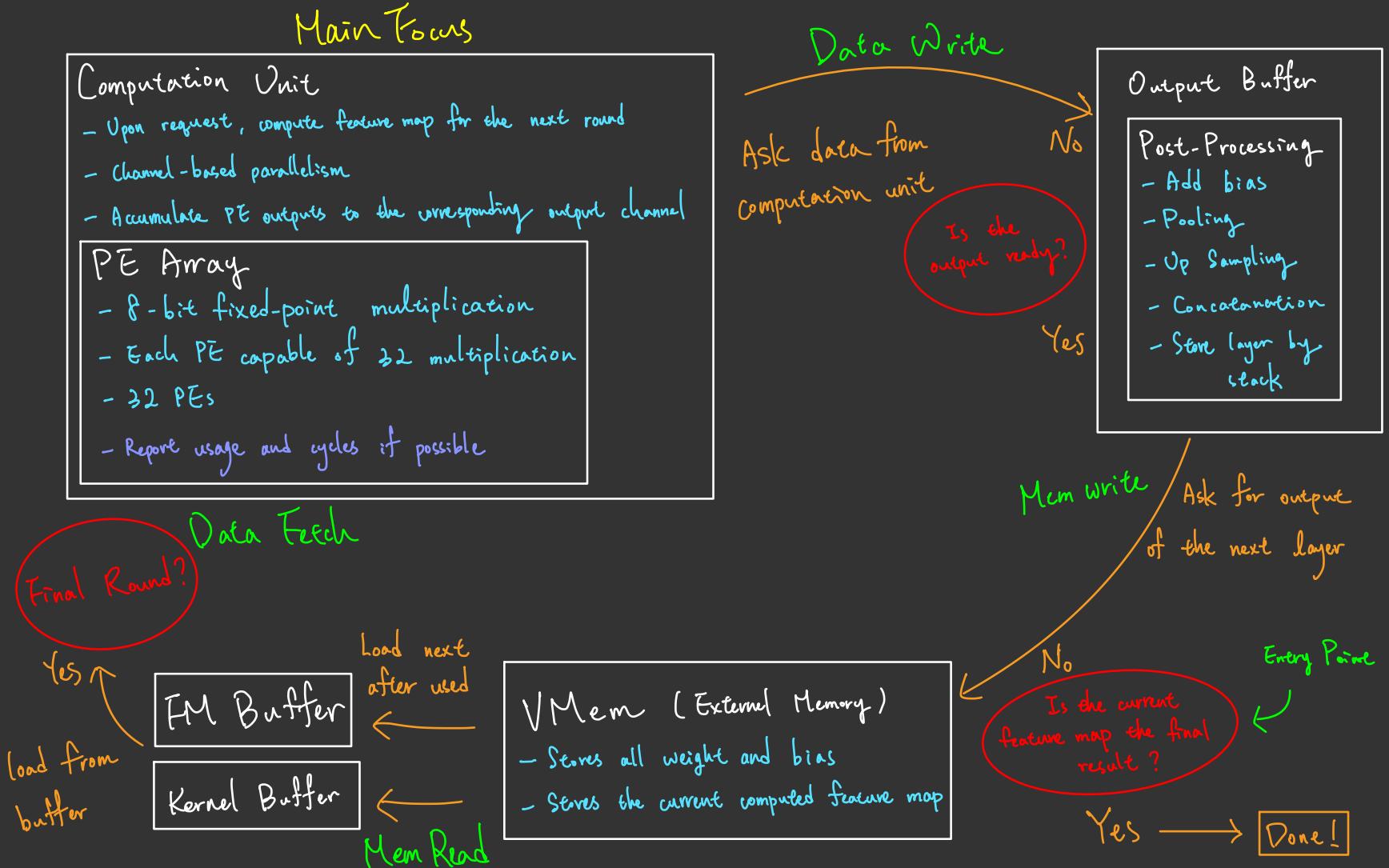
Main Focus
Computation Unit
- Upon request, compute feature map for the next round
- Channel-based parallelism
- Accumulate PE outputs to the corresponding output channel
PE Array
- 8-bit fixed-point multiplication
- Each PE capable of 32 multiplication
- 32 PEs
- Report usage and cycles if possible
Data Write
Ask data from computation unit
Is the output ready?
No
Yes
Output Buffer
Post-Processing
- Add bias
- Pooling
- Up Sampling
- Concatenation
- Store layer by stack
Mem write
Ask for output of the next layer
Entry Point
No
Is the current feature map the final result ?
Yes
Done!
VMem (External Memory)
- Stores all weight and bias
- Stores the current computed feature map
Load next after used
FM Buffer
Kernel Buffer
Mem Read
Yes
load from buffer
Final Round?
Data Fetch