

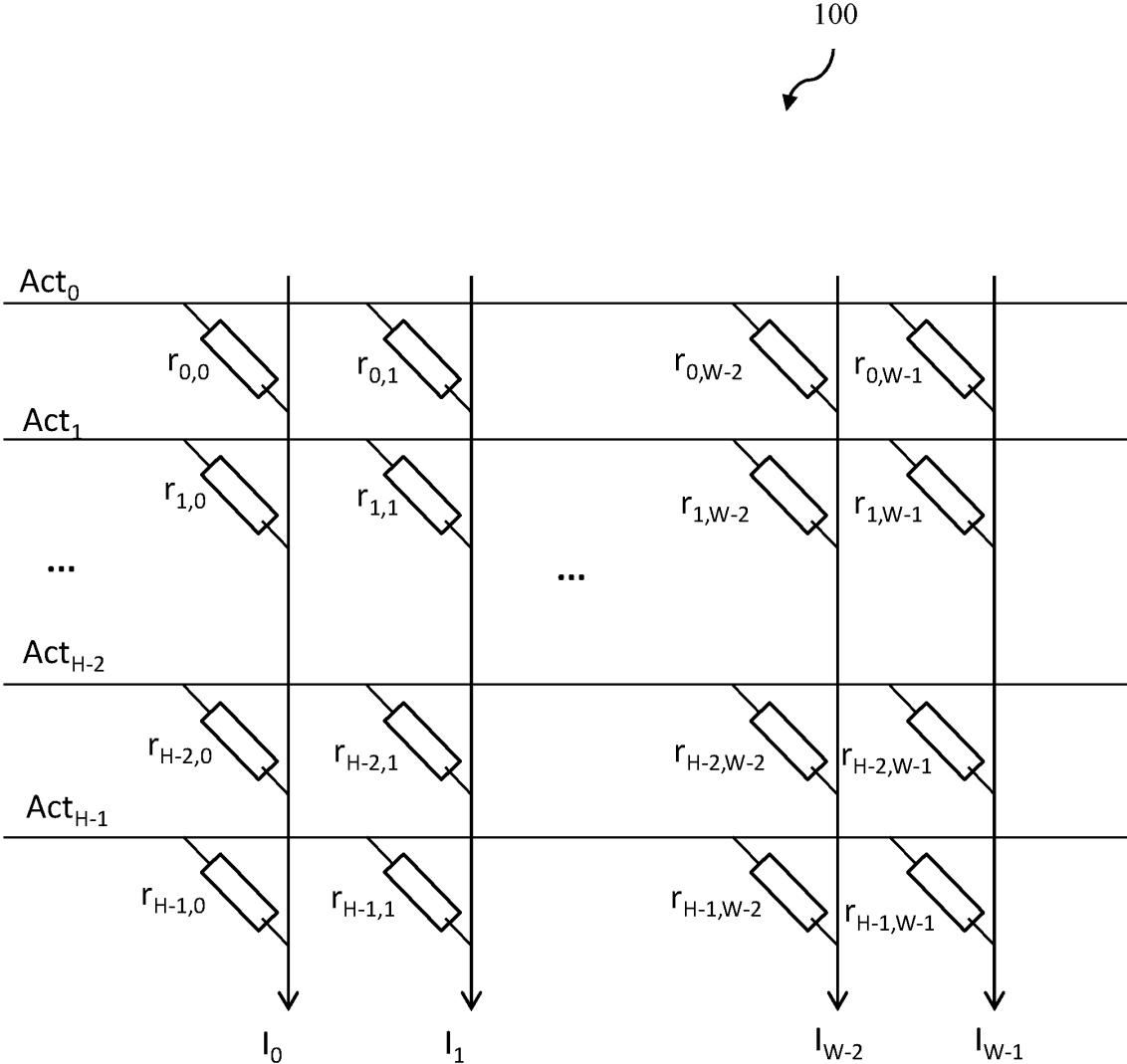


FIG. 1

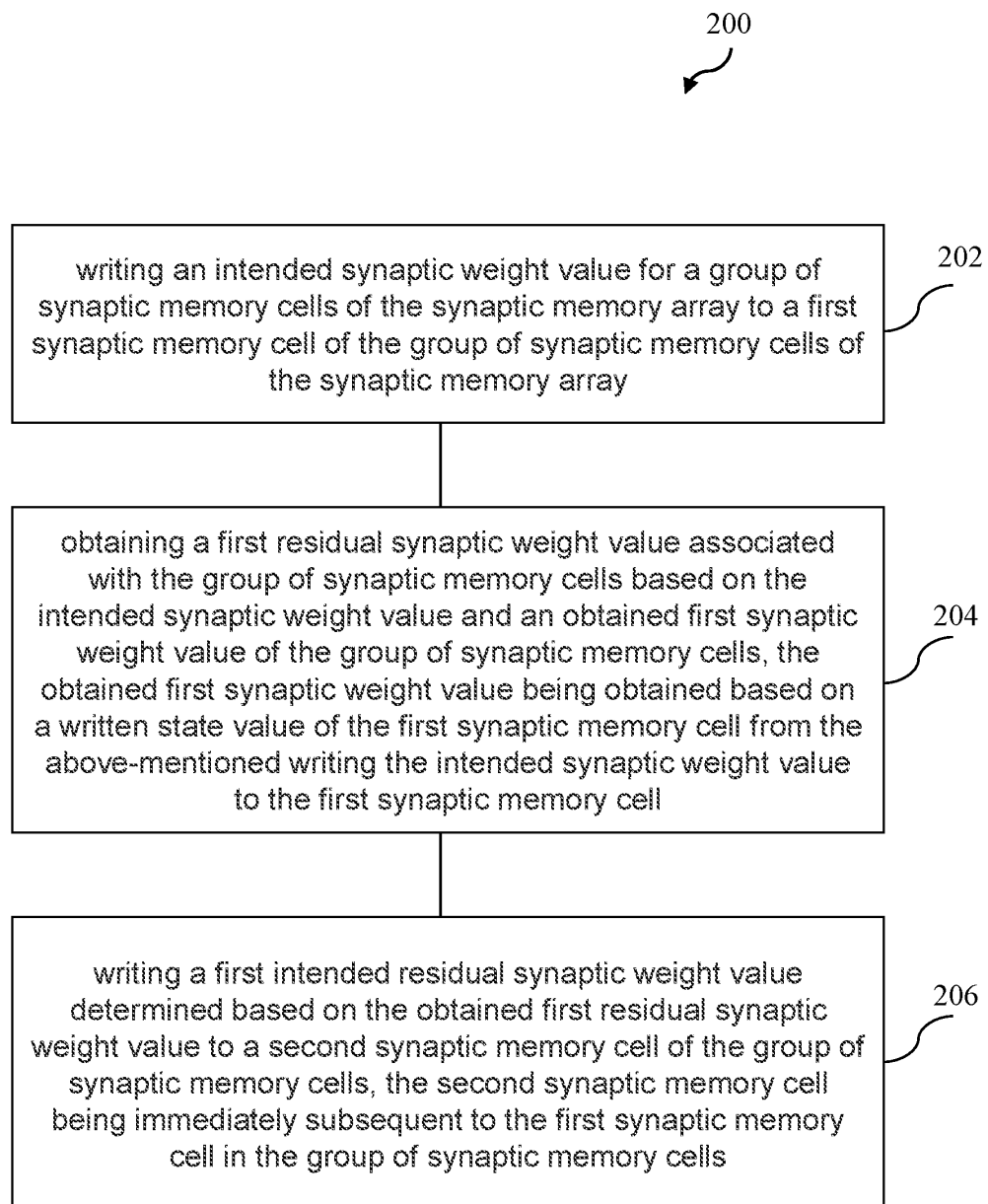


FIG. 2

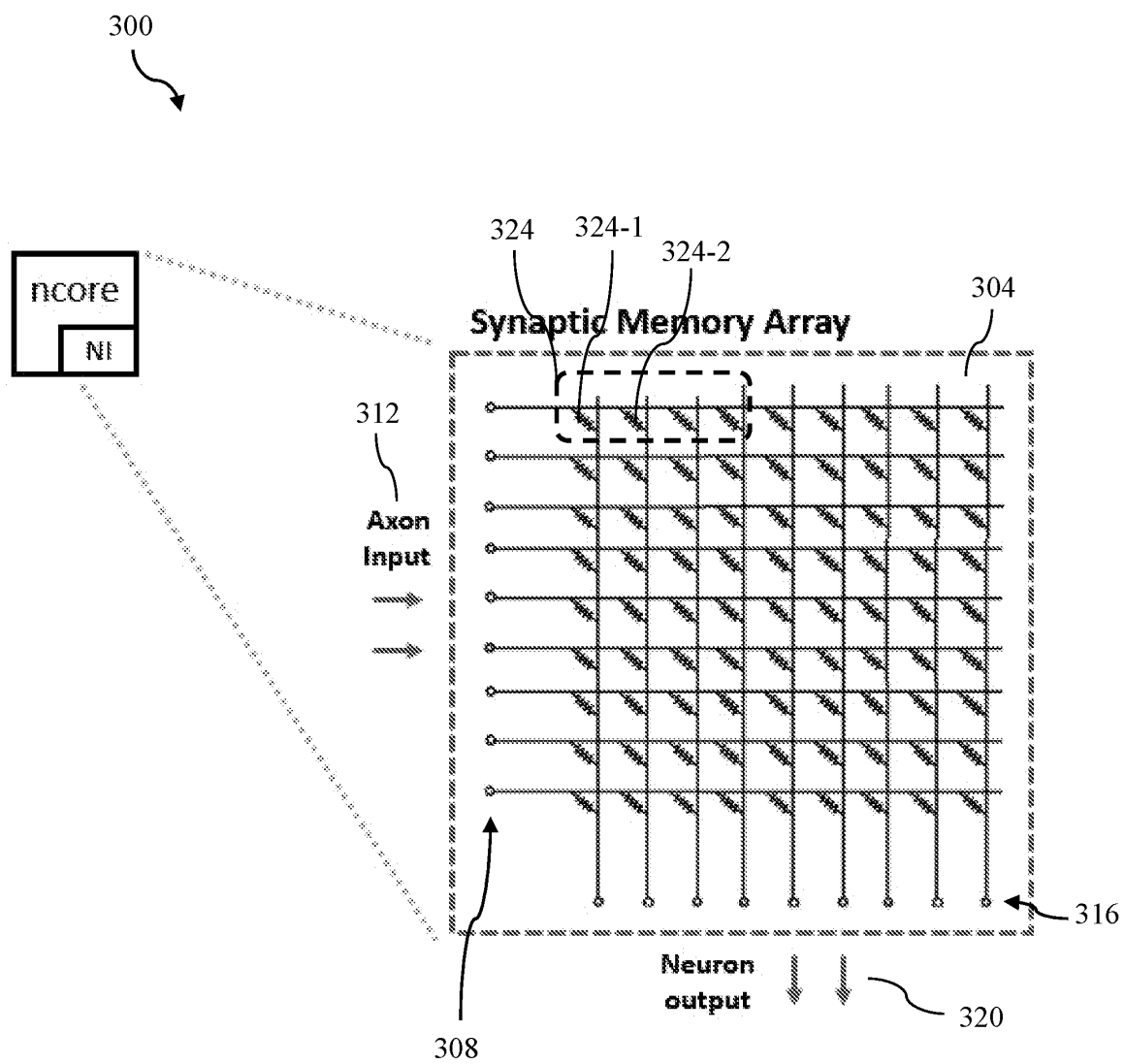
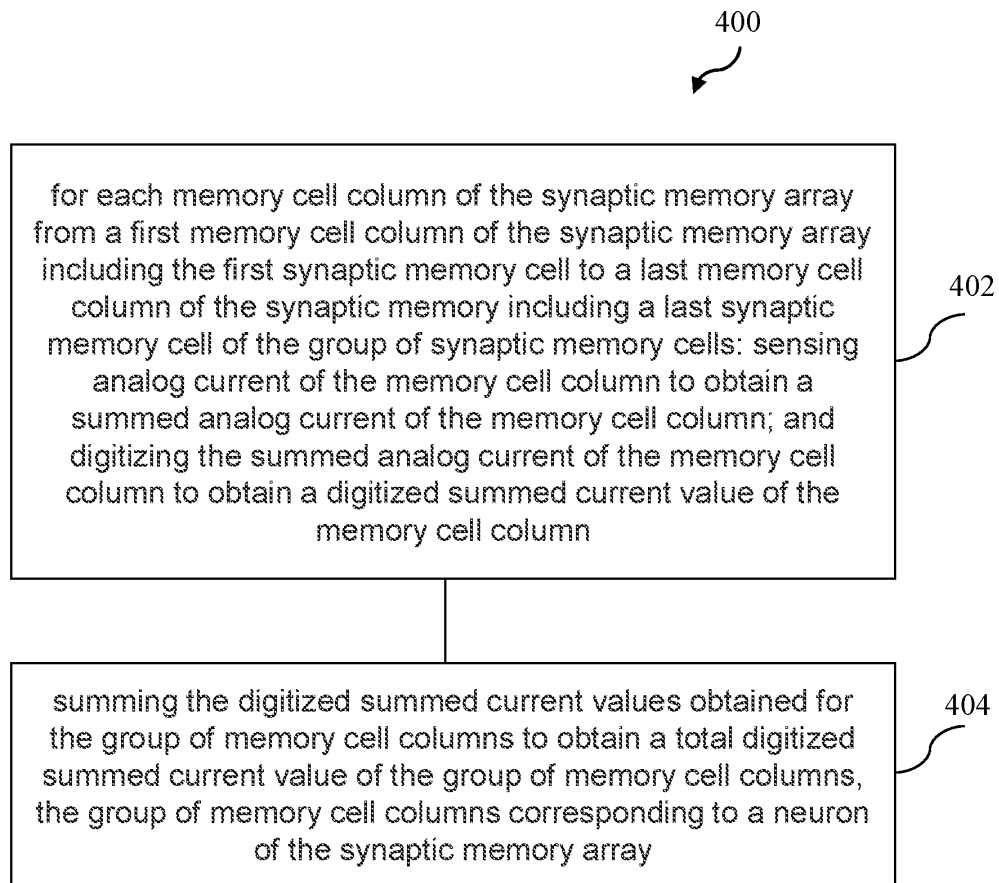


FIG. 3

**FIG. 4**

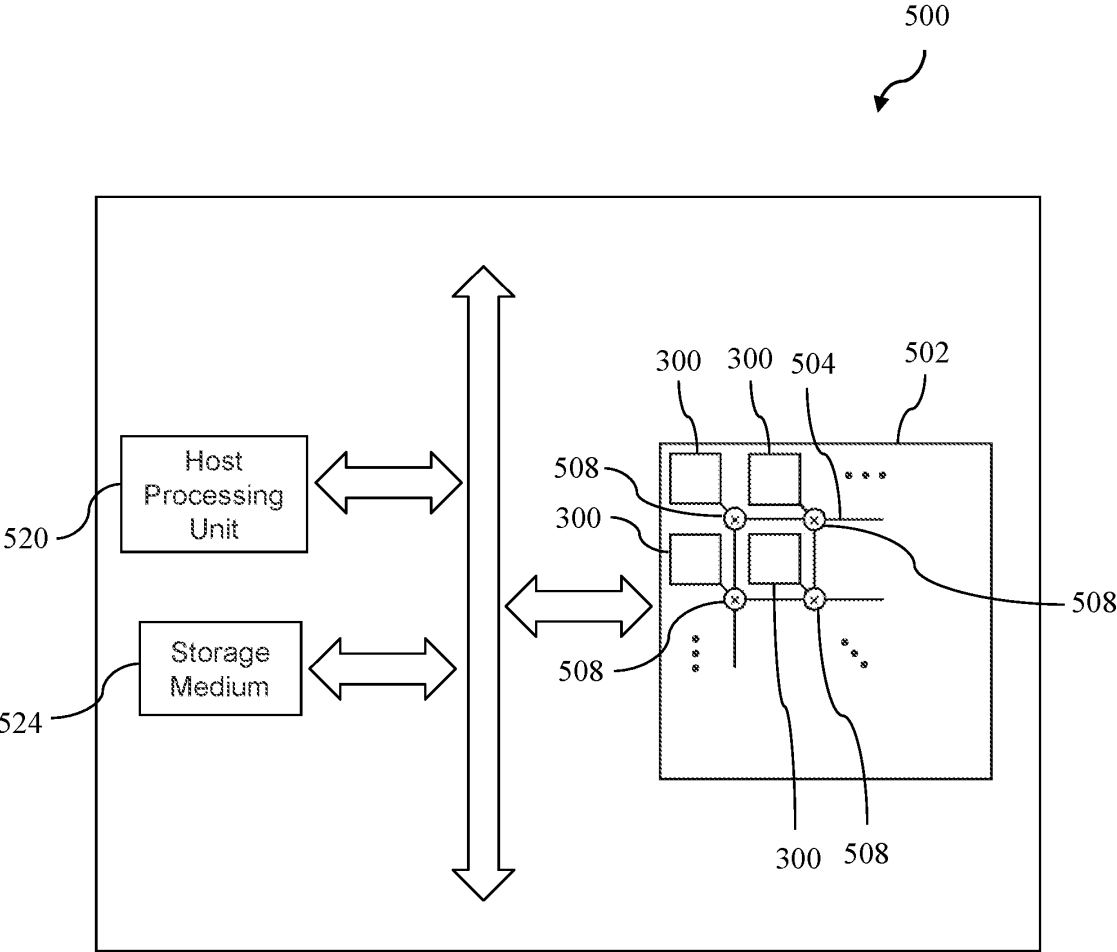


FIG. 5A

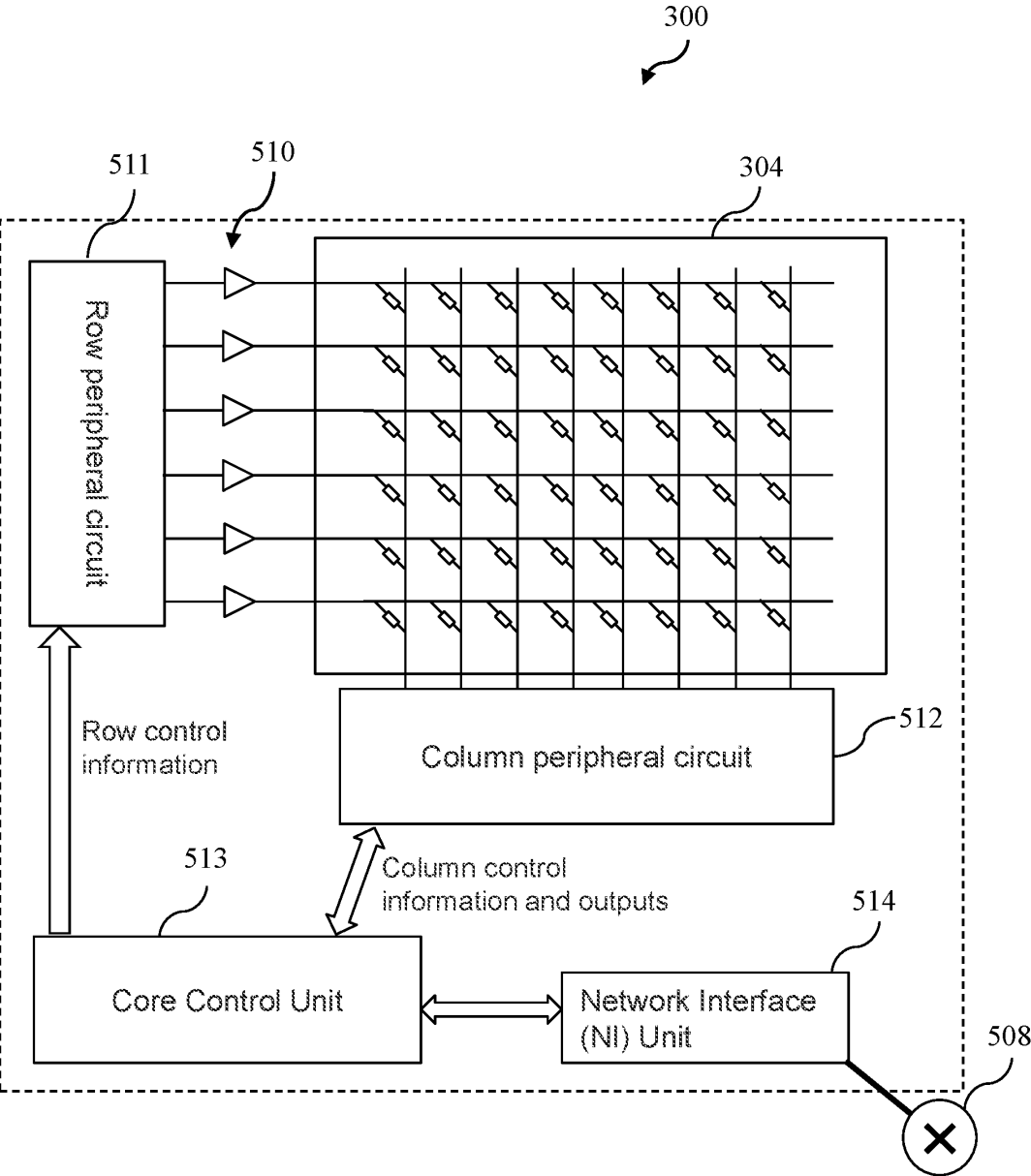


FIG. 5B

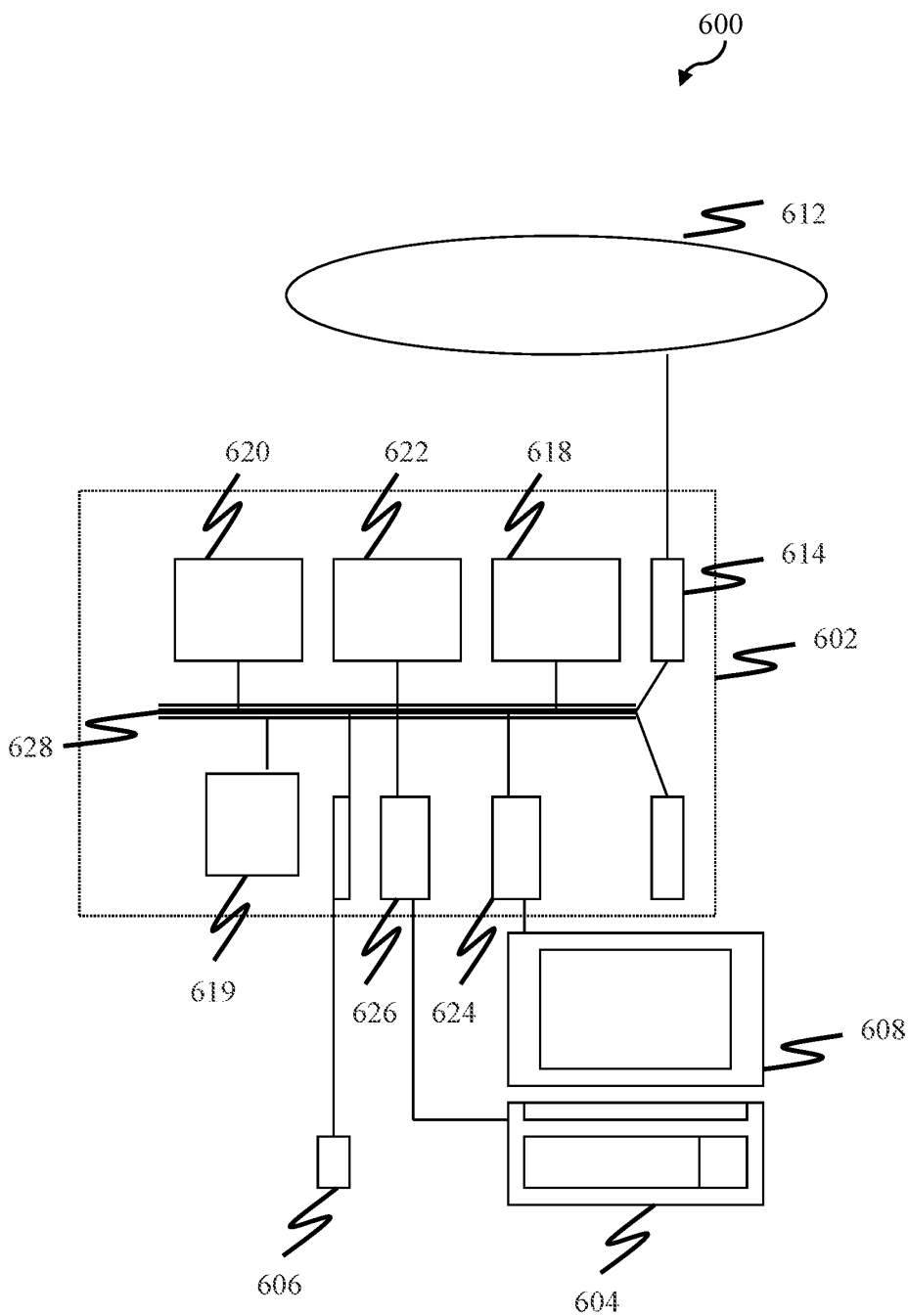
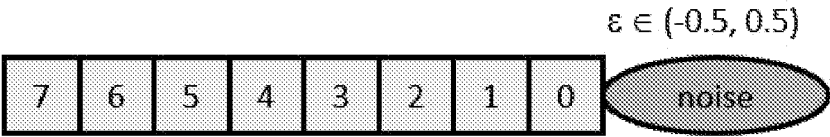
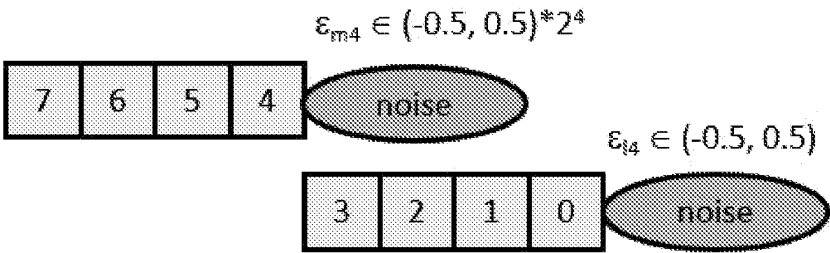


FIG. 6



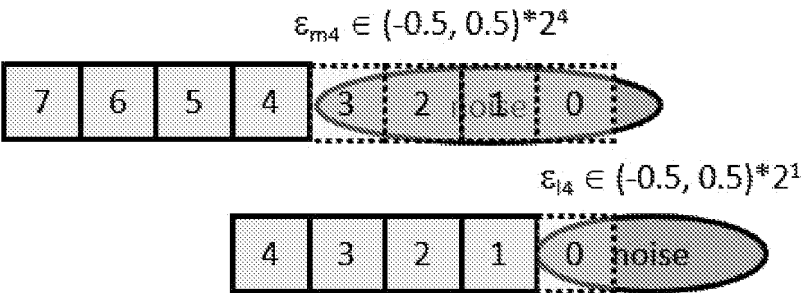
A native 8-bit cell

FIG. 7A



Naïve concatenation of
2 native 4-bit cells

FIG. 7B



Cell-stitching of 2 native
4-bit cells with 1-bit
overlap

FIG. 7C

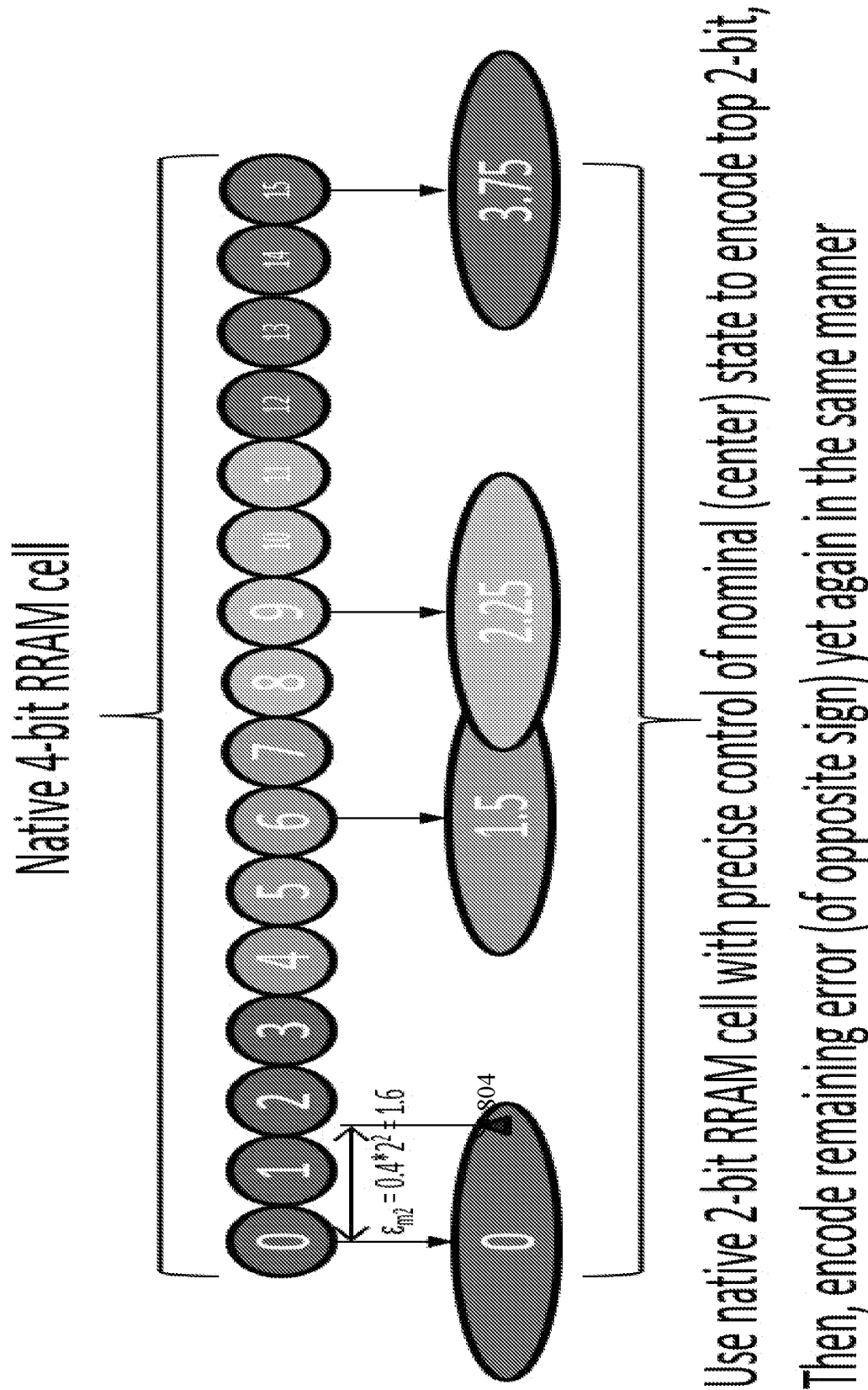


FIG. 8

Table 1

Desired state (4-bit unsigned)	MSB cell (2-bit unsigned), no mid ref nominal, scaling factor 2^2				LSB cell (2-bit signed), mid ref nominal = $2^1-0.5$, scaling factor 2^0				Actual state (4-bit)
	Nominal	Actual	ε	ε_{m2}	Nominal	Actual	ε	ε_{l2}	
0	0	0.4	0.4	1.6	-1.6	-1.9	-0.3	-0.3	-0.3
6	1.5	1.2	-0.3	-1.2	1.2	1.65	0.45	0.45	6.45
9	2.25	2.5	0.25	1.0	-1.0	-1.4	-0.4	-0.4	8.6
15	3.75	3.3	-0.45	-1.8	1.8	2.0	0.2	0.2	15.2

FIG. 9**Table 2**

Desired state (4-bit signed)	MSB cell (2-bit), mid ref nominal = 2^1 , scaling factor 2^2					LSB cell (2-bit), mid ref nominal = $2^1-0.5$, scaling factor 2^0			
	Raw state		Mid ref actual	ε	ε_{m2}	Raw state		Middle ref actual	ε or ε_{l2}
	nominal	actual				nominal	actual		
0 (-8)	0	0.4	1.9	0.5	2.0	-0.5	-0.3	1.4	0.3
6 (-2)	1.5	1.2	2.1	-0.4	-1.6	3.1	3.65	1.6	0.45
9 (1)	2.25	2.5	1.9	0.35	1.4	0.1	0.5	1.4	0.5
15 (7)	3.75	3.4	2.1	-0.45	-1.8	3.3	2.9	1.6	-0.5

FIG. 10

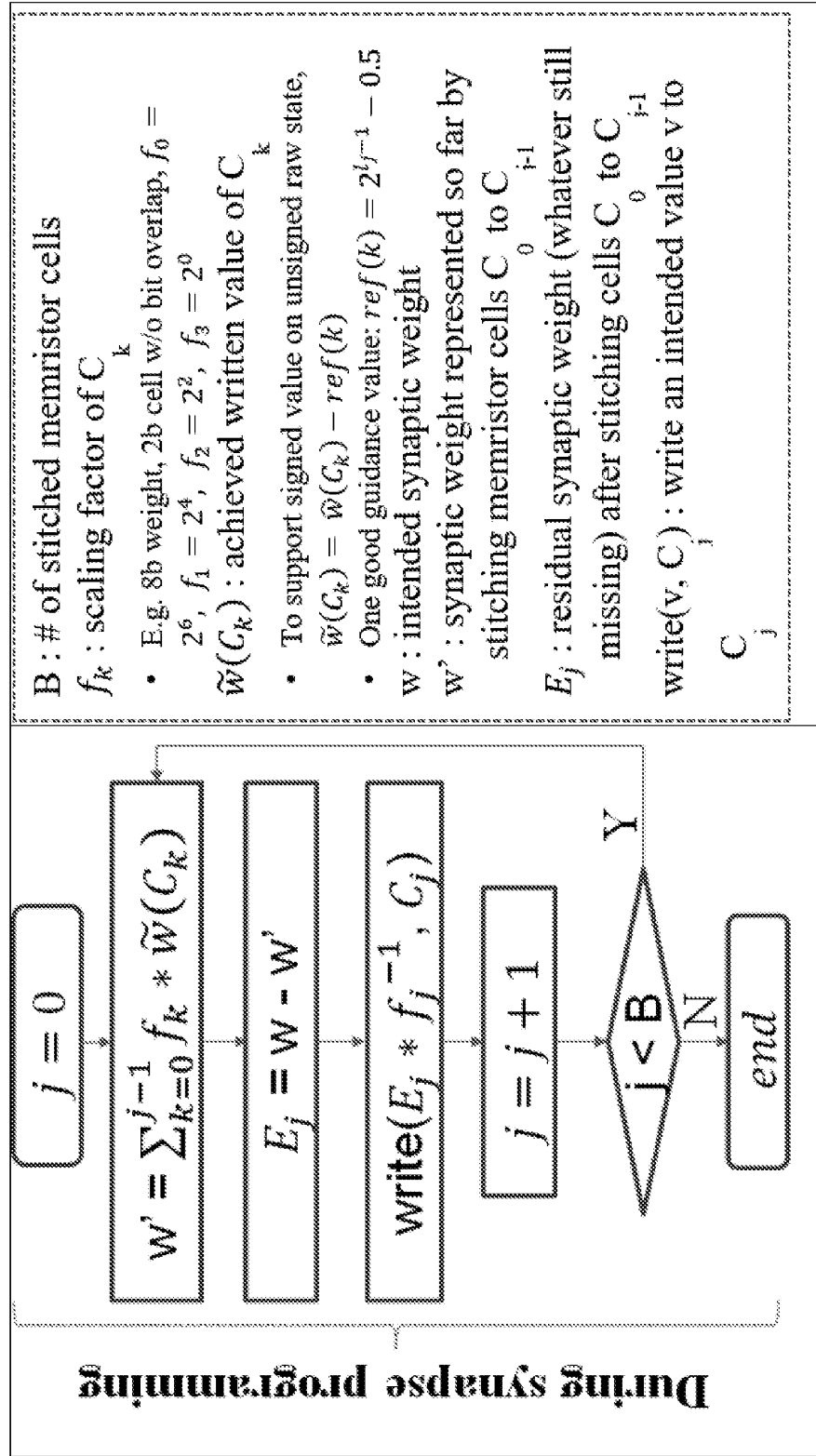


FIG. 11A

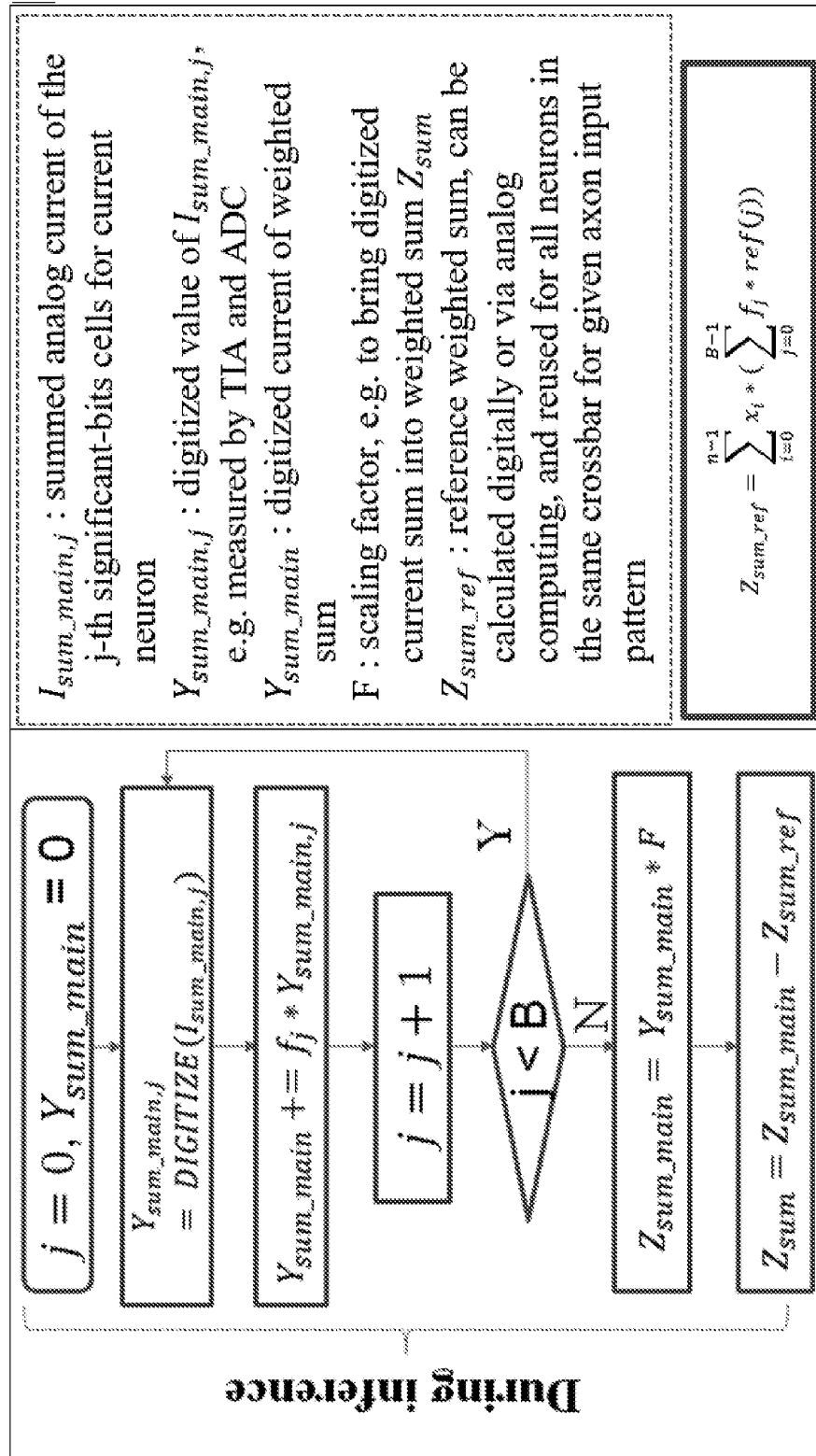


FIG. 11B

1210

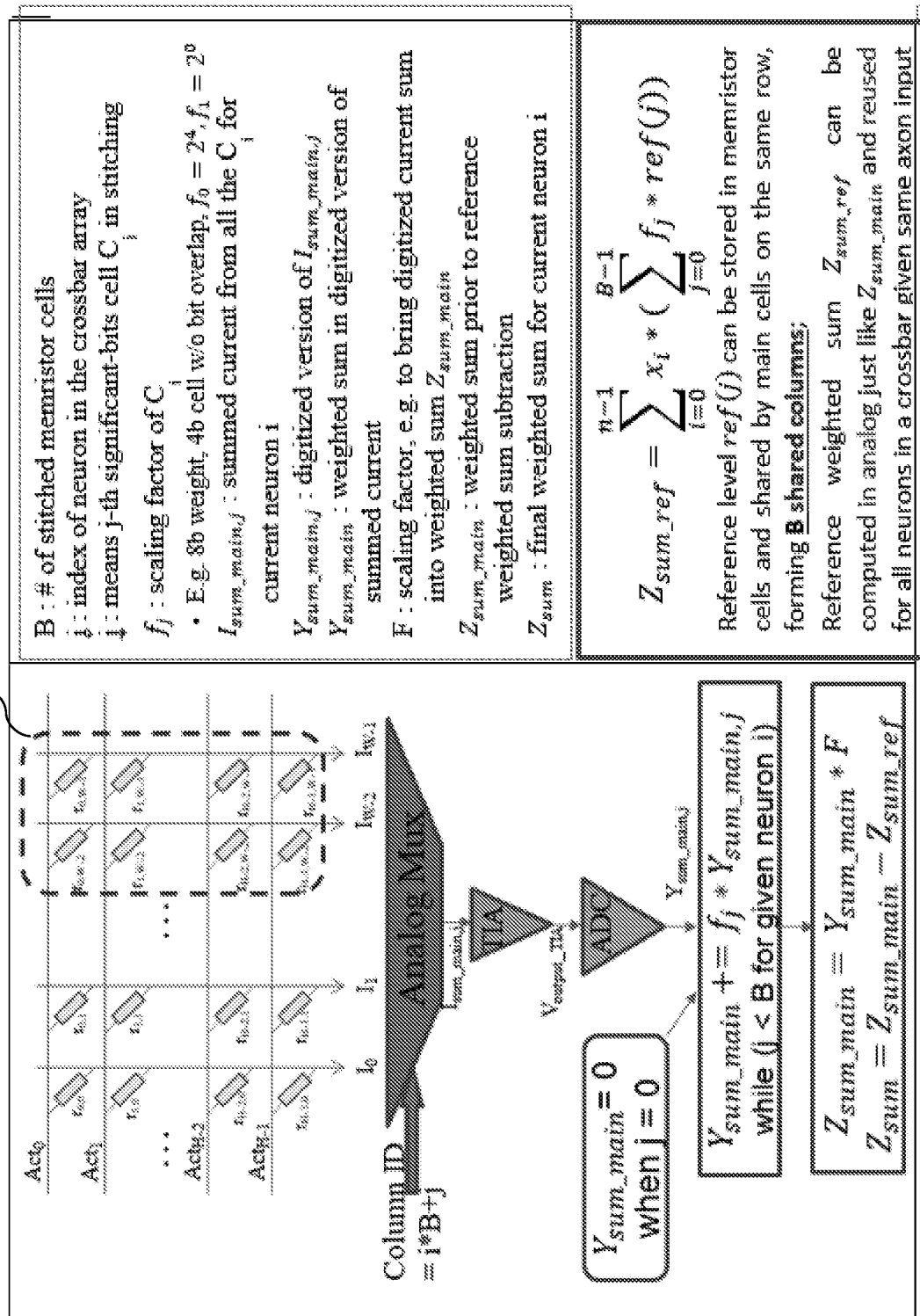


FIG. 12

Table 3A

AlexNet		Classification Error, on ILSVRC-2012 (ImageNet 1000-class problem)	
		Top-1	Top-5
FP32		40.3%	
DoReFa 8,8,8 (8-bit weight, 8-bit activation, 8-bit training gradient)	baseline	42%	20%
	Naïve 2 4-bit concatenation	51.97%	27.97%
	1 4-bit	52%	27.86%
	Stitched 2 4-bit, 1-bit overlap	42.19%	20.05%

FIG. 15A**Table 3B**

AlexNet		Classification Error, on ILSVRC-2012 (ImageNet 1000-class problem)	
		Top-1	Top-5
FP32		40.3%	
DoReFa 6,6,32 (6-bit weight, 6-bit activation, 32-bit training gradient)	baseline	42.31%	19.97%
	Naïve 2 3-bit concatenation	65.97%	42%
	1 3-bit	64.76%	40.54%
	Stitched 2 3-bit, no bit overlap	42.63%	20.03%

FIG. 15B

Table 3C

AlexNet		Classification Error, on ILSVRC-2012 (ImageNet 1000-class problem)	
		Top-1	Top-5
FP32		40.3%	
DoReFa 5,5,32 (5-bit weight, 5-bit activation, 32-bit training gradient)	baseline	42.45%	20.21%
	Naïve 2 2-bit concatenation	94.22%	84.49%
	1 2-bit	93%	82.45%
	Stitched 2 2-bit, no bit overlap	46.25%	23.25%

FIG. 15C**Table 3D**

AlexNet		Classification Error, on ILSVRC-2012 (ImageNet 1000-class problem)	
		Top-1	Top-5
FP32		40.3%	
DoReFa 4,4,32 (4-bit weight, 4-bit activation, 32-bit training gradient)	baseline	42.56%	20.01%
	Naïve 2 2-bit concatenation	87.54%	72%
	1 2-bit	85.82%	68.58%
	Stitched 2 2-bit, no bit overlap	45.11%	22.07%

FIG. 15D