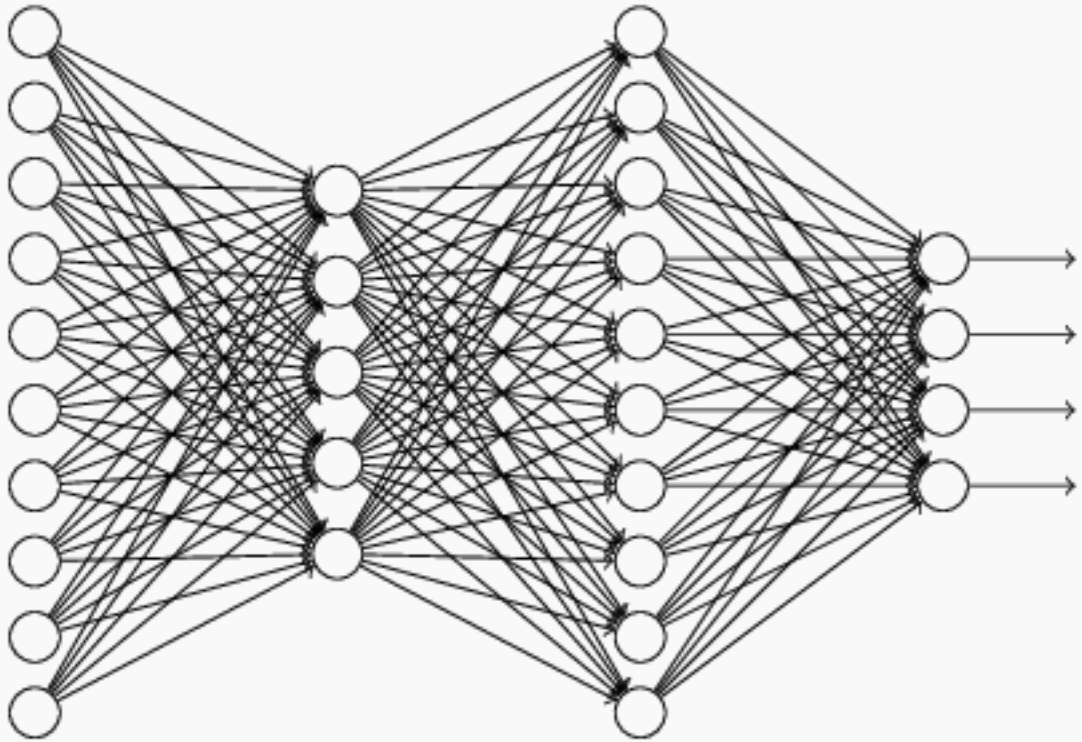
Implementing a Neural Network on FPGAs



Project Seminar

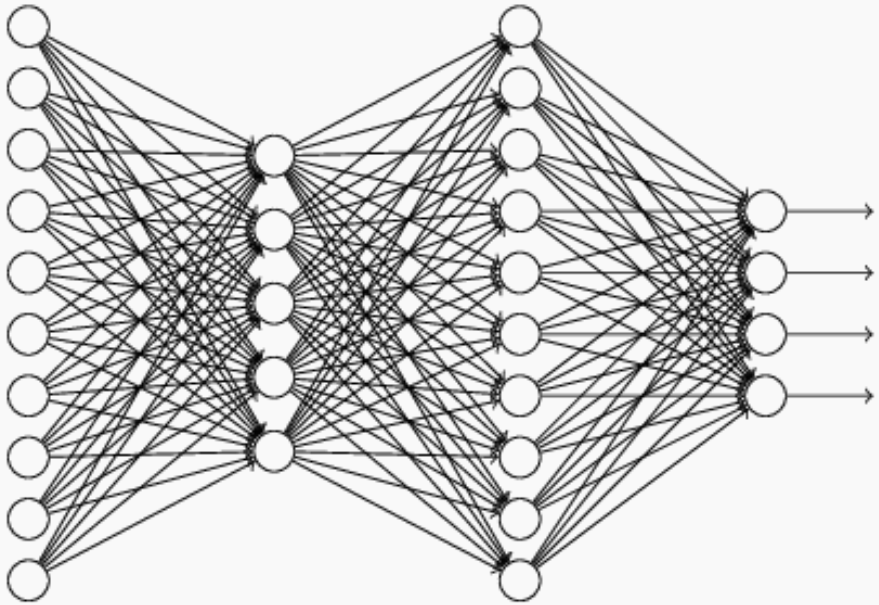
Author: Oliver Kugel

Supervisor: Dirk Koch

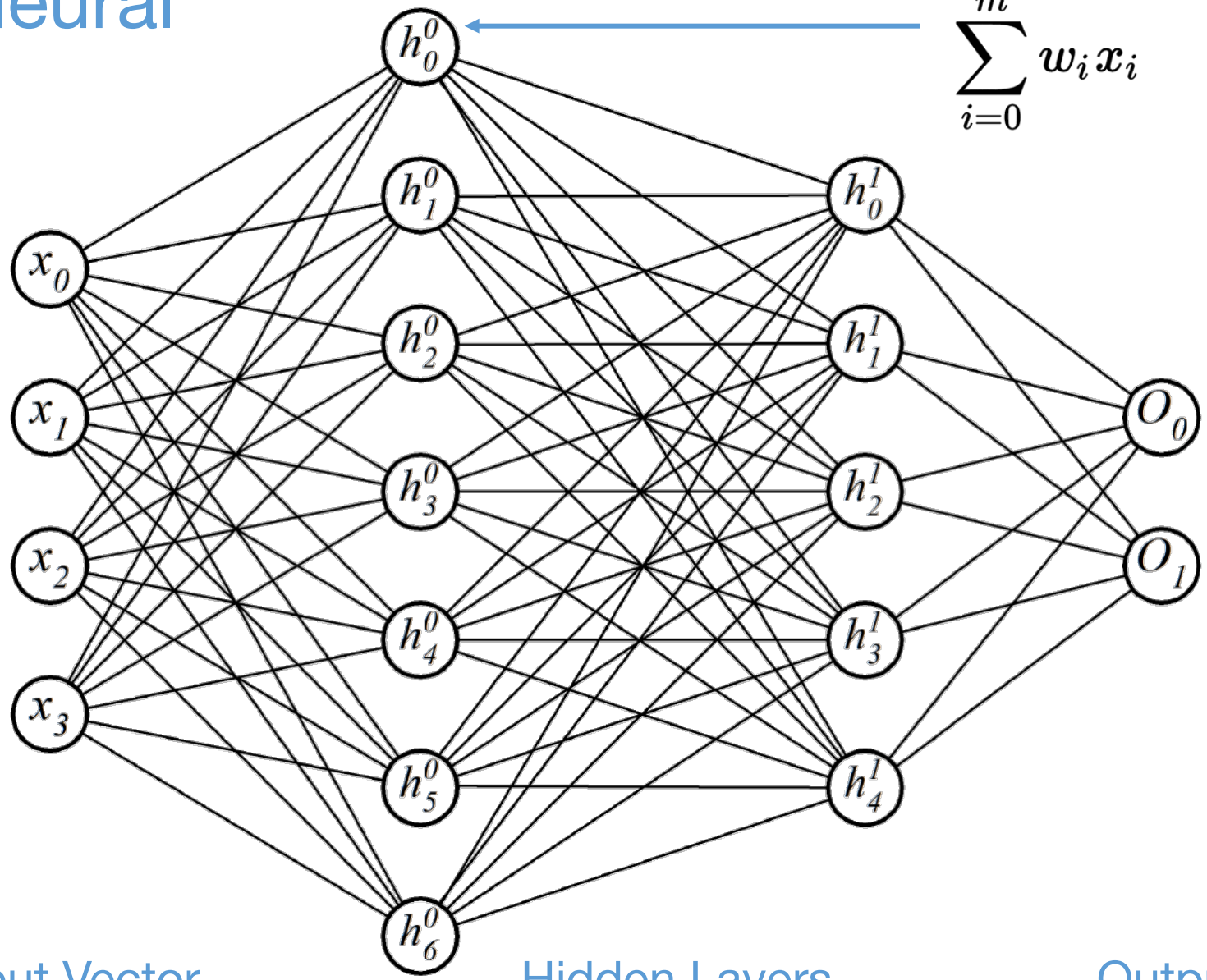
24th November 2016

Outline

1. Brief explanation of Artificial Neural Networks (ANN)
2. Serialized FPGA-implementation of an ANN
3. Computation cell architecture
4. Dataflow Scheduling
5. Verilog Testbench
6. Where do I get the data from, where do I get the trained ANN from?
7. What is left to do?

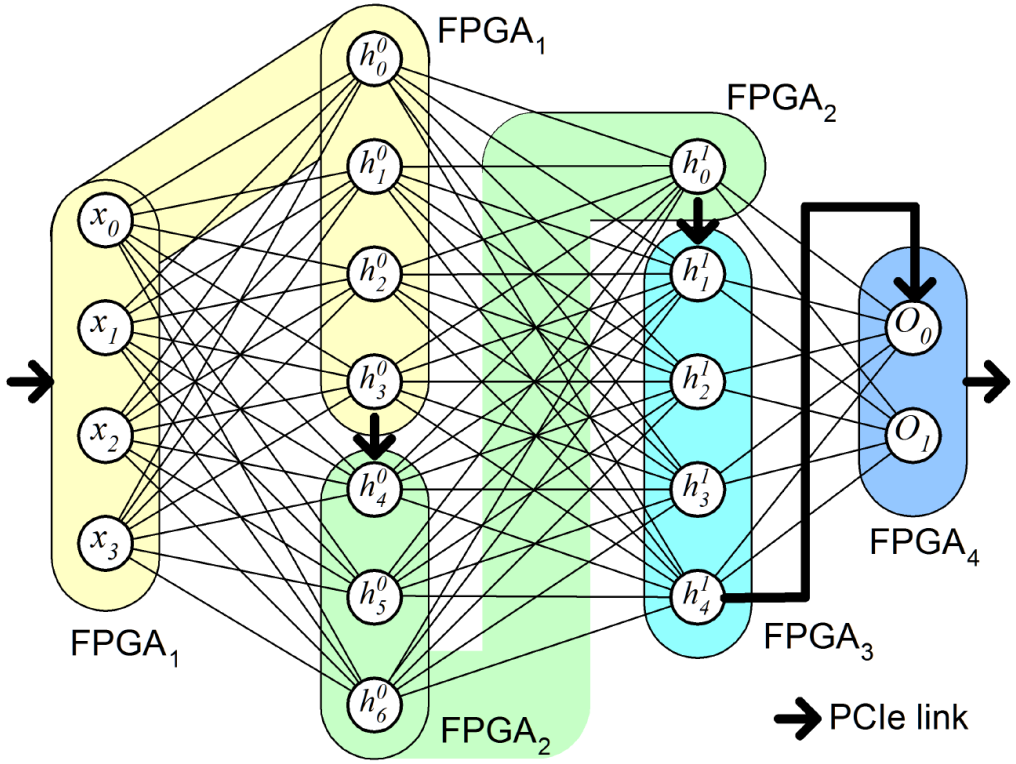


Artificial Neural Network



Advantages of Serialization

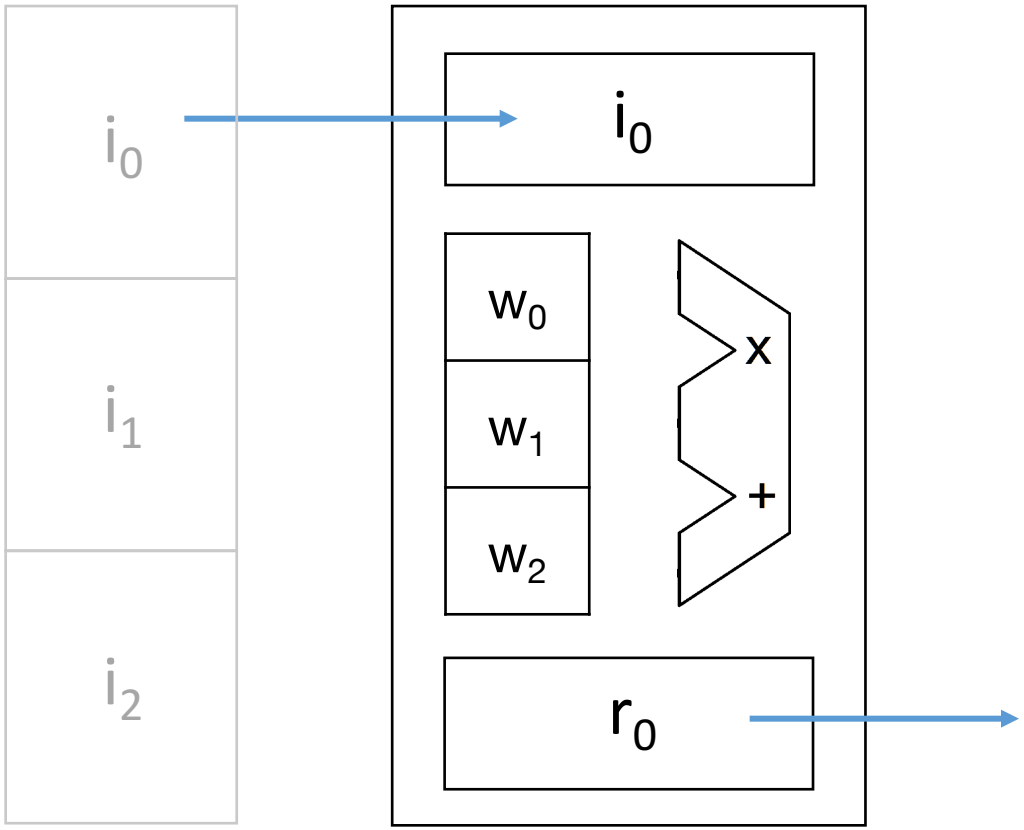
- Avoid problem of large fan-outs
- Computations and replications are of a fixed pre-known size
- Computation pipeline can be divided and mapped onto multiple FPGAs connected by PCIe links

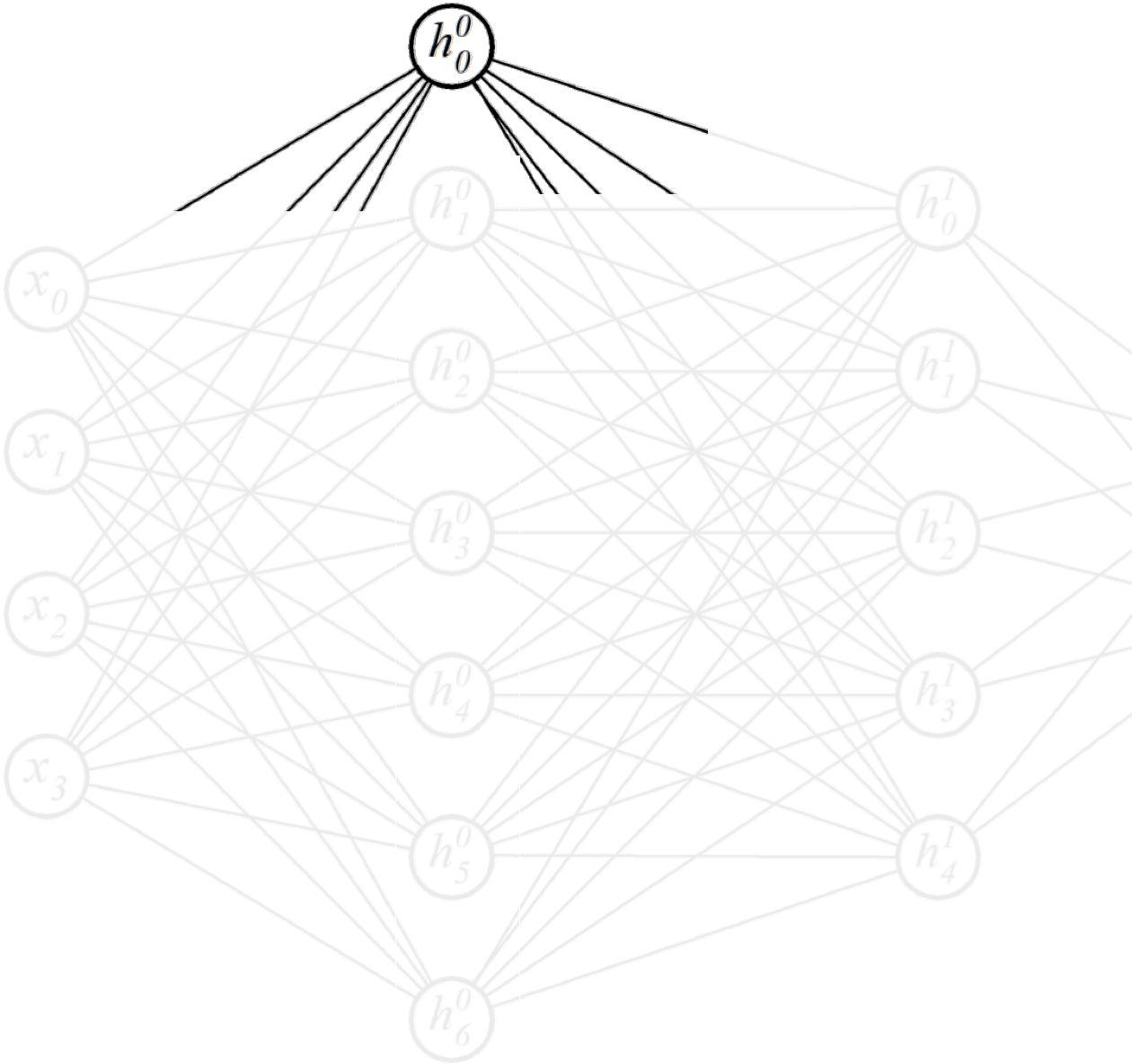


Challenges with Serialization

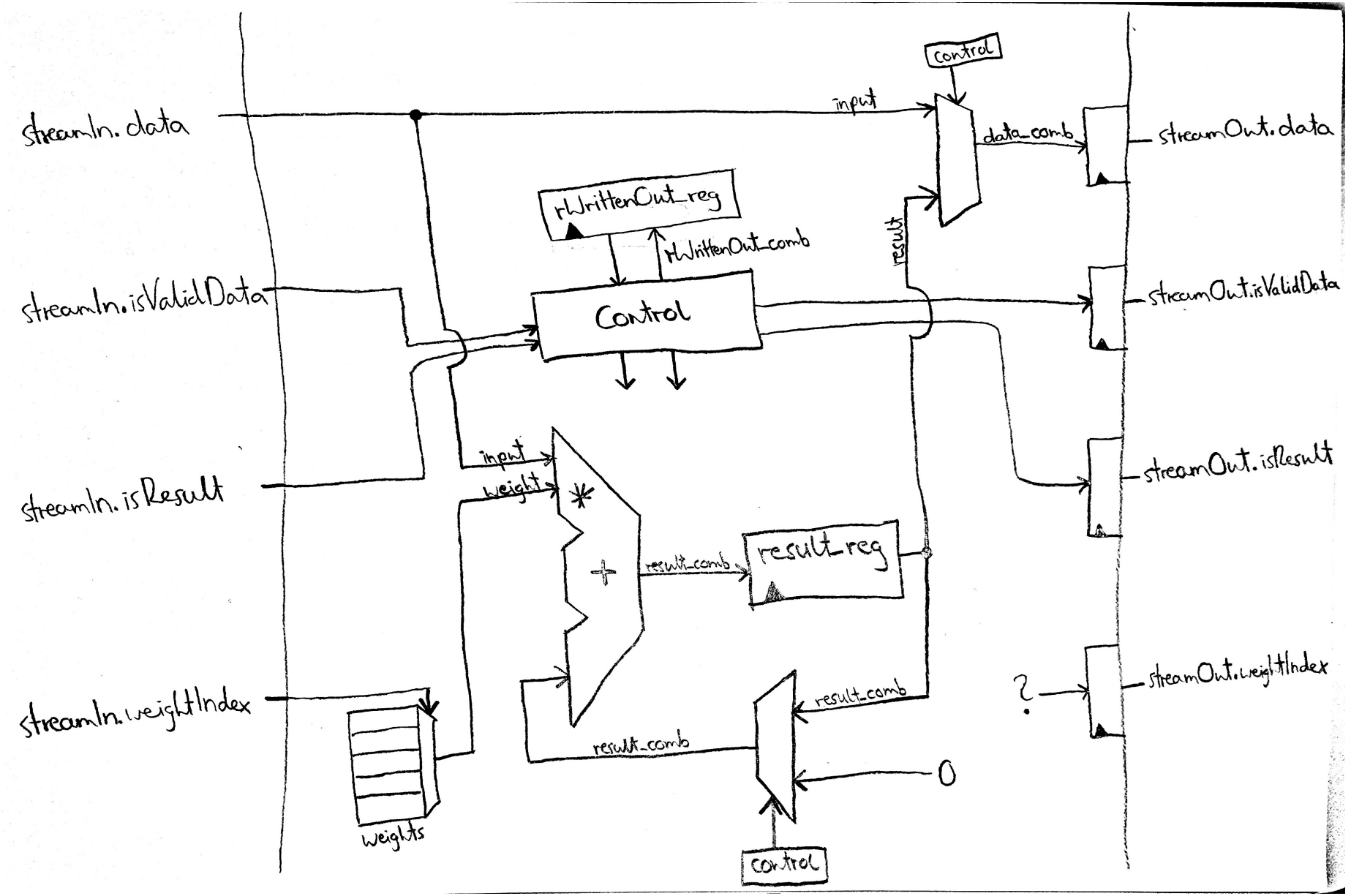
- Match producer rate to consumer rate to avoid large delays
- Get orchestration right
- Keep the pipe full at all times!

Computation Cell

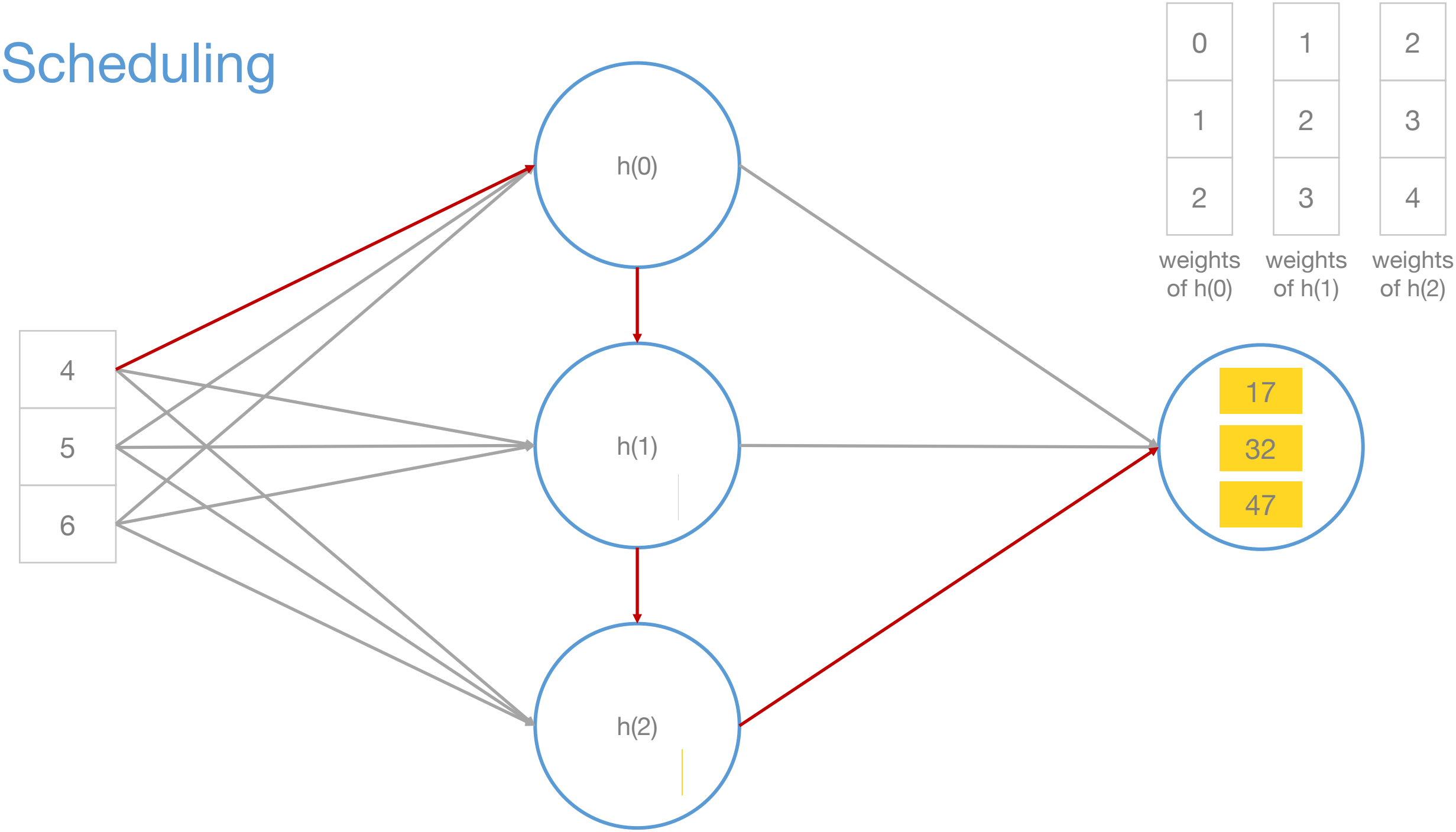




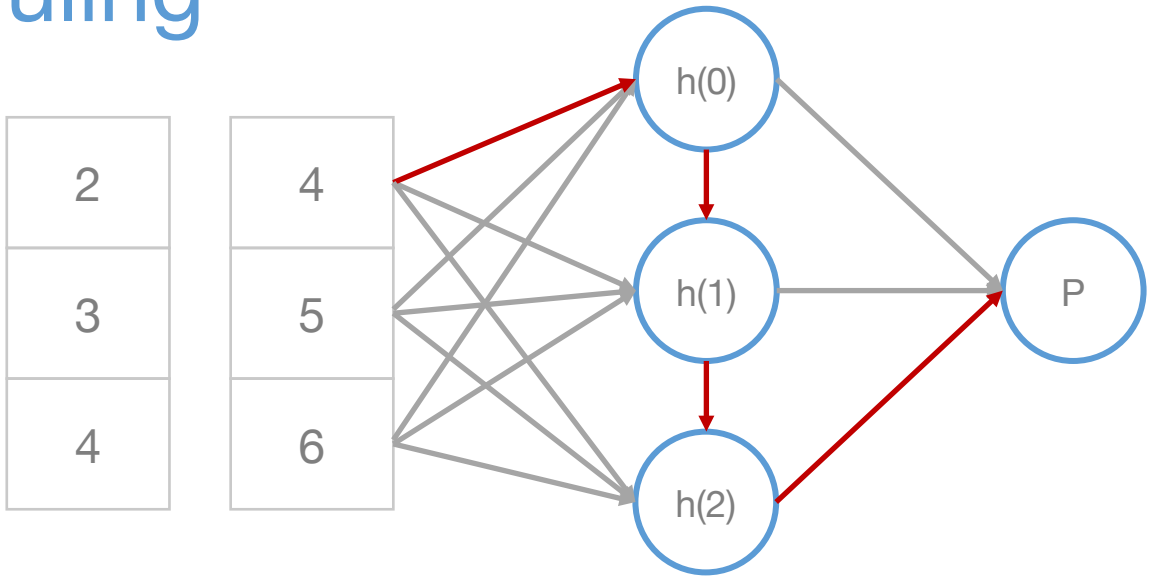
Computation Cell



Scheduling


0
1
2
1
2
3
2
3
4
weights of h(0)
weights of h(1)
weights of h(2)
h(0)
h(1)
h(2)
4
5
6
17
32
47

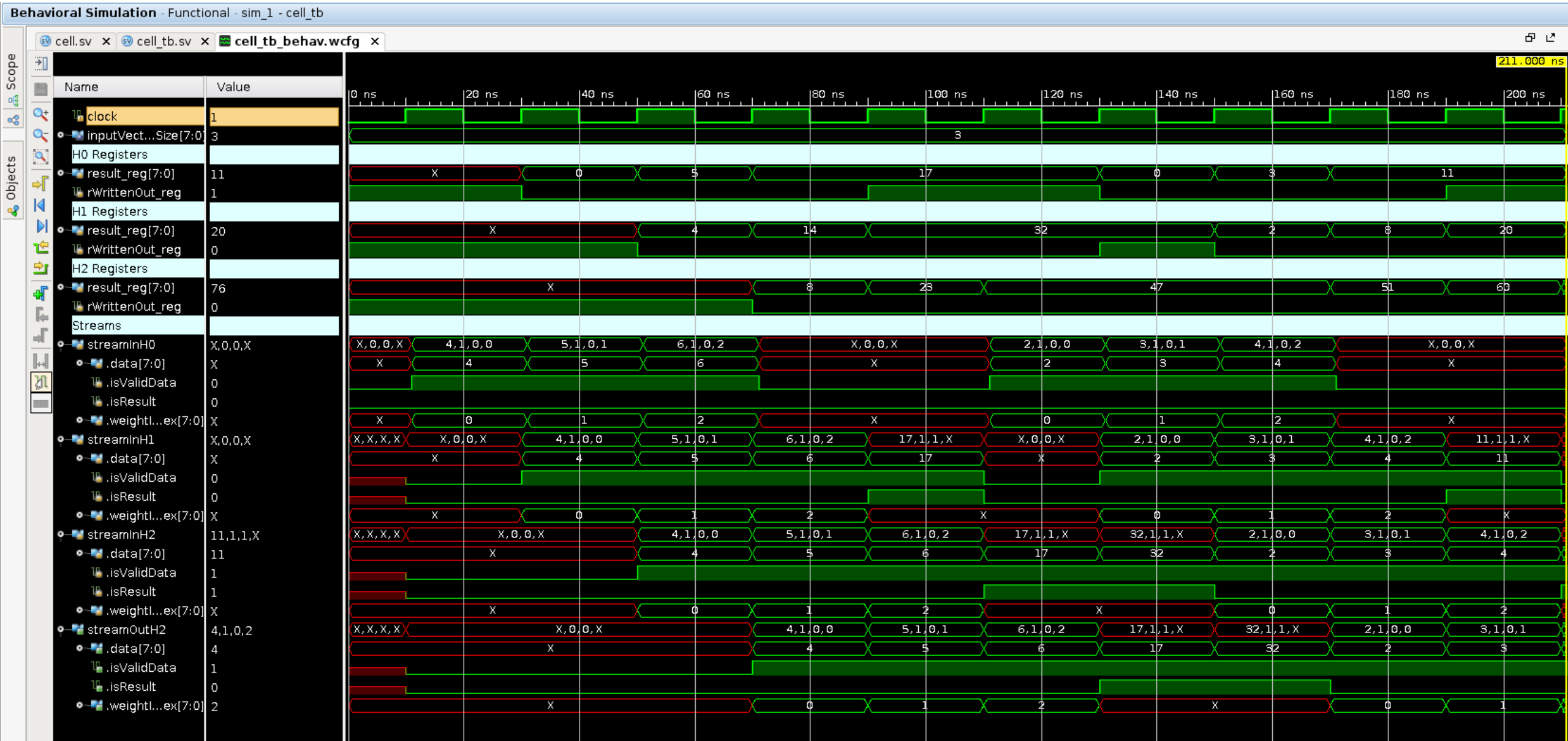
Scheduling



0	1	2
1	2	3
2	3	4
weights of h(0)	weights of h(1)	weights of h(2)

	T	t(0)	t(1)	t(2)	t(3)	t(4)	t(5)	t(6)	t(7)	t(8)	t(9)	t(10)	t(11)	t(12)	t(13)	t(14)
h(0)	input	4	5	6	*delay*		2	3	4	*delay*		…	…	…	*delay*	
	result		0	5	17			0	3	11			…	…	…	
h(1)	input		4	5	6	17		2	3	4	11		…	…	…	…
	result			4	14	32	32		2	8	20	20		…	…	…
h(2)	input			4	5	6	17	32	2	3	4	11	20	…	…	…
	result				8	23	47	47	47	4	13	29	29	29	…	…
P	input							17	32	47			11	20	29	

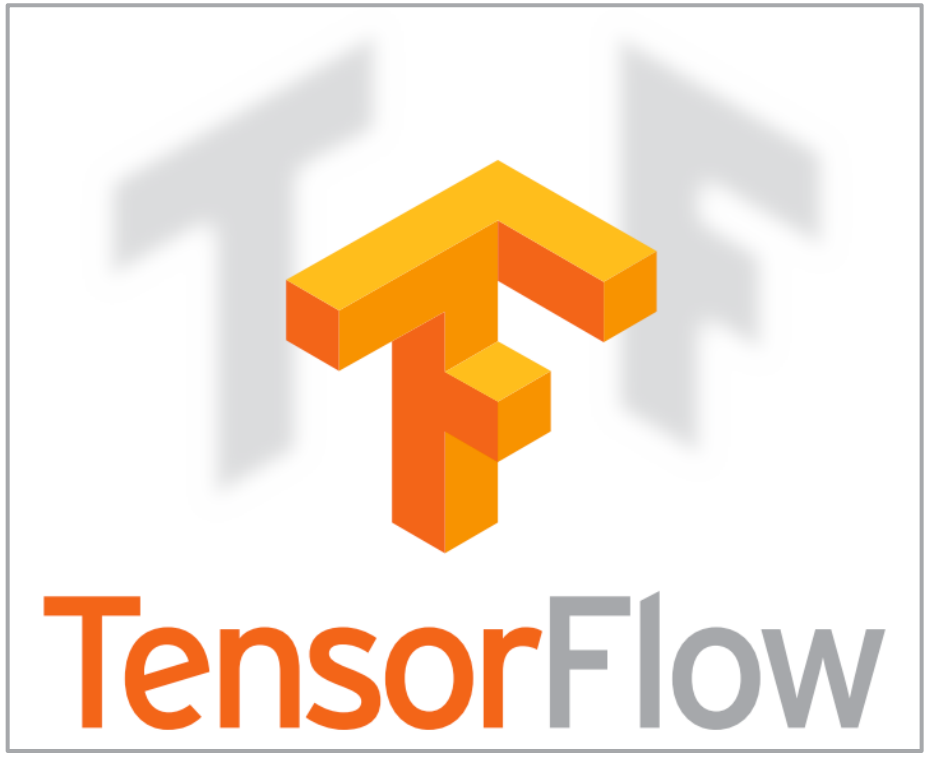
Verilog Testbench Waveform



Where does the Data come from?

- 14 million images
- 21,000 synsets
- Images are labelled
- Used heavily in academia and industry to train and evaluate artificial neural networks

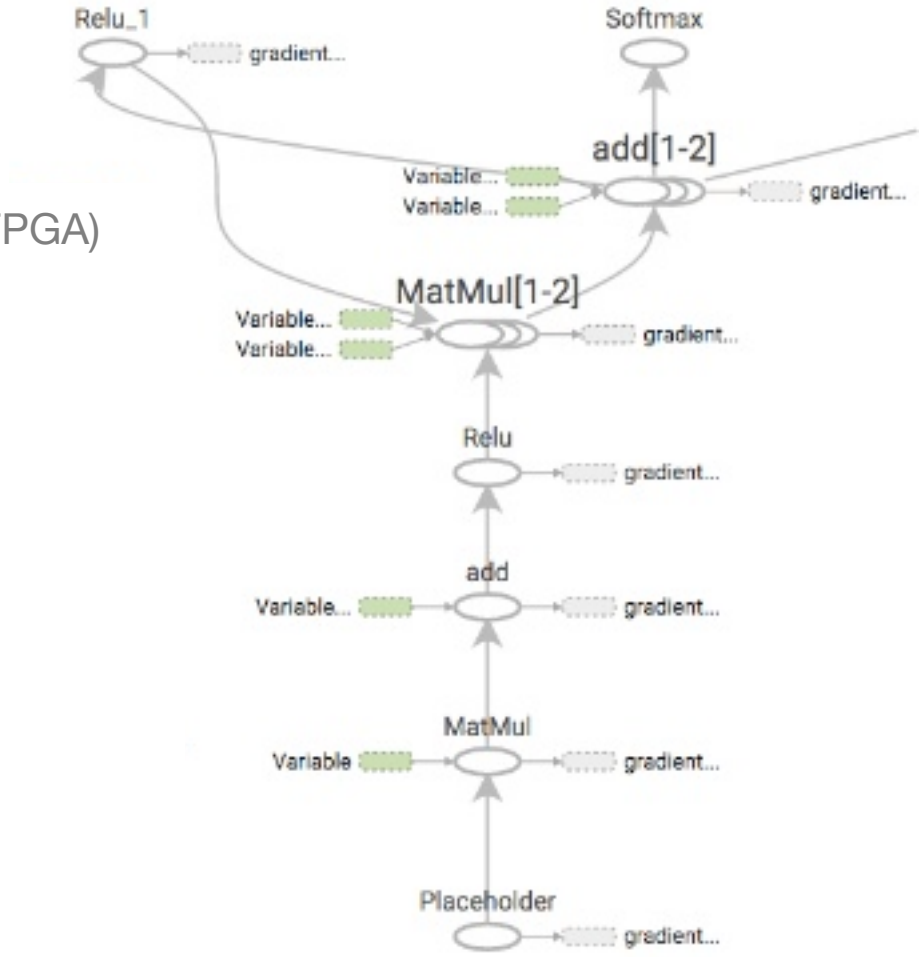
Where does the Neural Network come from?



Inception-v3

- Pre-trained neural network provided by Google
- 5 billion multiply-adds per inference
- Less than 25 million parameters (we can bring ~4 million on an FPGA)
- Best-in-class error rates

Network	Crops Evaluated	Top-5 Error	Top-1 Error
GoogLeNet [20]	10	-	9.15%
GoogLeNet [20]	144	-	7.89%
VGG [18]	-	24.4%	6.8%
BN-Inception [7]	144	22%	5.82%
PReLU [6]	10	24.27%	7.38%
PReLU [6]	-	21.59%	5.71%
Inception-v3	12	19.47%	4.48%
Inception-v3	144	**18.77%**	**4.2%**



What is done

- My knowledge on FPGAs, Verilog, Tensorflow and Neural Networks has grown
- ImageNet data is locally available
- Tensorflow is built and works fine
- Compressed Inception-v3 network
- Designed and implemented computation cell
- Working testbench for computation cells
- Accumulated considerable number of sources and papers to reference in my thesis

What is left to do

- Functions implemented by Inception-v3 must be translated to Verilog (e.g. Relu)
- Software pre-processing to format images
- Problem of variable number of weights must be solved
- Speed and cost comparison to GPP
- Thesis must be written
- Optional: Multi-FPGA solution

Thank you.

Any questions?

Project Seminar

Author: Oliver Kugel

Supervisor: Dirk Koch

24th November 2016