Field Programmable Gate Arrays for the Acceleration of Neural Network Computation

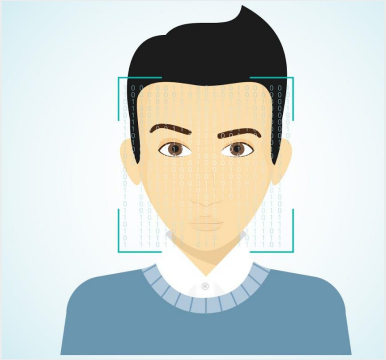
By: Joao Foltran, Michael Greer, and Justin Gutter

Today,
more than 40%
of adults use voice search engines at least once per day.

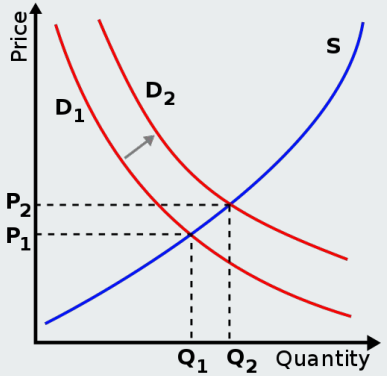
(Location World, 2018)

Artificial Neural Networks are the building blocks for current AI systems.

(Figure Tribune, 2017)



(Blogger, 2018)

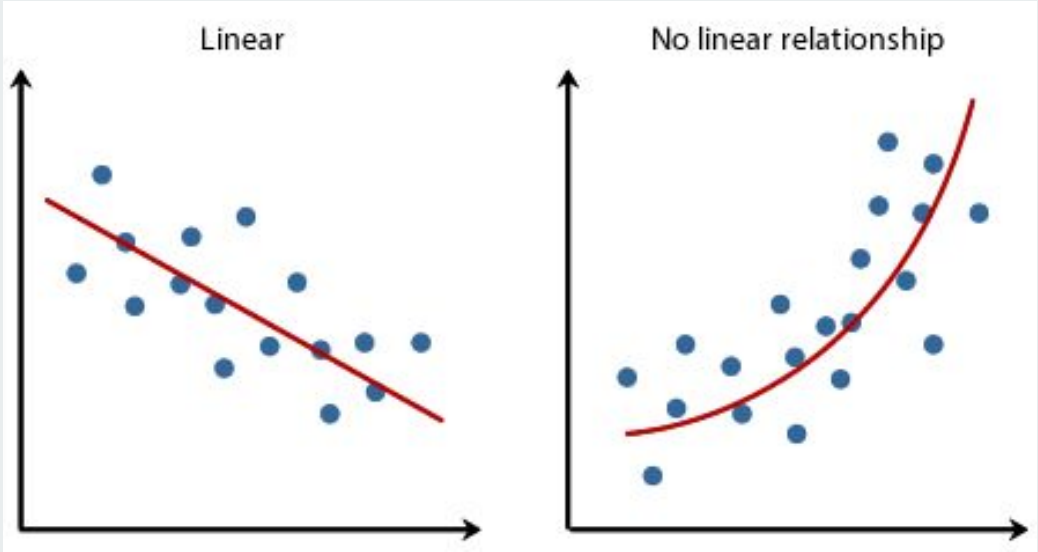


(Wikipedia, 2018)

Relations between inputs and outputs

Inputs	Outputs
2006 World Cup	Lost
2010 World Cup	Won
2014 World Cup	Lost
2018 World Cup	?

Training Data



(Laerd, 2018)

Example of Deep Neural Network

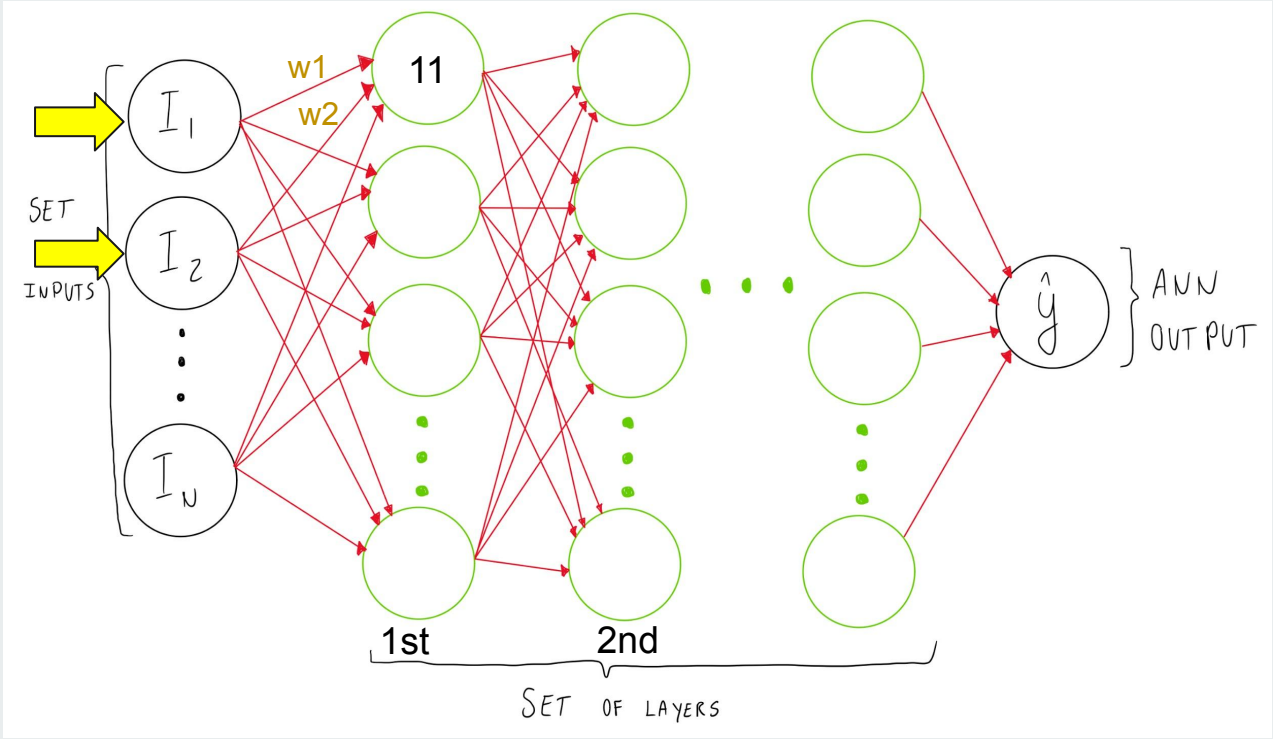


Figure by authors

Example of Activation Function

$$\sigma(z)_j = \frac{e^{z_j}}{\sum_{k=1}^{k} e^{z_k}}$$

(Yash, 2017)

Training of Artificial Neural Networks

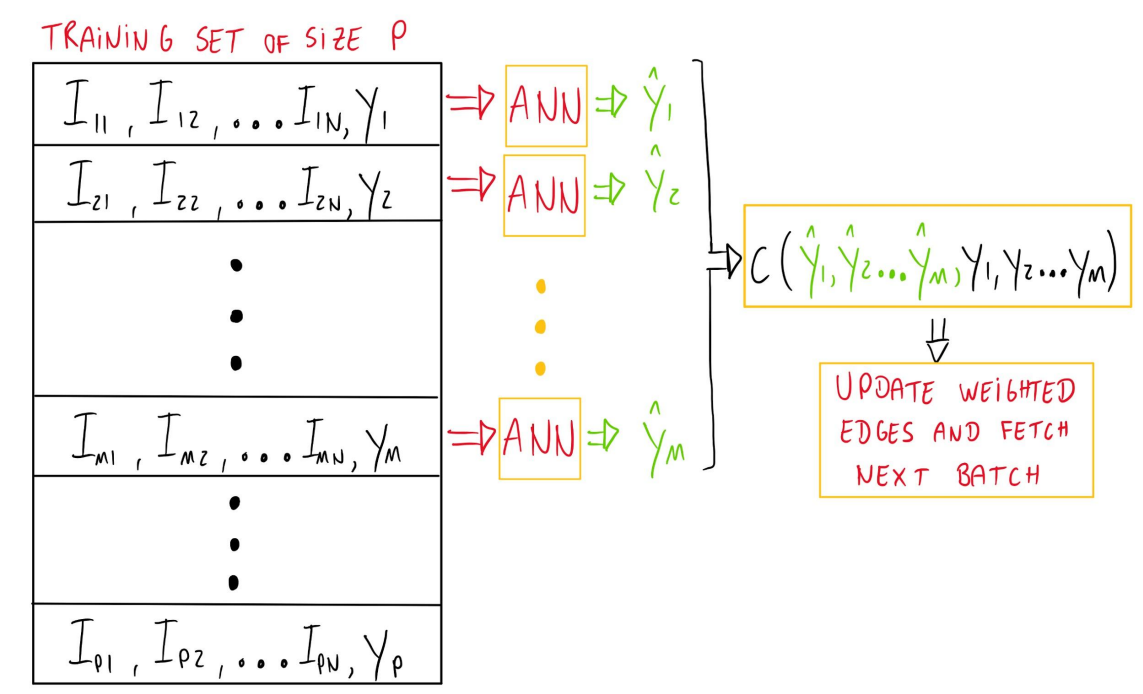
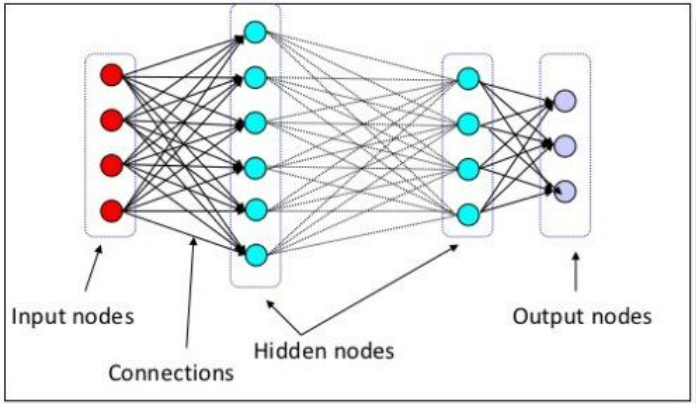


Figure by authors

Need for parallelism

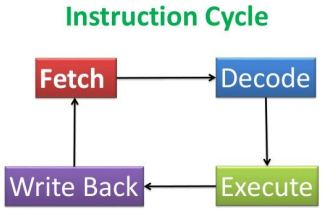


$$v_k = \sum_j w_{kj} i_{kj} + \theta_k$$

(Yash, 2017)

-Many calculations at a single point in time is the crux of the problem.

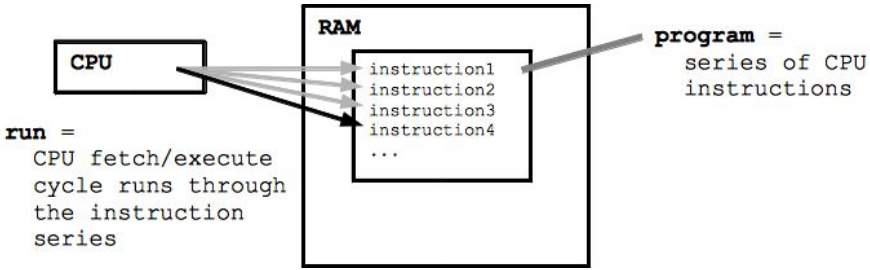
GPU LAYOUT

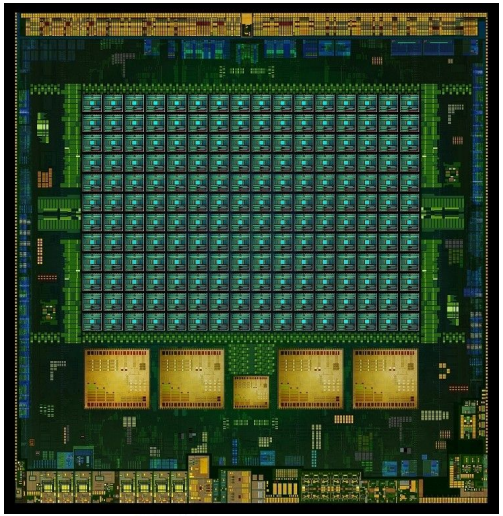


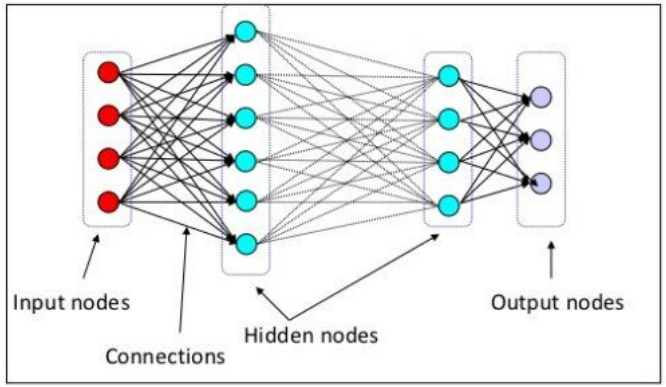
-Created for parallelism

-Each core runs software similar to CPU.

$$v_k = \sum_j w_{kj} i_{kj} + \theta_k$$

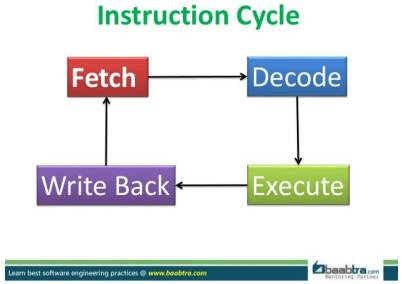






(Yash, 2017)

FPGA LAYOUT

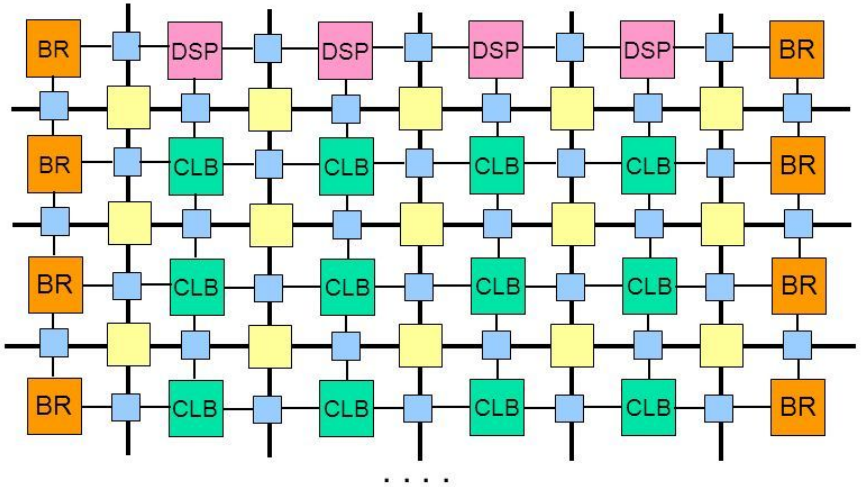


-Cluster of useful functional blocks

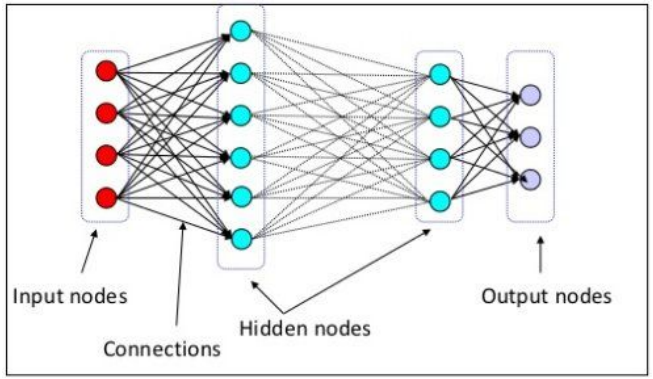
-BRAM

-DSP block

-CLB block



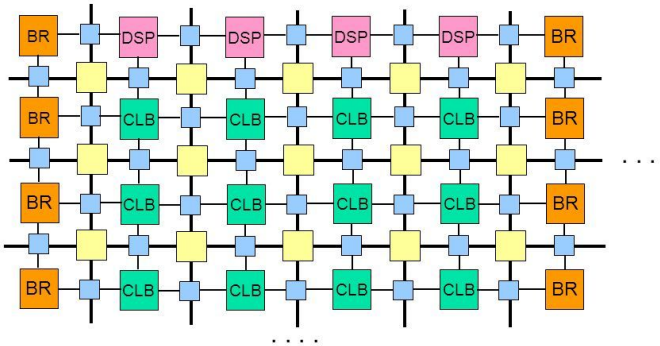
$$v_k = \sum_j w_{kj} i_{kj} + \theta_k$$

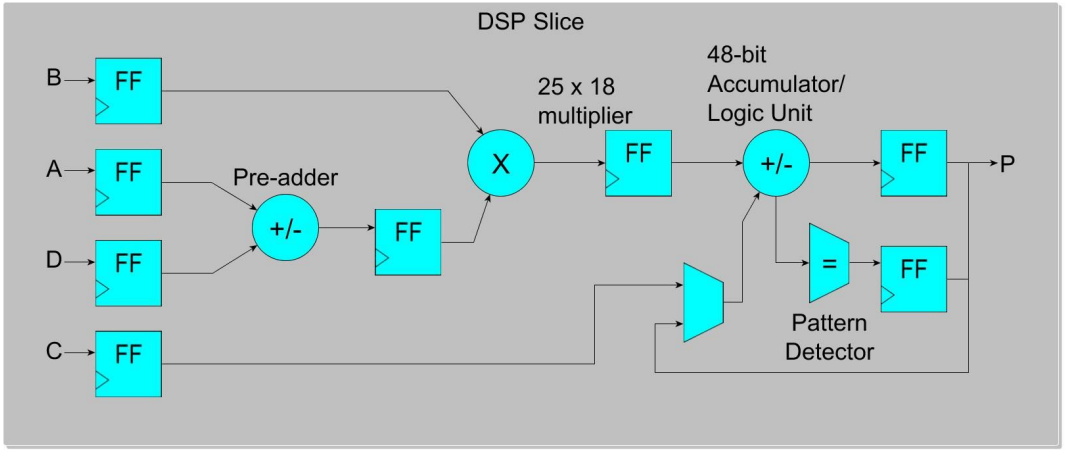


(Yash, 2017)

(Digital Signal Processing) DSP Block

-Performs multiplication, needed for NN mathematics.

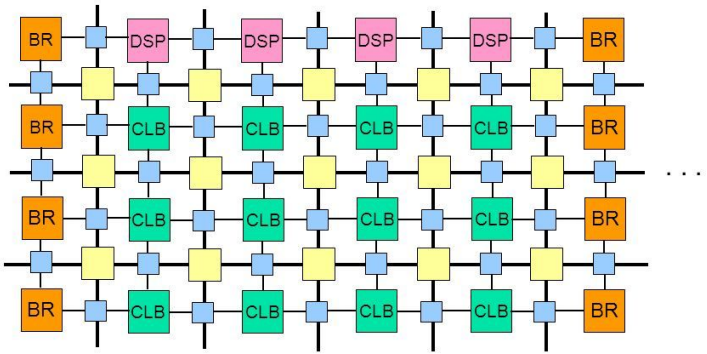




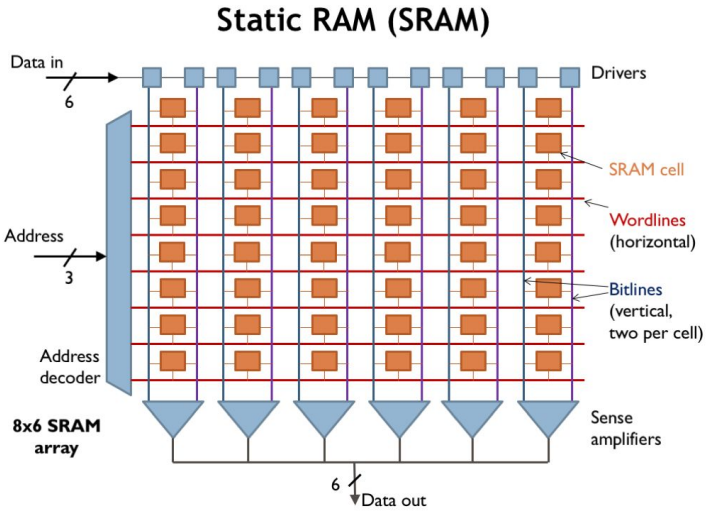
(Yash, 2017)

(Block RAM)BRAM

-Memory modules used for storage

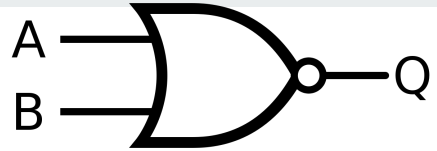


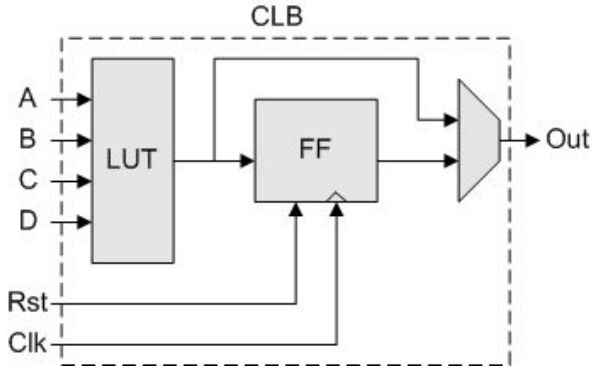
(Yash, 2017)

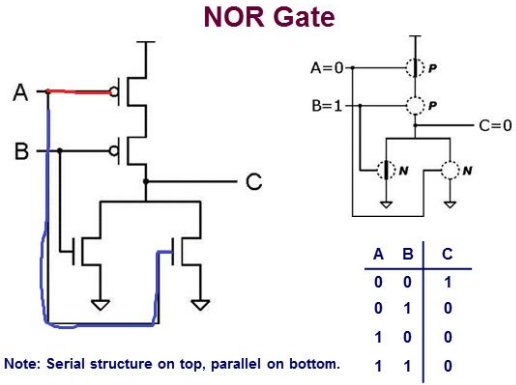


(Configurable Logic Block) CLB

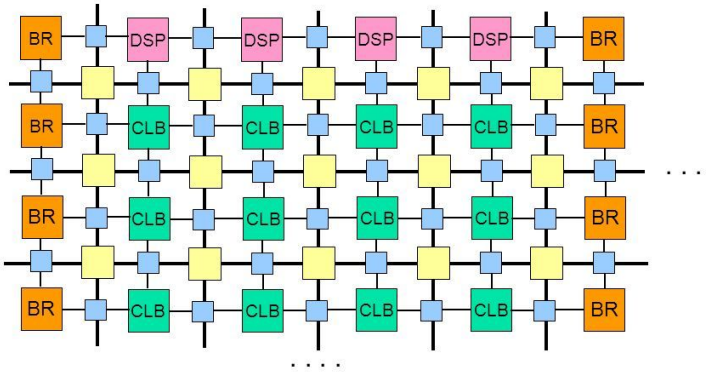
-Uses memory device that selects a memory location using the inputs, this outs a saved value.



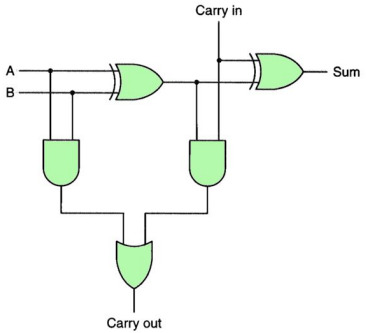




A	B	C
0	0	1
0	1	0
1	0	0
1	1	0



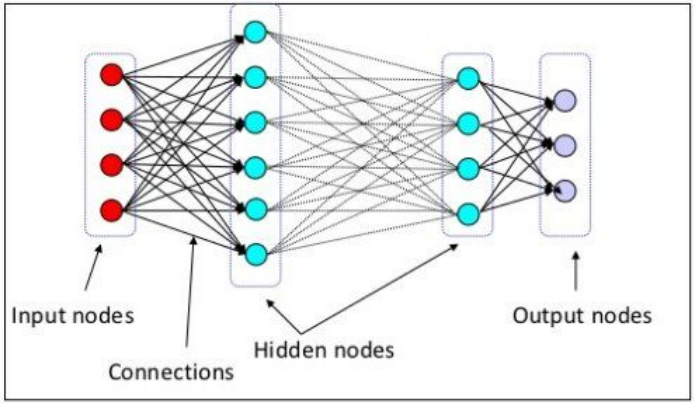
A	B	Carry in	Sum	Carry out
0	0	0	0	0
0	0	1	1	0
0	1	0	1	0
0	1	1	0	1
1	0	0	1	0
1	0	1	0	1
1	1	0	0	1
1	1	1	1	1



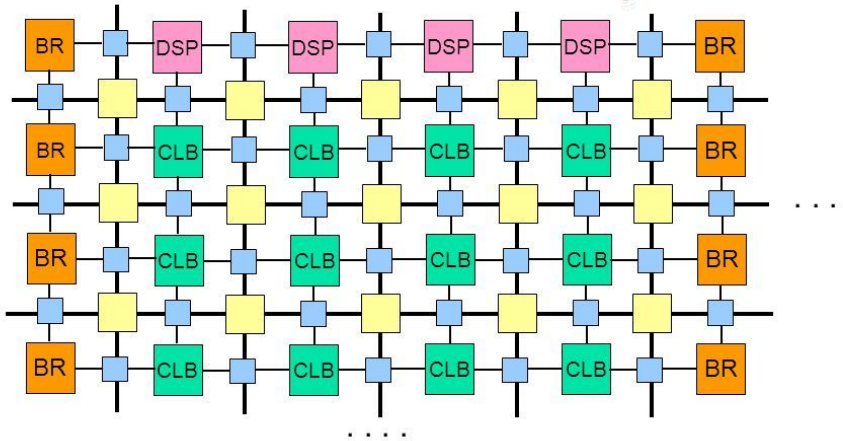
(Yash, 2017)

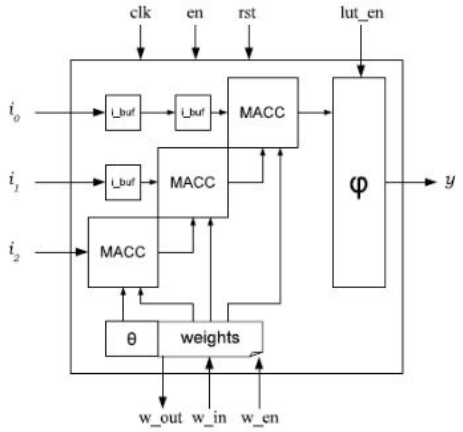
Implimenting NN on FPGA

Each node is a piece of configured hardware



$$v_k = \sum_j w_{kj} i_{kj} + \theta_k$$

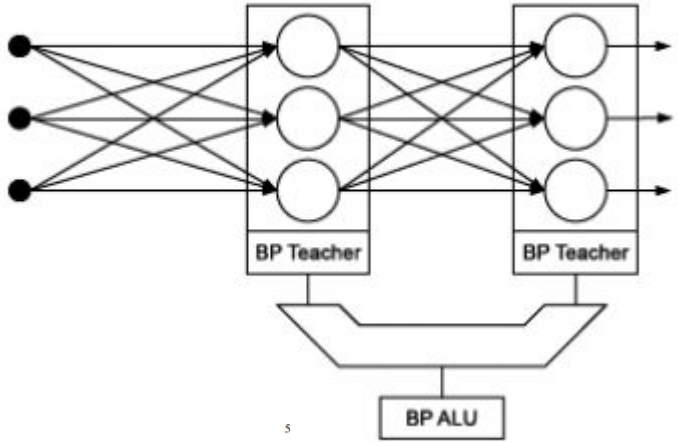




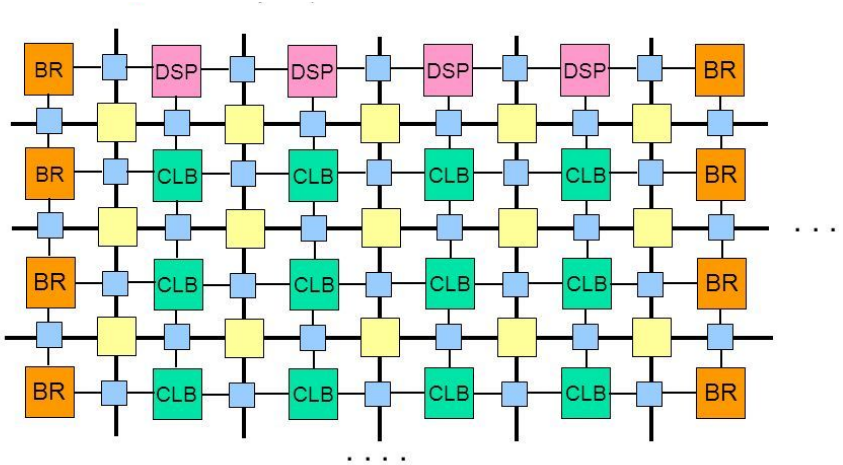
(Yash, 2017)

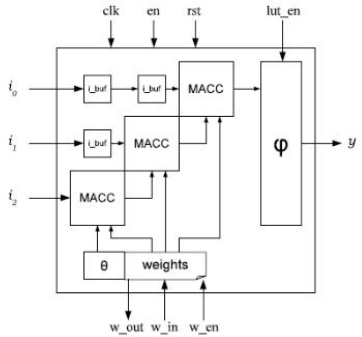
Training Algorithm

-Custom configured hardware is used

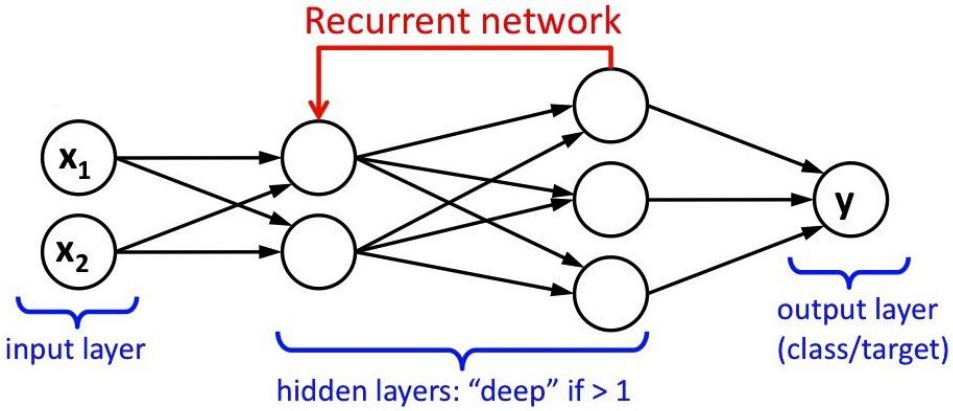


(Yash, 2017)

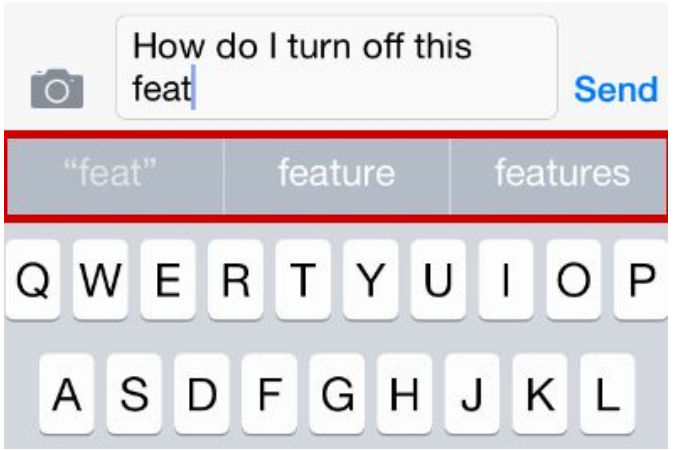




Recurrent Neural Networks (RNNs) are state-of-the-art constructs for analysis of sequenced data



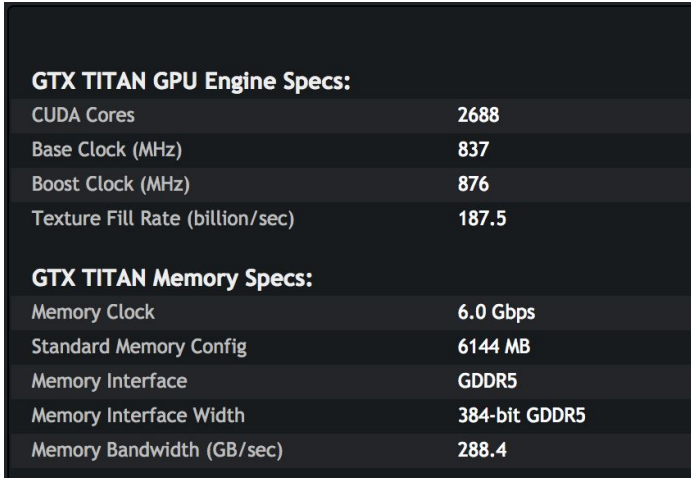
(Santos, 2017)



Specifications of the GPU and FPGA accelerators used in performance evaluation

Accelerator Designs	FPGA	ALMs	DSPs	M20Ks
8 clusters, 256 FMAs	Stratix V	296K	256	4MB
32 clusters, 1024 FMAs	Arria 10	224K	1024	4MB

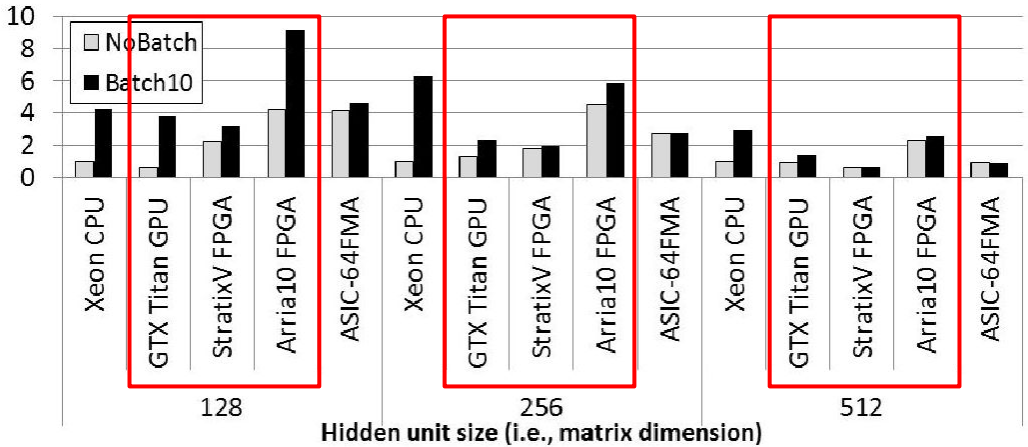
(Nurvitadhi, 2016)



GTX TITAN GPU Engine Specs:	
CUDA Cores	2688
Base Clock (MHz)	837
Boost Clock (MHz)	876
Texture Fill Rate (billion/sec)	187.5
GTX TITAN Memory Specs:	
Memory Clock	6.0 Gbps
Standard Memory Config	6144 MB
Memory Interface	GDDR5
Memory Interface Width	384-bit GDDR5
Memory Bandwidth (GB/sec)	288.4

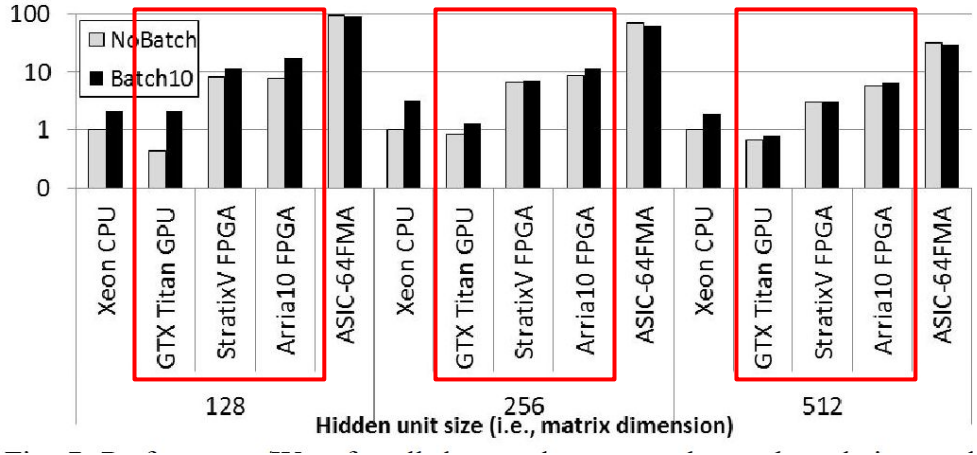
(Nvidia Corporation, 2016)

FPGA and GPU execution performance of Recurrent Neural Network relative to CPU baseline



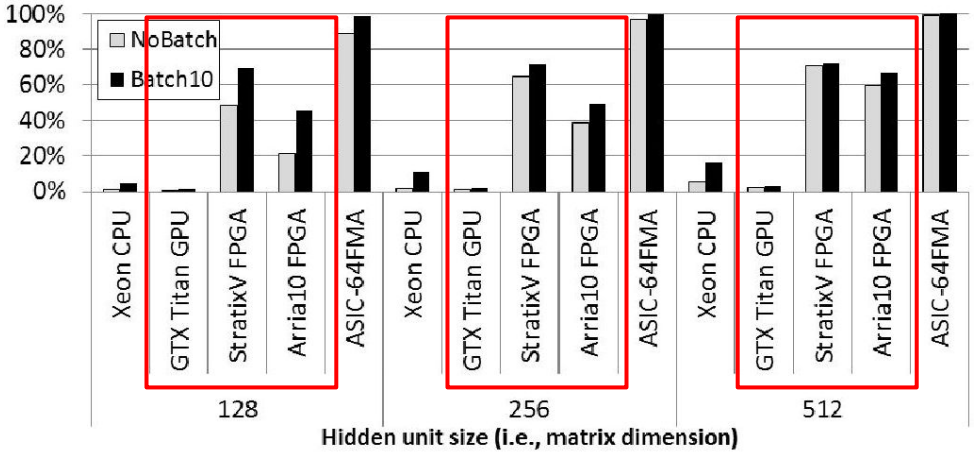
(Nurvitadhi, 2016)

FPGA and GPU efficiency of Recurrent Neural Network



(Nurvitadhi, 2016)

FPGA and GPU peak performance utilization of Recurrent Neural Network



(Nurvitadhi, 2016)

Conclusion

- Reconfigurability for specific network architectures
- Superior Efficiency
- Overall Improvements to specifications

Thank you!

Questions?