

Main Focus

Computation Unit

- Upon request, compute feature map for the next round
- Channel-based parallelism
- Accumulate PE outputs to the corresponding output channel

PE Array

- 8-bit fixed-point multiplication
- Each PE capable of 32 multiplication
- 32 PEs
- Report usage and cycles if possible

Ask data from
computation unit

No

Is the
output ready?

Yes

Output Buffer

Post-Processing

- Add bias
- Pooling
- Up Sampling
- Concatenation
- Store layer by stack

Mem write

Ask for output
of the next layer

No

Is the current
feature map the final
result?

Empty Prime

Yes

Done!

Data Feeds

Final Round?

Yes

load from
buffer

IM Buffer

Load next
after used

Kernel Buffer

Mem Read

VMem (External Memory)

- Stores all weight and bias
- Serves the current computed feature map