Lecture 18
Case Study: Eyeriss (Part 1)

Xuan ‘Silvia’ Zhang
Washington University in St. Louis

http://classes.engineering.wustl.edu/ese566/

Eyeriss (ISCA 2016)

- ISCA
  - International Symposium on Computer Architecture
  - top-tier conference on computer (micro)architecture

- Spatial Architecture for CNN
  - the most popular architecture for NN acceleration
  - a systematic approach for parameter reuse

Generalized Spatial Architecture

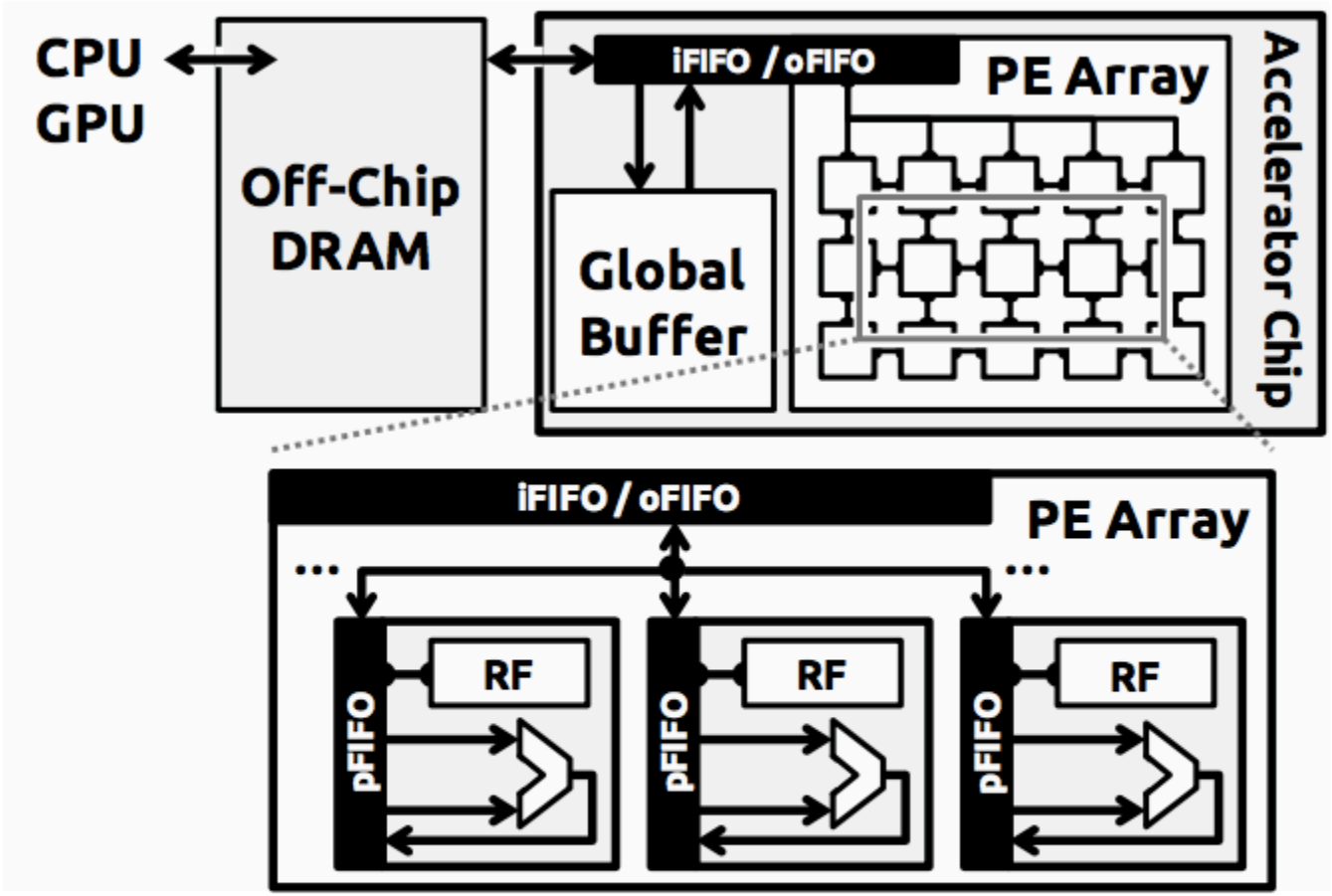


Figure 1. Block diagram of a general CNN accelerator system consisting of a spatial architecture accelerator and an off-chip DRAM. The zoom-in shows the high-level structure of a PE.

Generalized Algorithm Abstraction

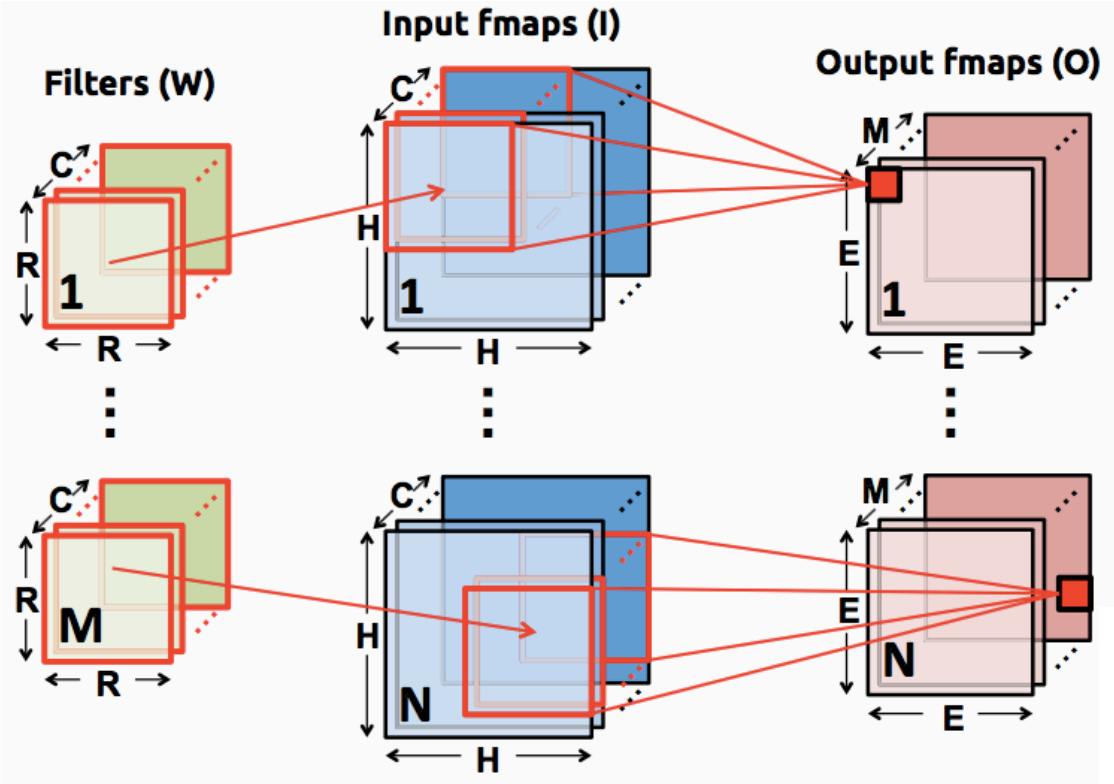


Figure 2. Computation of a CONV/FC layer.

Shape Parameter	Description
N	batch size of 3D fmaps
M	# of 3D filters / # of ofmap channels
C	# of ifmap/filter channels
H	ifmap plane width/height
R	filter plane width/height (= H in FC)
E	ofmap plane width/height (= 1 in FC)

Table I
SHAPE PARAMETERS OF A CONV/FC LAYER.

$$\mathbf{O}[z][u][x][y] = \mathbf{B}[u] + \sum_{k=0}^{C-1} \sum_{i=0}^{R-1} \sum_{j=0}^{R-1} \mathbf{I}[z][k][Ux+i][Uy+j] \times \mathbf{W}[u][k][i][j],$$

$$0 \le z < N, 0 \le u < M, 0 \le x, y < E, E = (H - R + U)/U.$$

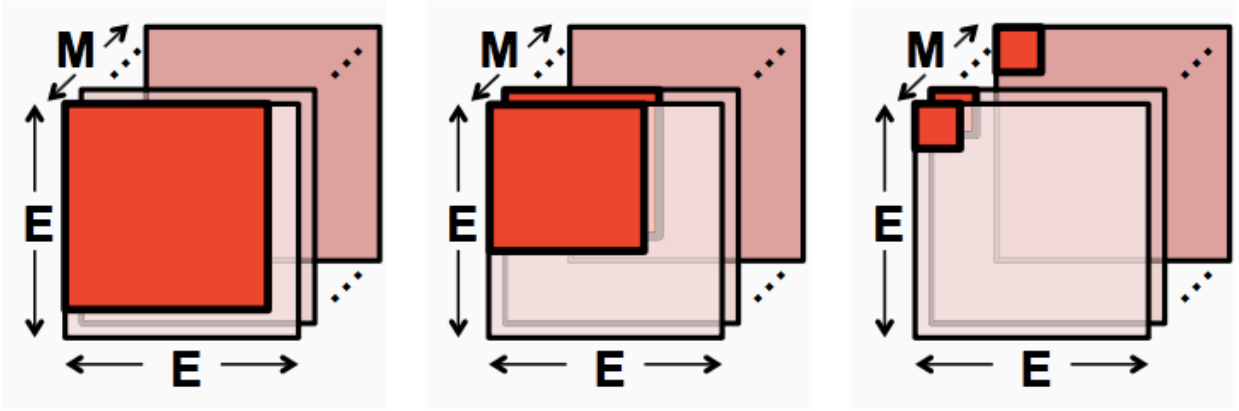
Data Handling Comparison

- Convolutional reuse
  - filter weight: E2
  - ifmap pixel: R2
- Filter reuse
  - across batch size of N
- ifmap reuse
  - across M filter channels

Data Handling Comparison

Dataflow	Data Handling
WS	Maximize *convolutional reuse* and *filter reuse* of weights in the RF.
SOC-MOP OS	Maximize *psum accumulation* in RF. *Convolutional reuse* in array.
MOC-MOP OS	Maximize *psum accumulation* in RF. *Convolutional reuse* and *ifmap reuse* in array.
MOC-SOP OS	Maximize *psum accumulation* in RF. *Ifmap reuse* in array.
NLR	*Psum accumulation* and *ifmap reuse* in array.

Table III
DATA HANDLING COMPARISON BETWEEN EXISTING DATAFLOWS.



(a) (b) (c)

Figure 3. Comparison of the three different OS dataflow variants: (a) SOC-MOP, (b) MOC-MOP, and (c) MOC-SOP. The red blocks depict the ofmap region that the OS dataflow variants process at once.

Proposed Row Stationary Technique

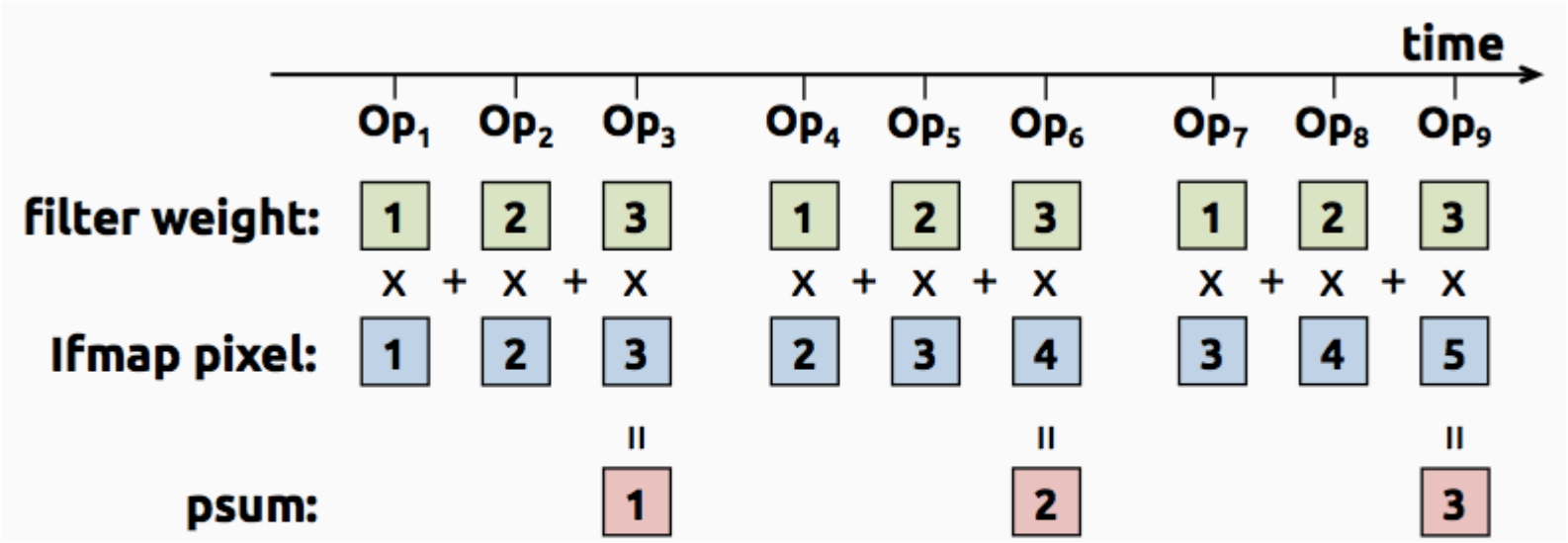
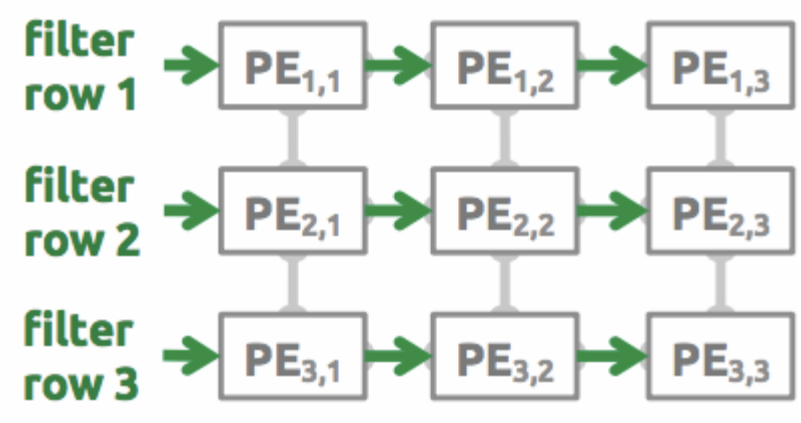
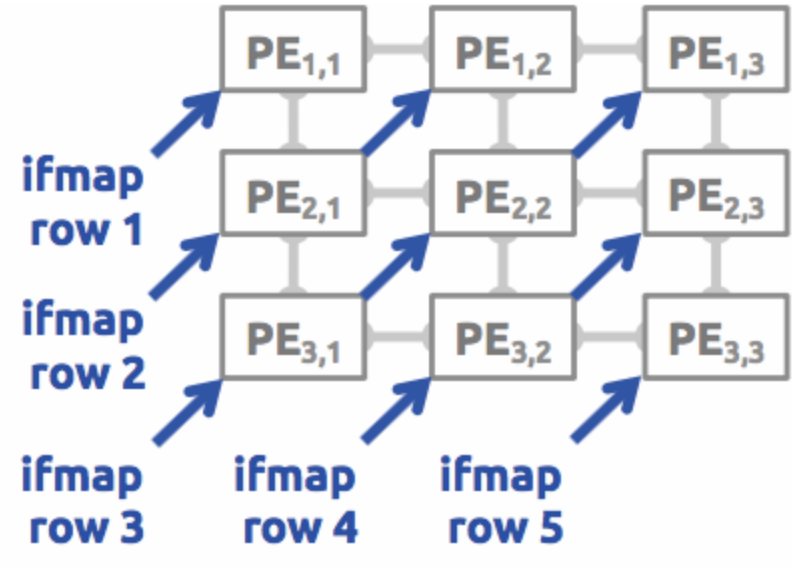


Figure 5. Processing of an 1D convolution primitive in the PE. In this example, $R = 3$ and $H = 5$.

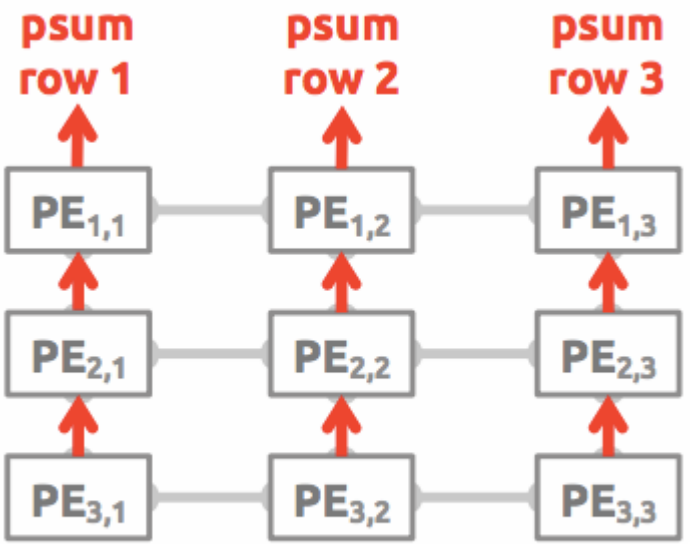
Proposed Row Stationary Technique



(a)



(b)



(c)

Hardware Implementation Trade-off: Area and Energy

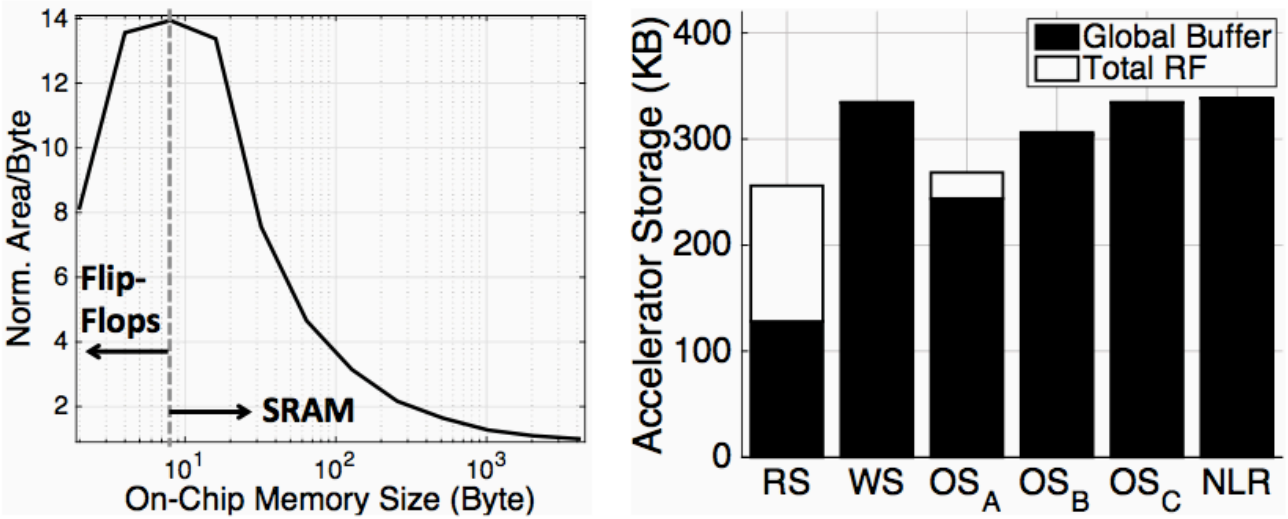


Figure 7. The trade-off between storage area allocation and the total storage size. (a) A smaller memory have a higher cost on area utilization. (b) Due to the area allocation between global buffer and RF, the total on-chip storage size varies between dataflows.

	DRAM	Global Buffer (>100kB)	Array (inter-PE) (1-2mm)	RF (0.5kB)
Norm. Energy	200×	6×	2×	1×

Table IV
NORMALIZED ENERGY COST RELATIVE TO A MAC OPERATION EXTRACTED FROM A COMMERCIAL 65NM PROCESS.

Hierarchical Reuse Examples

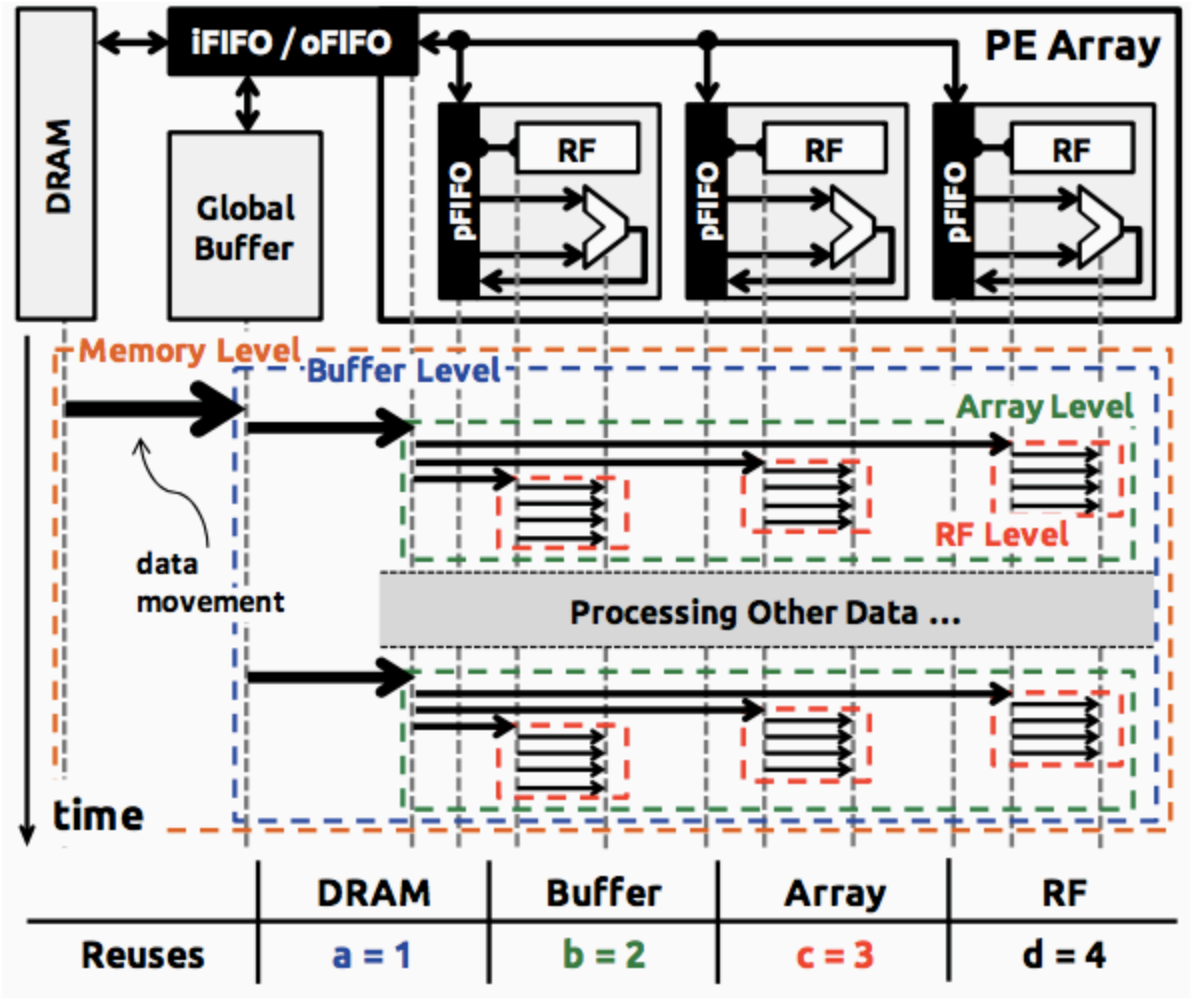


Figure 8. An example of the ifmap pixel or filter weight being reused across four levels of hierarchy.

Hierarchical Reuse Examples

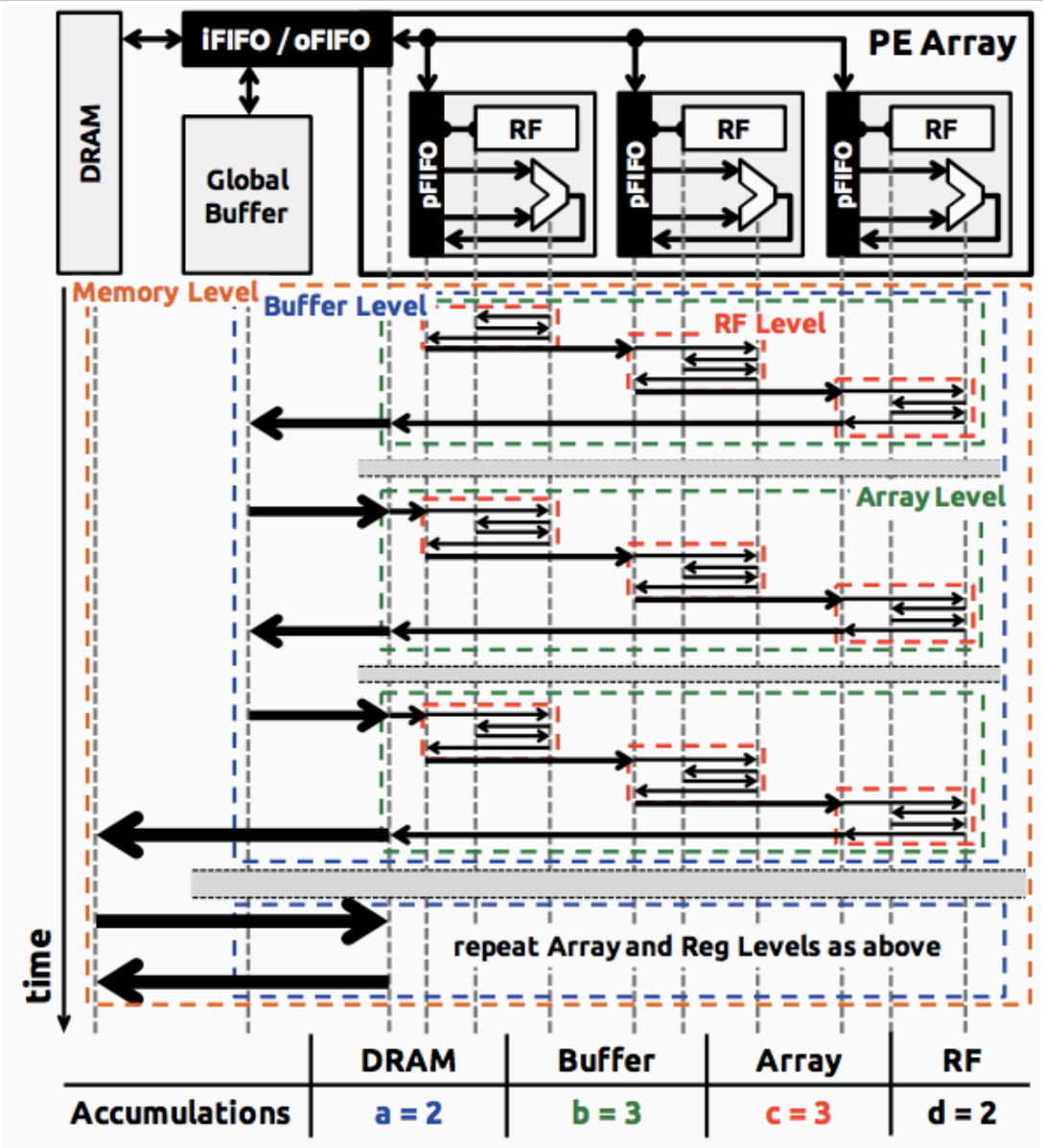


Figure 9. An example of the psum accumulation going through four levels of hierarchy.

Questions?

Comments?

Discussion?