Convolutional Neural Networks

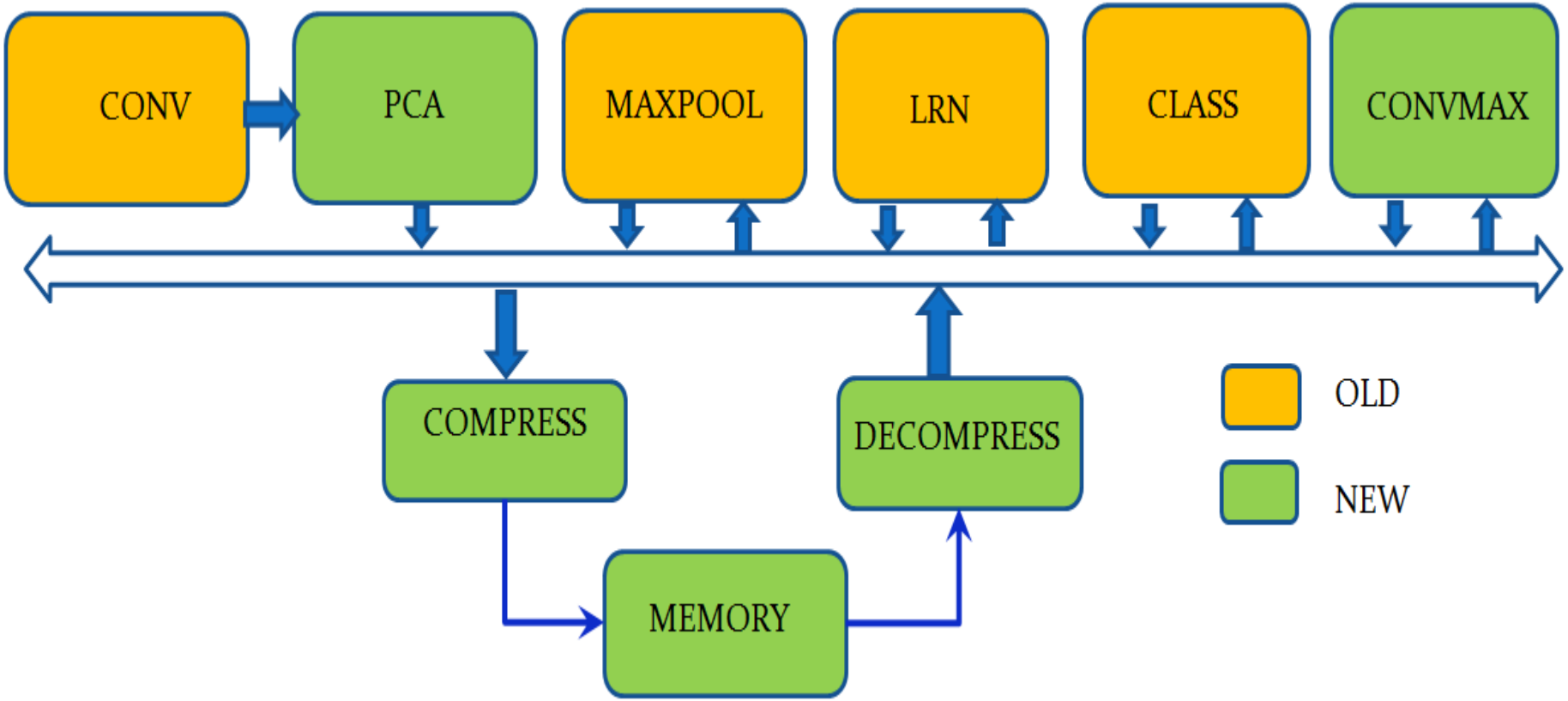
Arjun Gandhi, Mahmoud KhalafAlla, Venkatesh R.M.

April 23, 2015

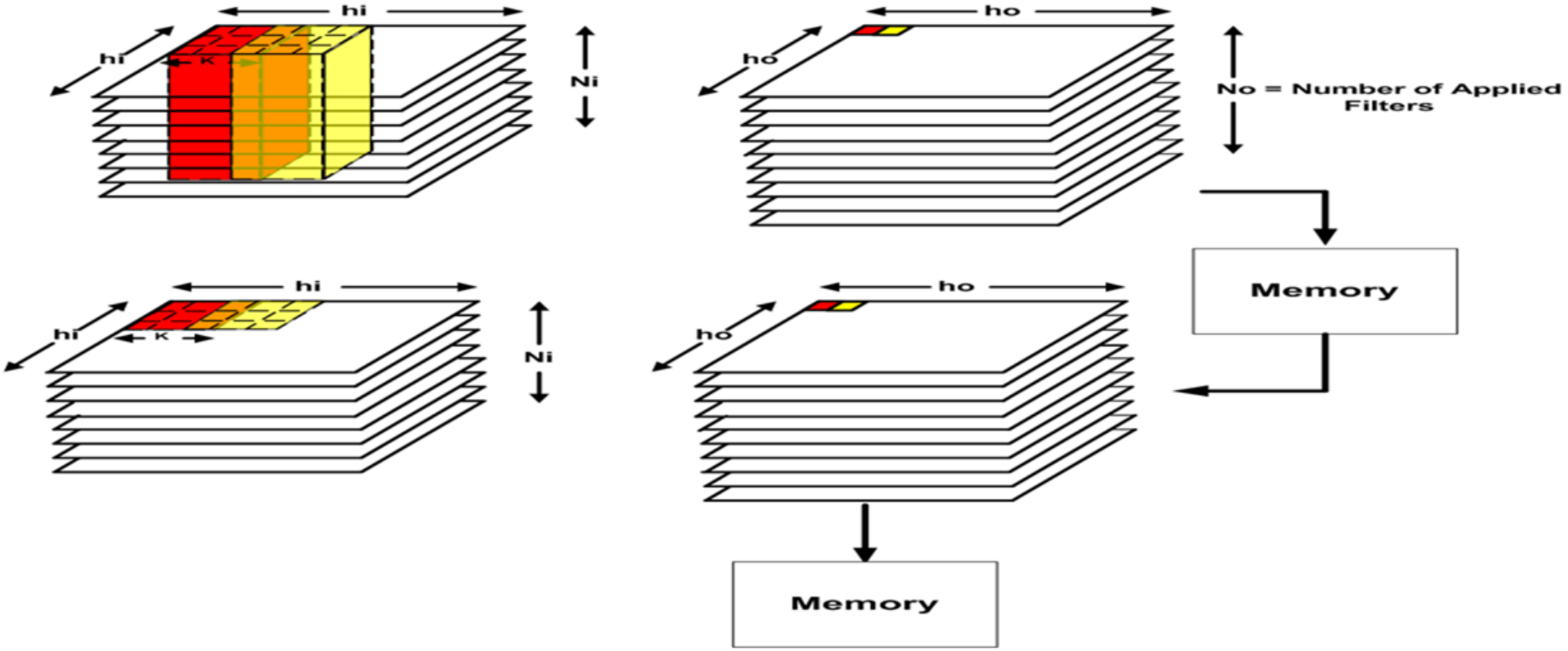
Agenda

- ❏ Computational Modules
- ❏ Increasing system parallelism
- ❏ Memory Architecture
- ❏ Discussion

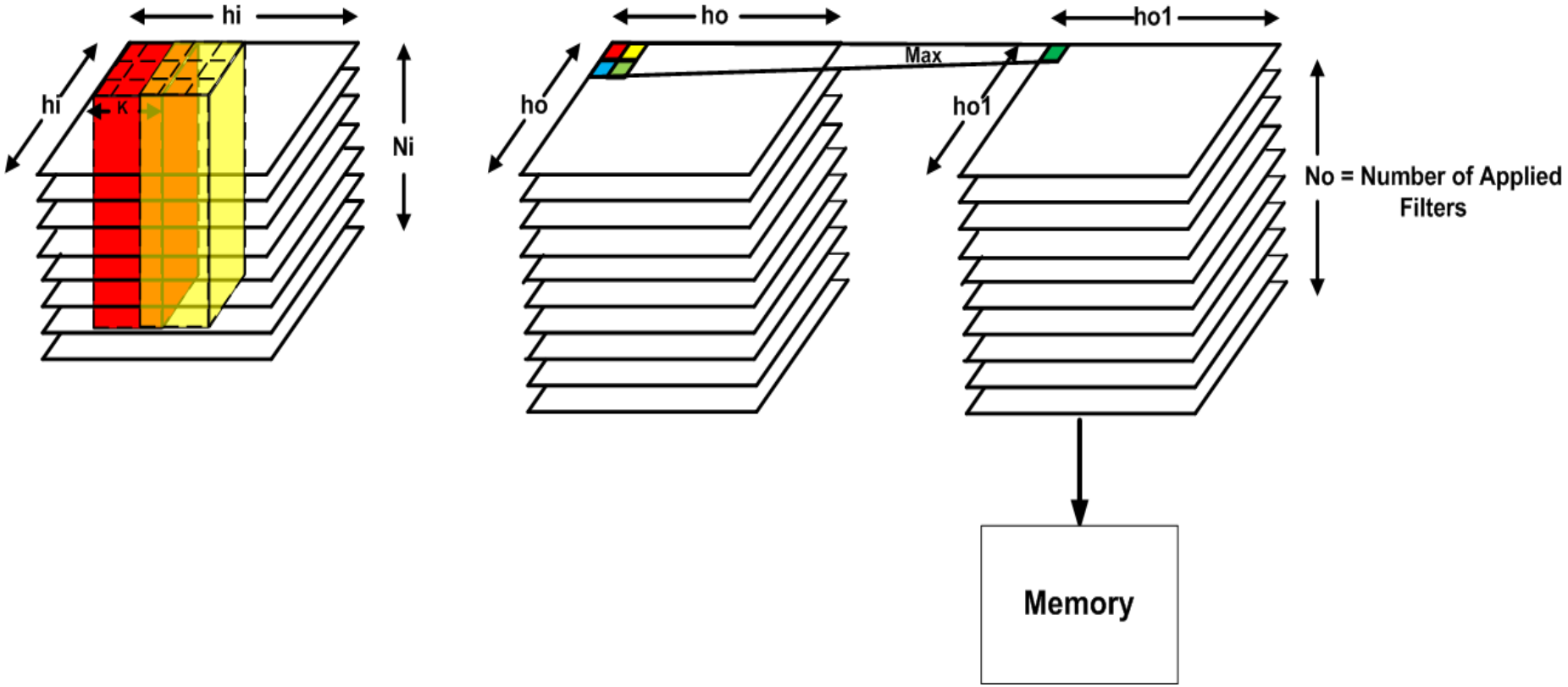
System-Level Diagram



ConvMax Layer



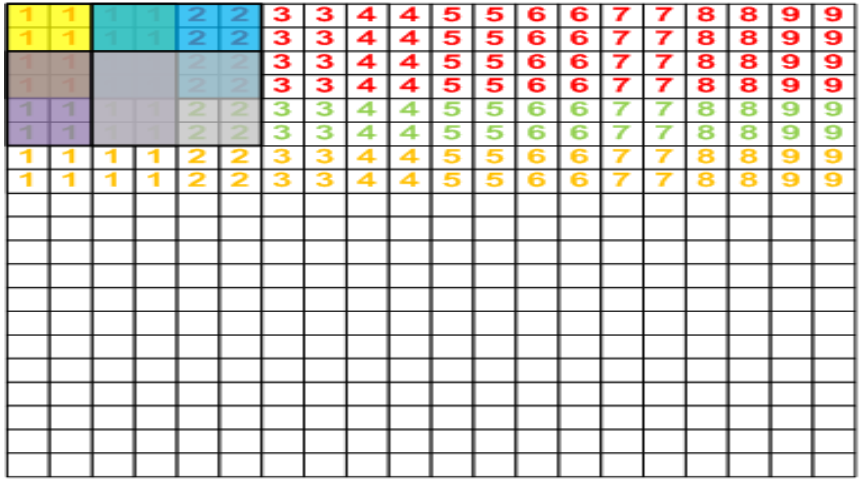
ConvMax Layer (Cont.)



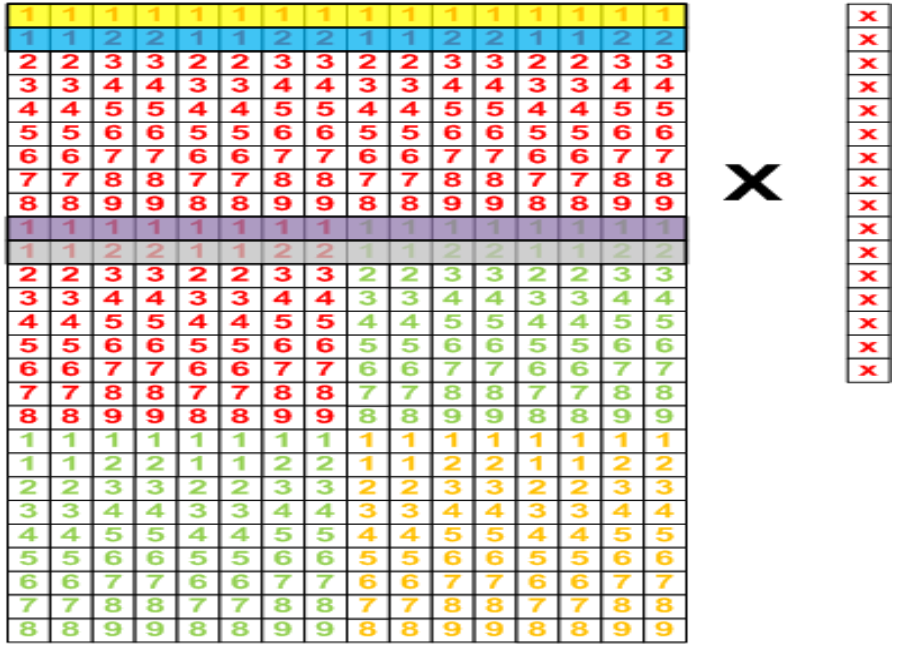
ConvMax Layer (Cont.)

Directives	Estimated Clock	Latency Min	Latency Max	Interval Min	Interval Max	BRAM_18K	DSP48E	FF	LUT
No Directive	8.09	227	227	228	228	0	5	903	1944
Outer Loop Pipeline + Data flow	8.09	53	53	41	41	0	5	1595	1446
Outer Loop Pipeline	8.63	65	65	66	66	0	5	1134	1515
Outer Loop Pipeline + Data flow + Intermediate Array partition	8.63	51	51	40	40	0	5	1861	2776

Caffe GPU implementation

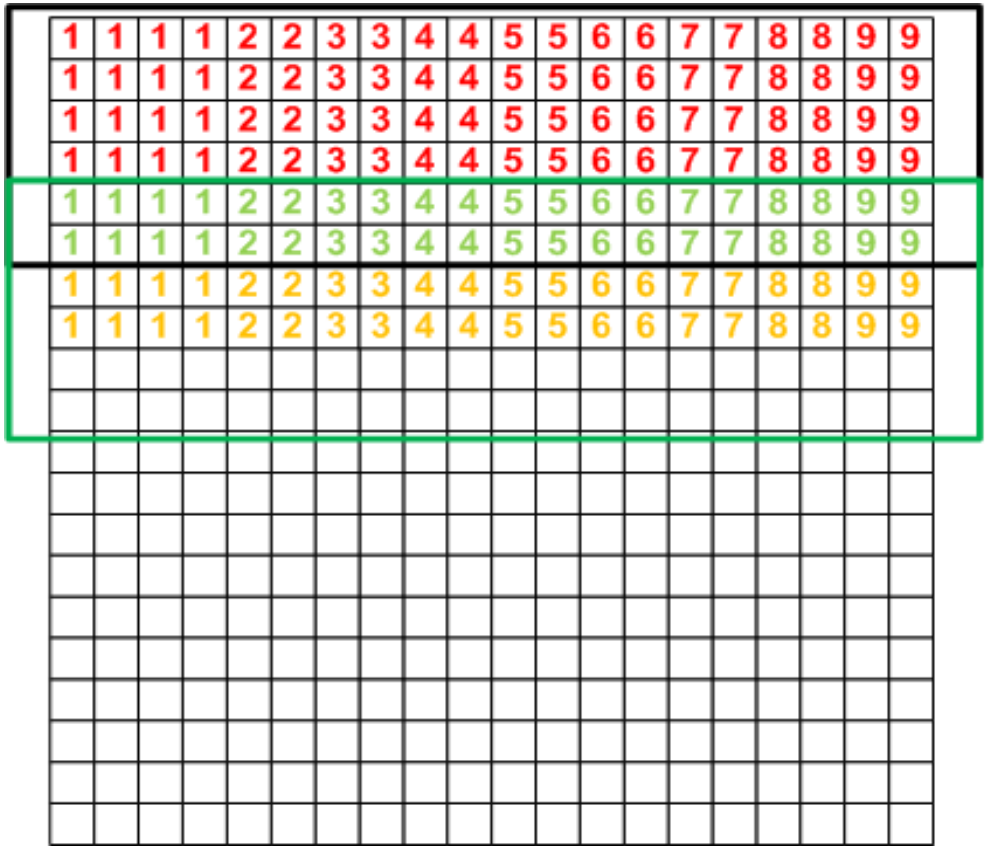


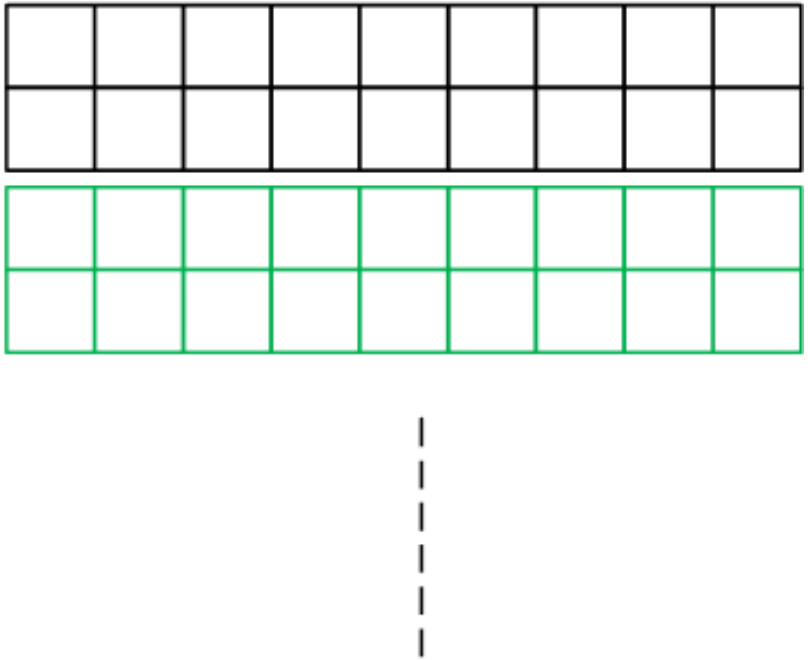
20 x 20 = 400



16 x 81 = 1296

Extending parallelism





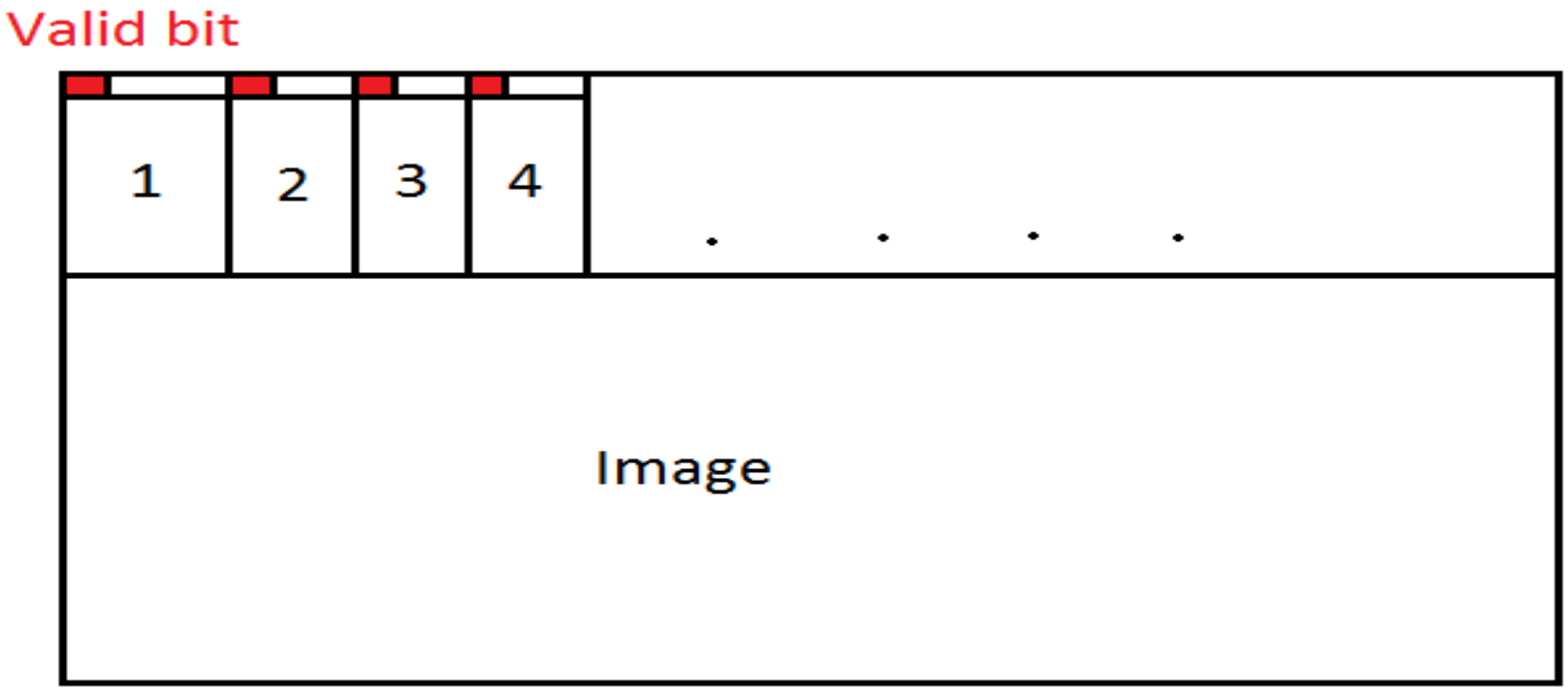
Memory Interface

- This is a programmable interface.
- Each module has an interface associated with it.
- The interface accesses the memory, retrieves the data and passes it on to the module.
- Similarly it can access the memory to write the output of the module.

Continue ...

- The controller programs the starting and ending address of the memory for each interface.
- The handshaking requires monitoring the valid bit set at the starting address by the preceding module.
- Upon accessing the data, the bit is cleared by the succeeding block.

Memory diagram



Continue ...

- This allows preceding block to write the new output to the same address if required.

Synthesis results

Performance Estimates

- **Timing (ns)**

Clock		solution1	solution2
default	Target	10.00	10.00
	Estimated	7.22	6.92

- **Latency (clock cycles)**

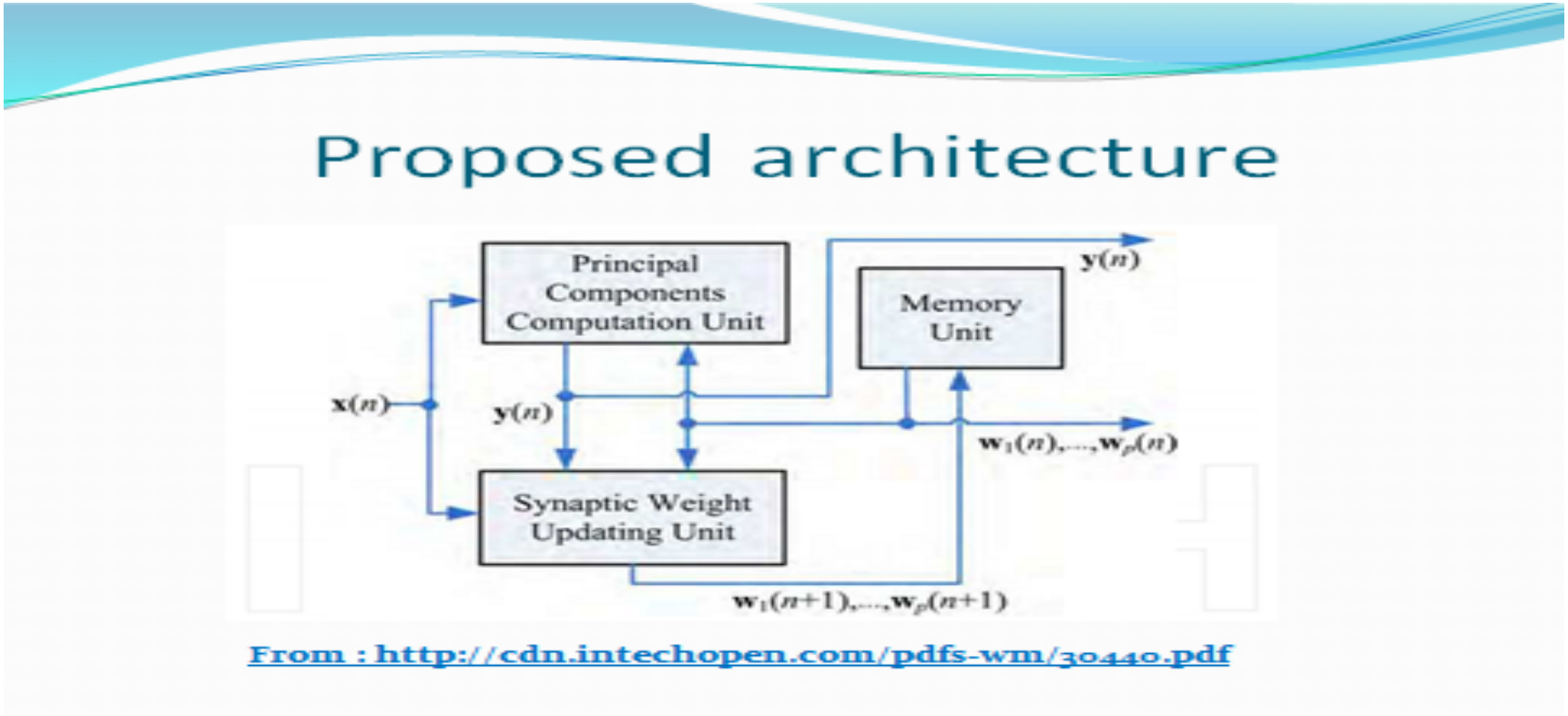
		solution1	solution2
Latency	min	5	5
	max	11	9
Interval	min	6	6
	max	12	10

Cont ...

Utilization Estimates

	solution1	solution2
BRAM_18K	0	0
DSP48E	0	0
FF	83	102
LUT	325	409

PCA proposed architecture



From : http://cdn.intechopen.com/pdfs-wm/30440.pdf

Time plan

- System Integration (Computational modules + Memory) (May -10)
- Development of test suite to measure system performance (May - 20)
- Replication of computational modules to extend parallelism (May - 31)