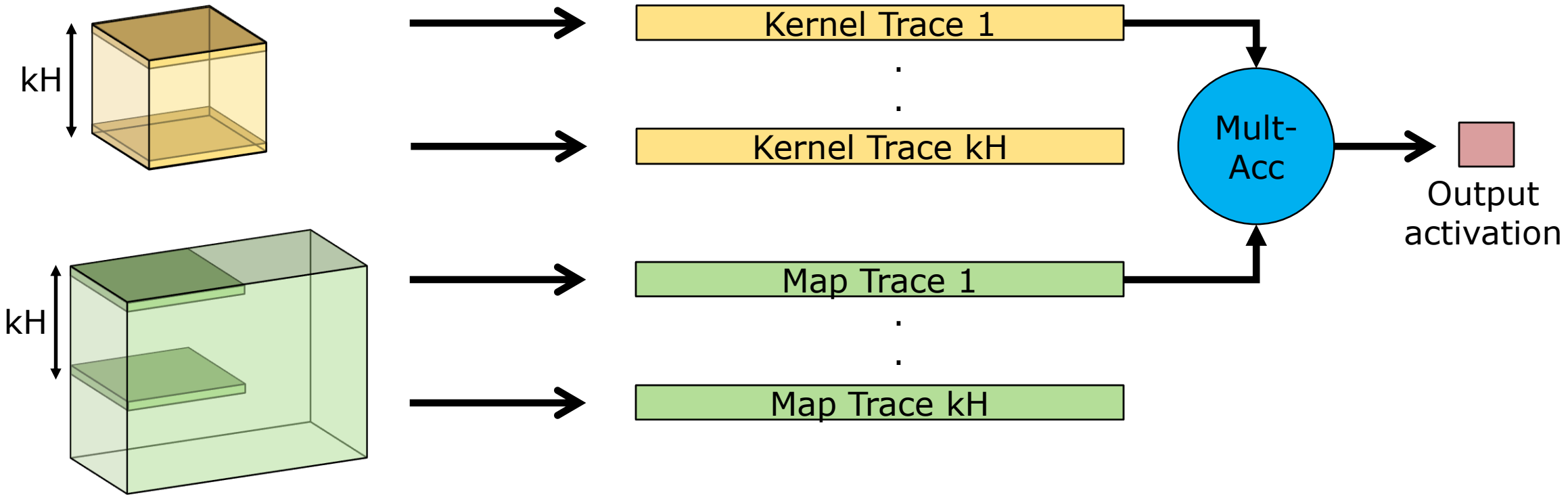
Map Trace

KW * N

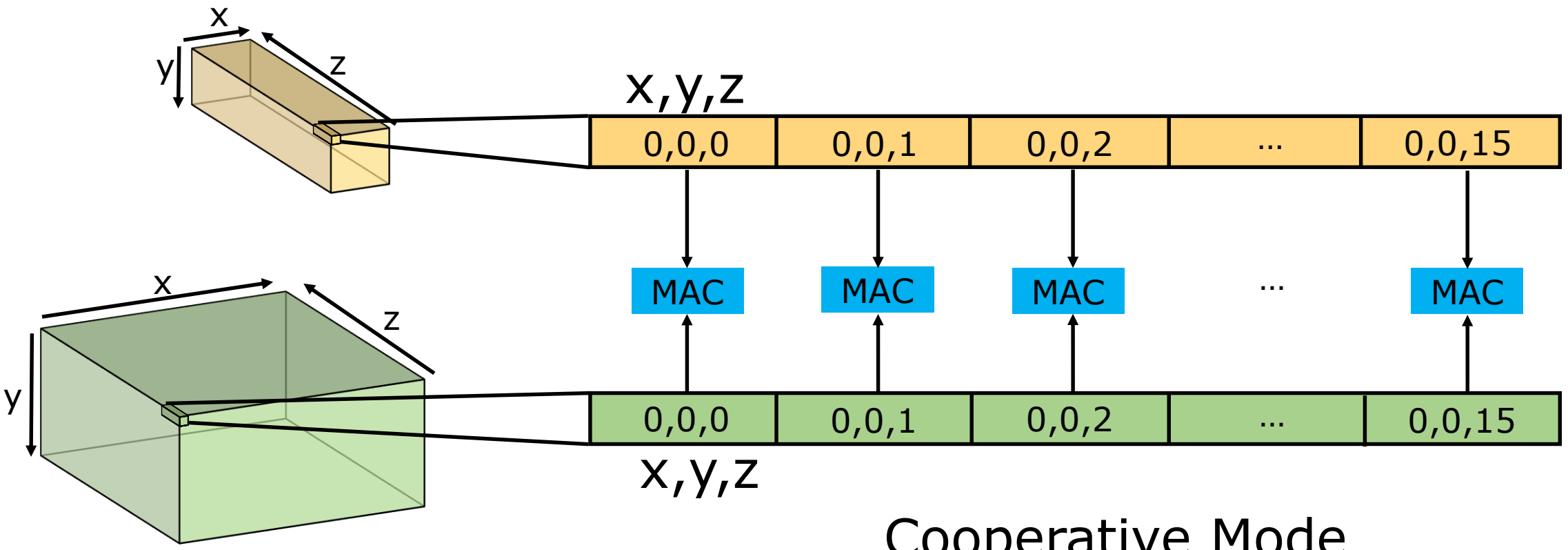
Vec 1 | Vec 2 | … | Vec M

Subdivided into M vectors

Data Organization



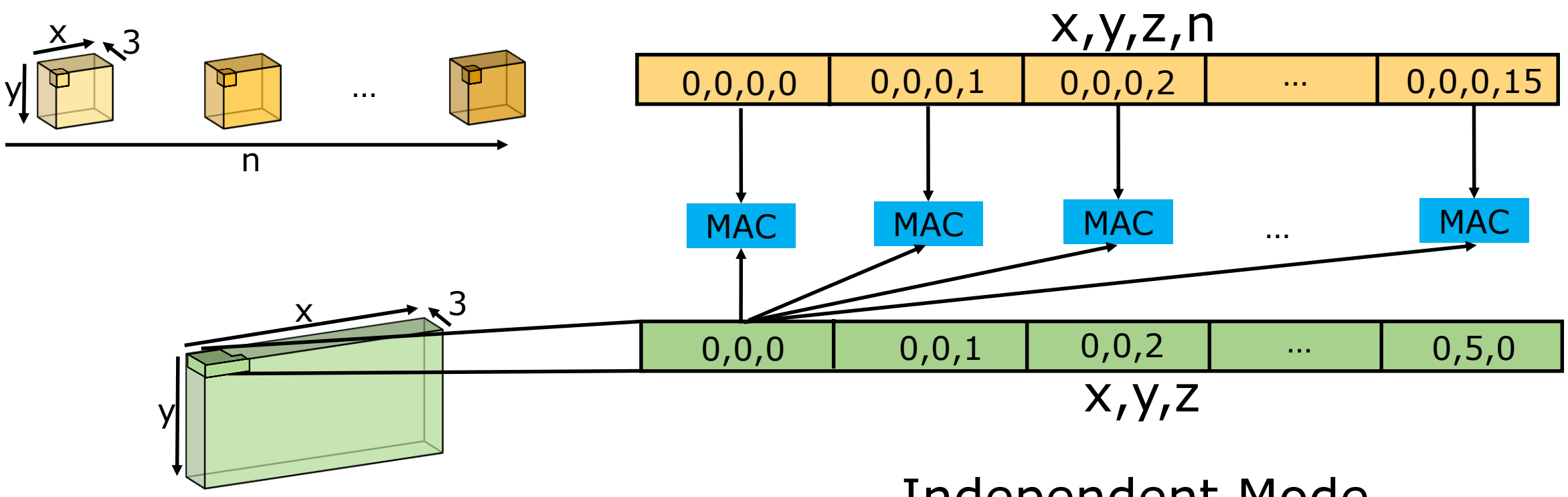
Intra-map, Intra-activation



Cooperative Mode

- Kernel shared by all MACs
- Single bias but need to reduce partials

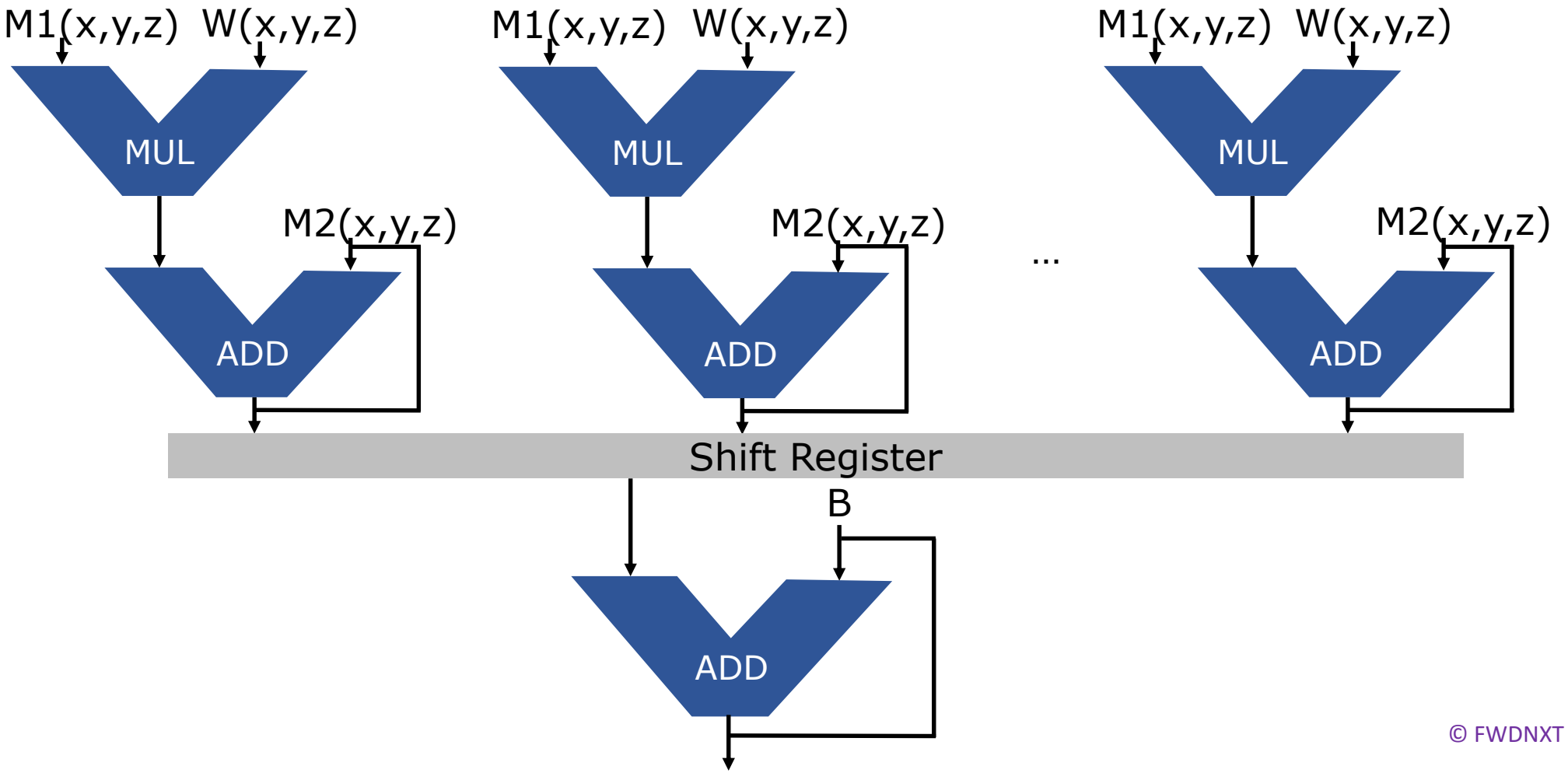
Inter-map



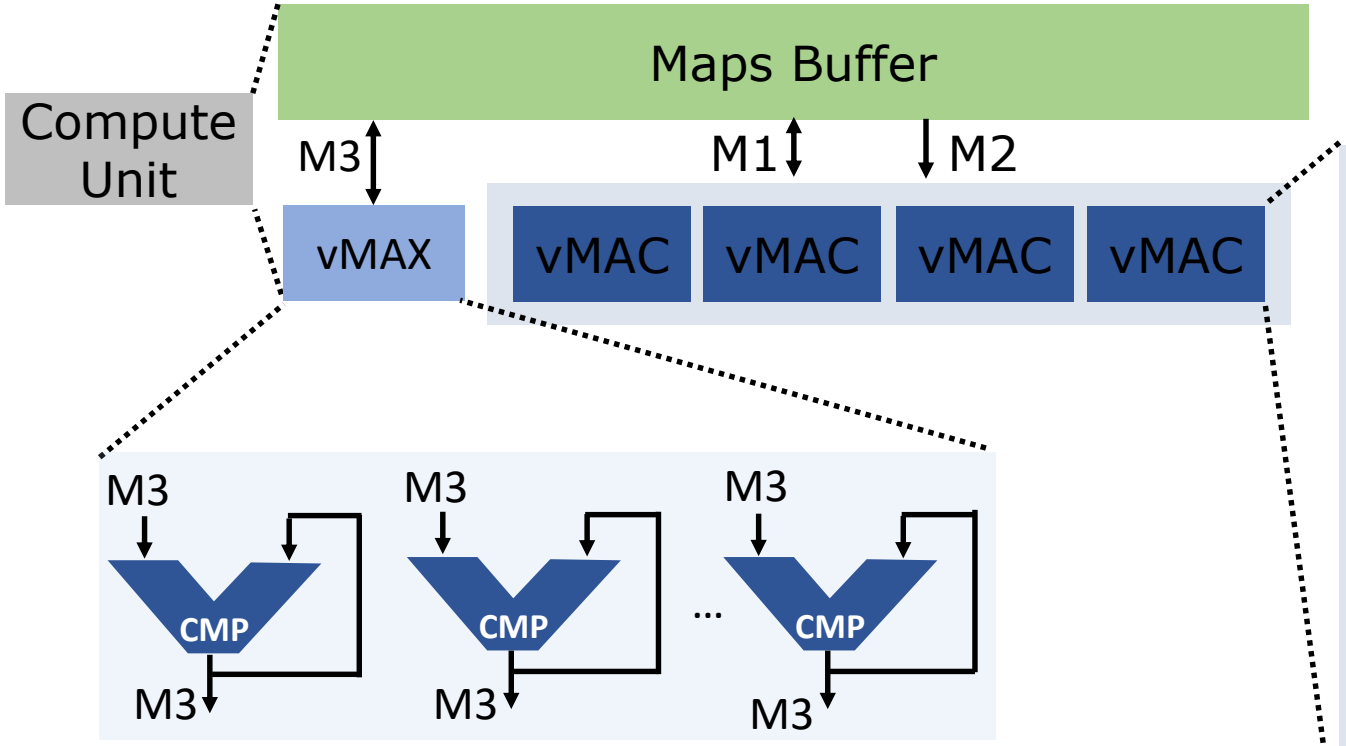
Independent Mode

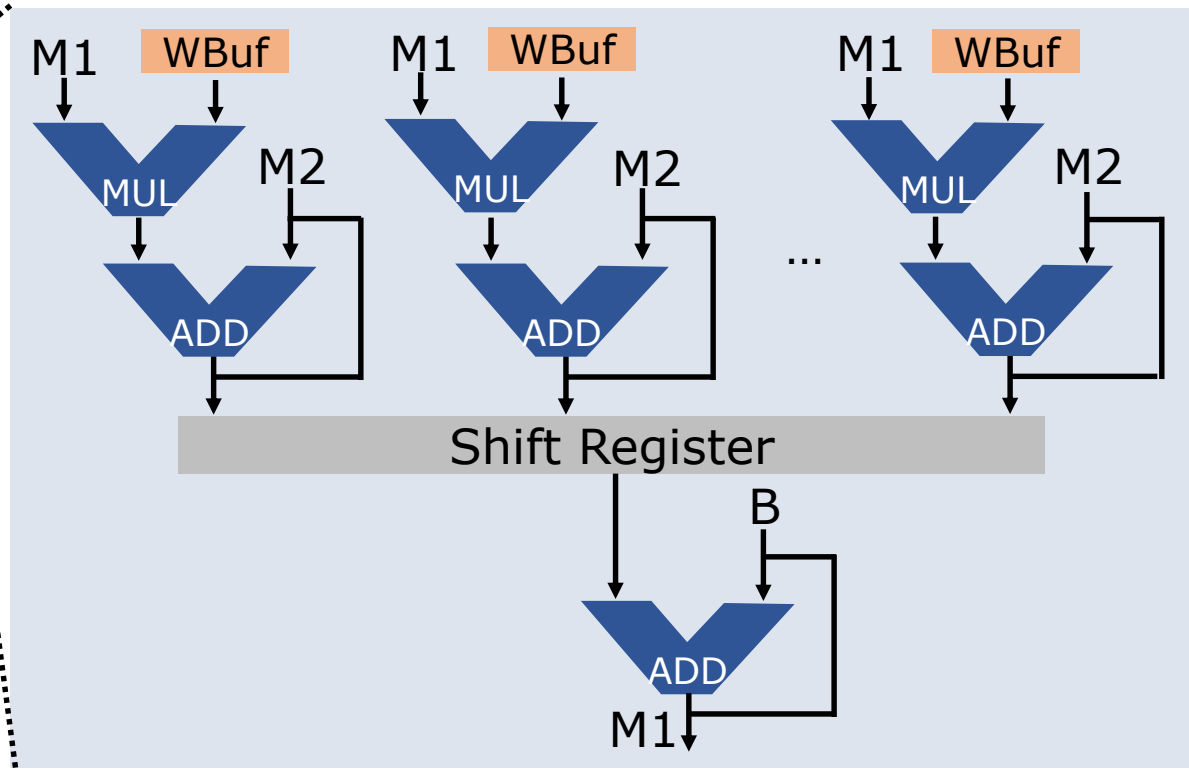
- Independent kernel per MAC
- 16 bias values but no reduction op

Vector Multiply-Accumulate (vMAC)



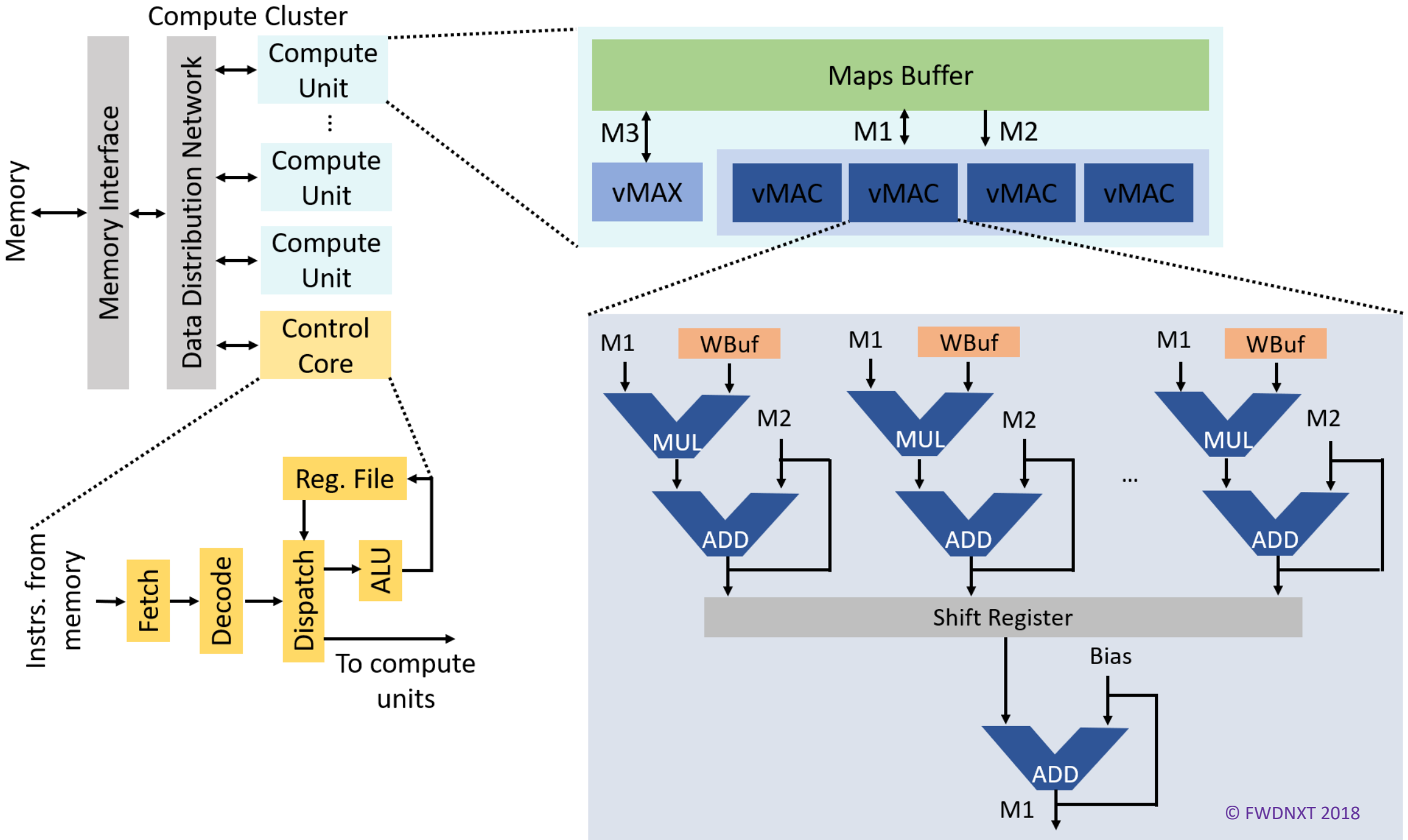
Scaling Up with Compute Units





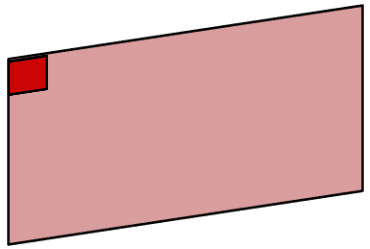
Compute unit

- 4 vMACs per CU
- 16 MACs per vMAC
- 1 KB weights buffer per MAC
- 64 KB maps (double) buffer per CU


Compute Cluster
Memory
Memory Interface
Data Distribution Network
Compute Unit
Compute Unit
Compute Unit
Control Core
Maps Buffer
M3
M1
M2
vMAX
vMAC
vMAC
vMAC
vMAC
Instrs. from memory
Fetch
Decode
Dispatch
Reg. File
ALU
To compute units
M1
WBuf
MUL
M2
ADD
M1
WBuf
MUL
M2
ADD
...
M1
WBuf
MUL
M2
ADD
Shift Register
Bias
ADD
M1
© FWDNXT 2018

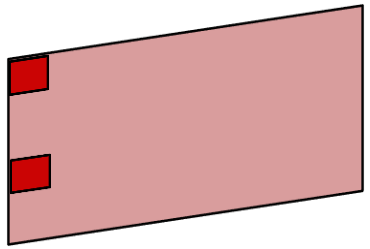
Types of Parallelism Revisited

Intra-map, intra-activation



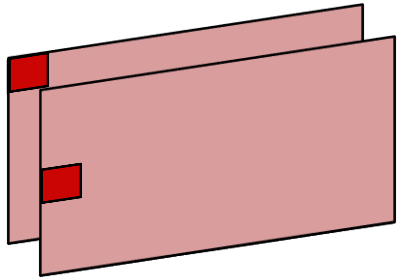
A vMAC in COOP mode

Intra-map, inter-activation



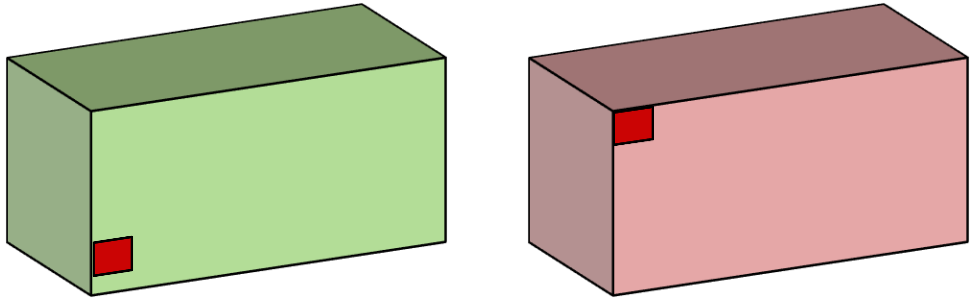
vMACs across CUs

Inter-map



vMACs within a CU

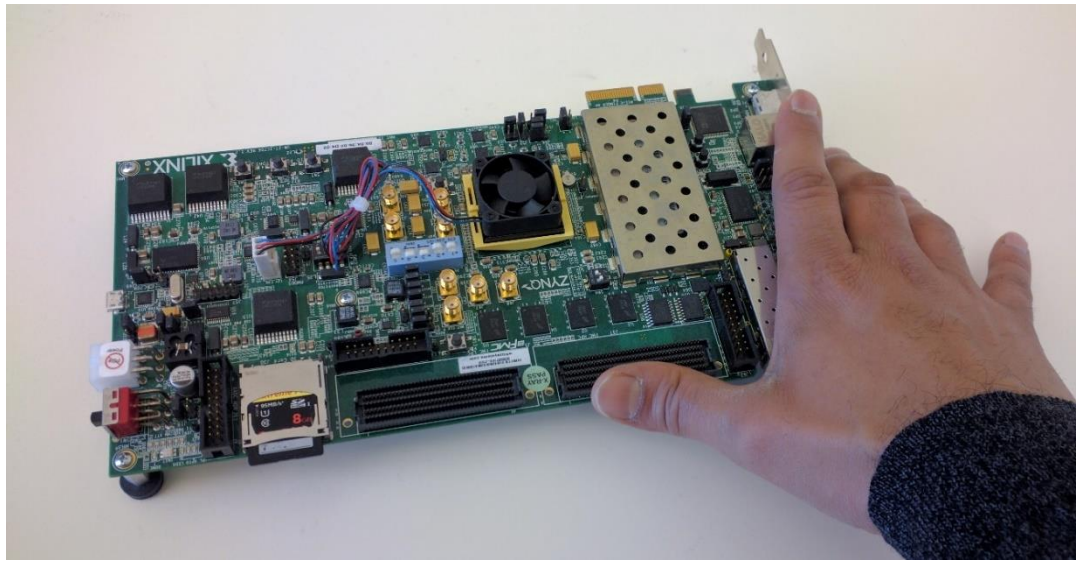
Inter-layer



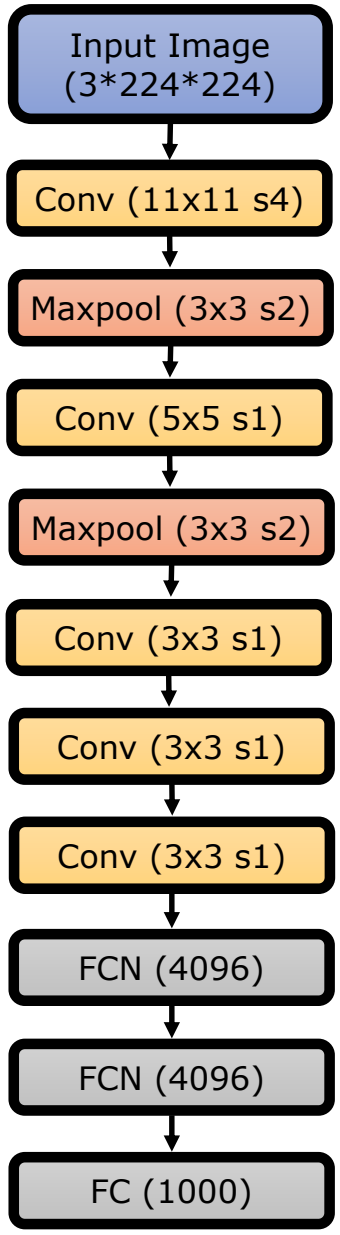
vMACs across clusters

System Specifications

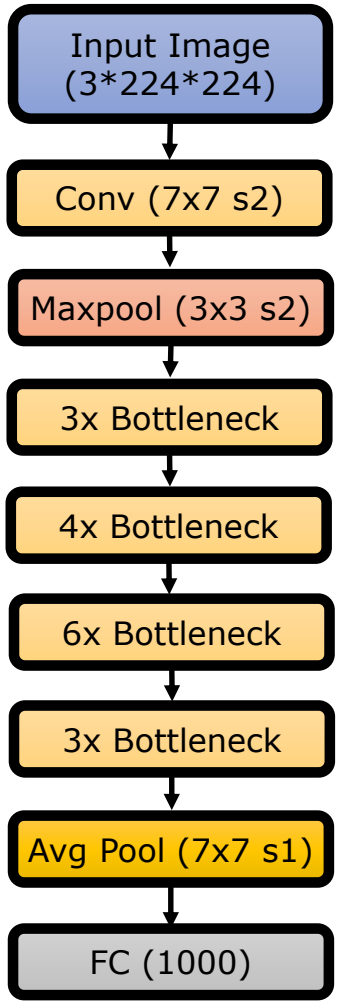
Host CPU	2x ARM Cortex-A9 @800 MHz
Accelerator cores	256 MAC units @ 250 MHz
Peak Throughput	128 G-ops/s
Memory	1GB DDR3 @ 533 MHz
Memory B/W	4.2 GB/s
Power (Board)	12 W
Power (Zynq + mem)	7 W



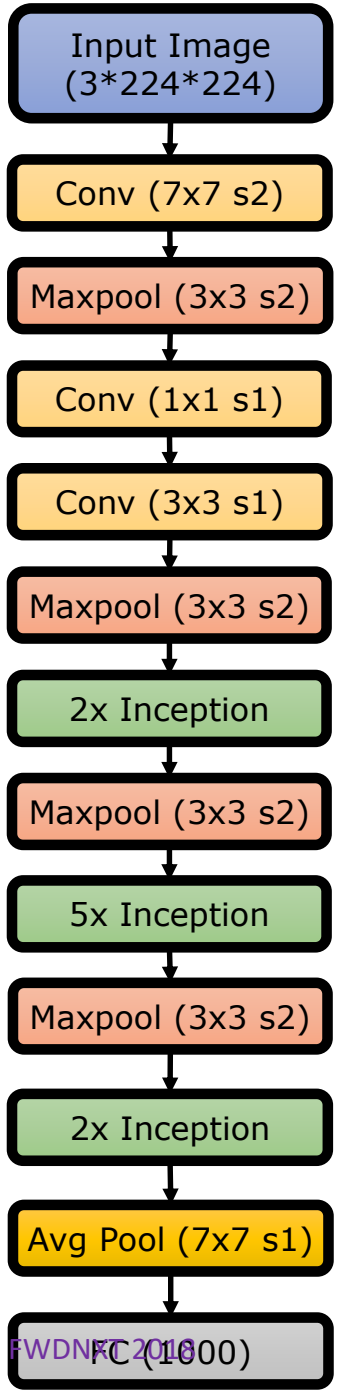
Benchmarks



AlexNet
1.4 G-ops/frame

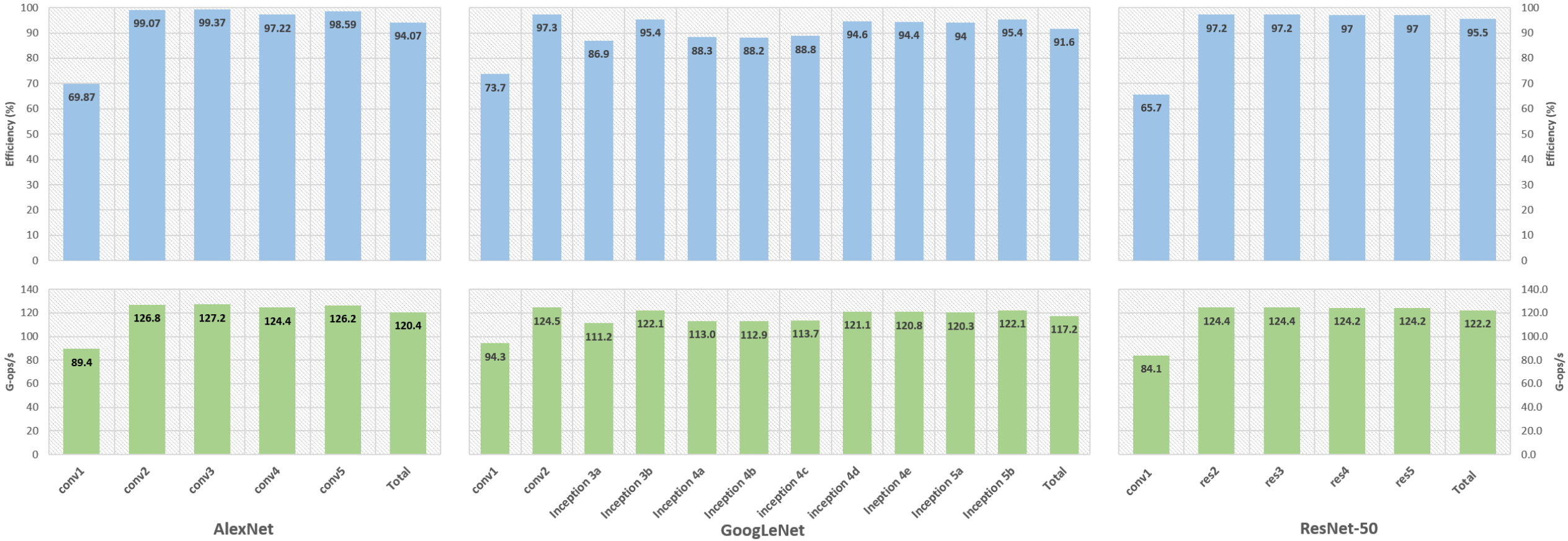


ResNet-50
7.7 G-ops/frame

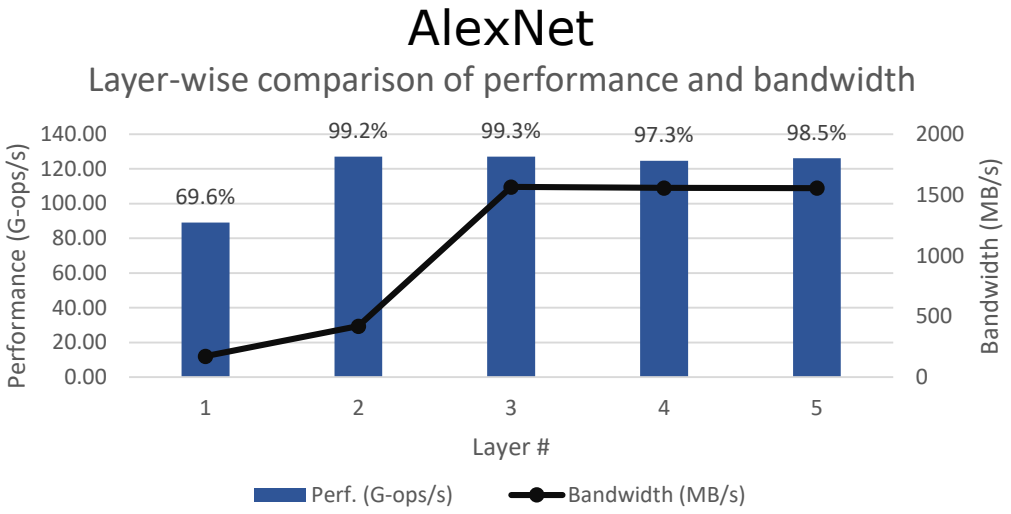


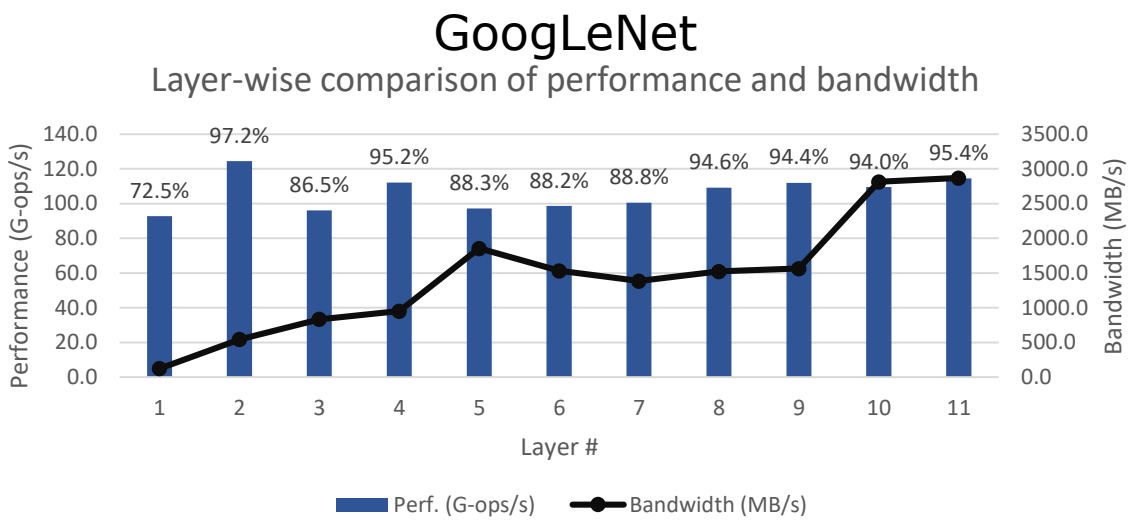
GoogLeNet
3.2 G-ops/frame

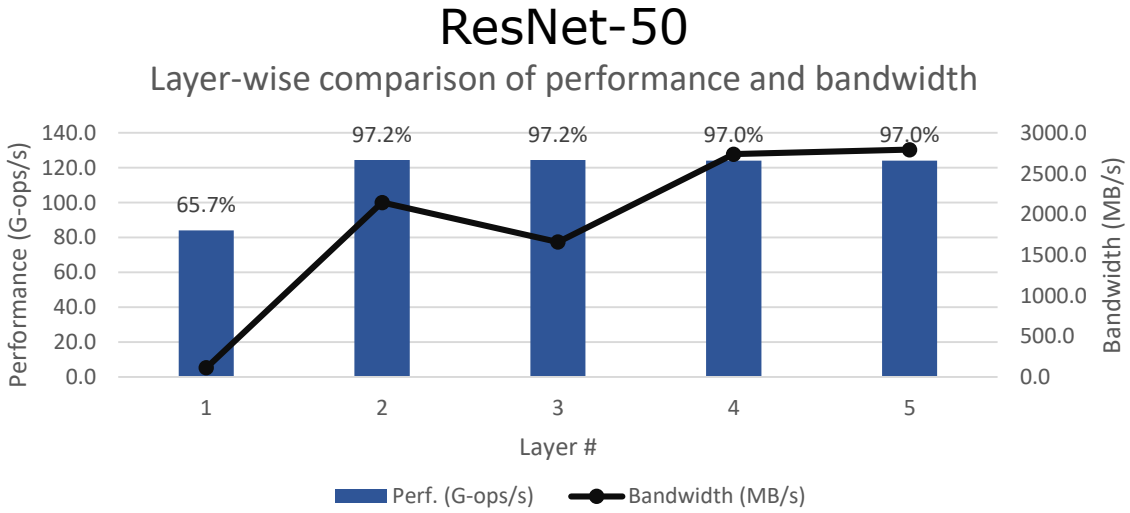
Performance



Comparison of Perf. and B/W







Classification Results (top-5)

ambulance, minivan, minibus, golfcart, motor scooter

car, motorcycle, bicycle, watch, shoe

jaguar, dalmatian, banded gecko, leopard, bonnet

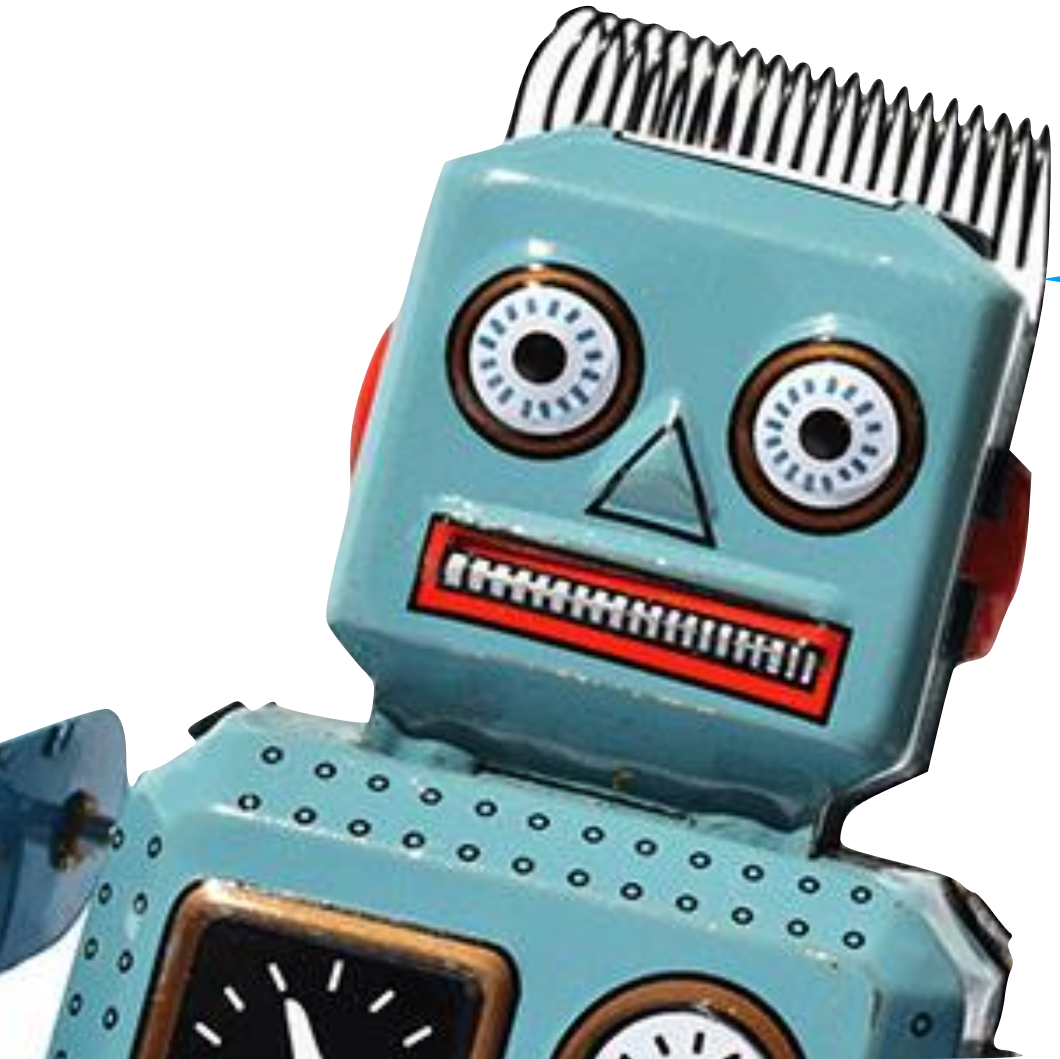
lionfish, jellyfish, sea slug, sea anemone, chambered nautilus

koala, wombat, sloth bear, mongoose, madagascar cat

plastic bag, cauliflower, broccoli, swab, zucchini

motorcycle, bicycle, car, toy, watch

lion, cougar, hippopotamus, chimpanzee, book jacket



Thank you