
# Going for Brain-Scale Integration using FPGAs,TSVs and NOC-based Artificial Neural Networks- A Case Study

Nowshad Painda Mand, Johnny Öberg

Electronic System Dept

School of information and Communication Technology

Royal Institute of Technology(KTH), Sweden

---

FPGAWorld Conference 9th Sep Stockholm

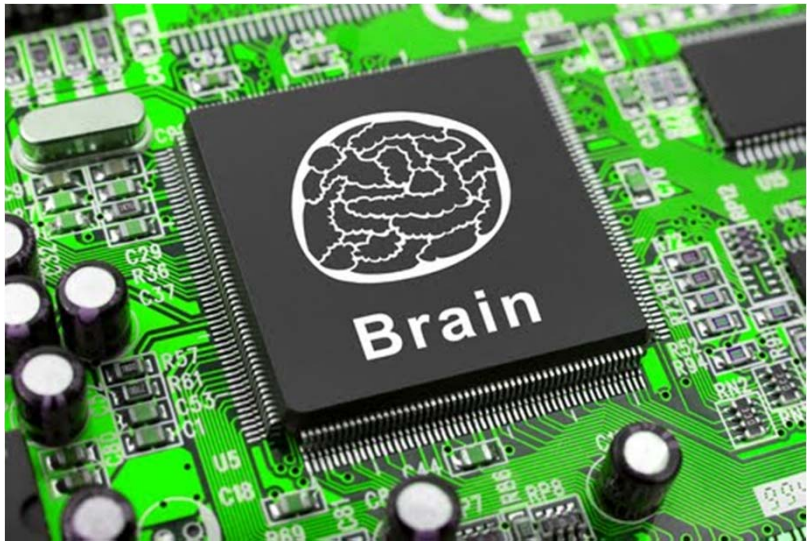
ROYAL INSTITUTE OF TECHNOLOGY

# Outline

- Motivation
- Related Work
- Introduction
- Neural Networks structure
- Vision and Architecture
- Experimental Results
- Conclusion and Future Work

ROYAL INSTITUTE OF TECHNOLOGY

# Let’s build a brain…



ROYAL INSTITUTE OF TECHNOLOGY

# Motivation-Tasks

## (Biological systems do with great speed)

- Common sense reasoning
- Simple tasks like walking/talking
- Imprecise info interpretation
- Generalization/Learning
- Voice, image,Video processing etc

ROYAL INSTITUTE OF TECHNOLOGY

# Motivation

| Biological Systems | |
|---|---|
| Processor | • Simple<br>• Low speed<br>• Large Number |
| Memory | • Distributed<br>• Content Addressable |
| Computation | • Distributed<br>• Parallel<br>• Self-learning |
| Power Consum | • Low |
| Reliability | • Very Robust |
| Complex Arith | • No |

ROYAL INSTITUTE OF TECHNOLOGY

# Motivation-Realizations

- Better understanding of brain
  - Individual cell
  - Part of brain
  - Malfunction/Diseases/Better drugs
- Artificial Intelligence
  - Pattern recogition
  - Speech and image processing
  - Robotics
- New computer architicture
  - Based on massive parrellism
    - For problem not solvable by current architectures

ROYAL INSTITUTE OF TECHNOLOGY

# Related Work-i

- SpiNNaker (Manchester Uni)
  - Tiny ARM968 CPUs on Chip,On chip/ inter-chip network, 65,536-18(1Million Cores), (Partially Completed,2013).
- C2S2(IBM+5Uni,2008)
  - First SW simulator now turned HW
  - Based on crossbar and not scalable
- FACET and BraiscaleS (19 European Uni)
  - Analog/digital Hybrid,Limited neurons can be simulated

ROYAL INSTITUTE OF TECHNOLOGY

# Related Work-ii

- BioRC (Southern California)
  - Synthtic Cortex using nano tubes and Graphene
  - Started very recently
- IBM Brain-Inspired Computer
  - One million Neurons
  - 256 million synapses
  - 5.4BT ,2nd largest CMOS Chip in the World
  - 4096 distributed cores On-Chip Mesh network
  - Over 400 million bits of local on chip memory, around 100kbits of memory per core

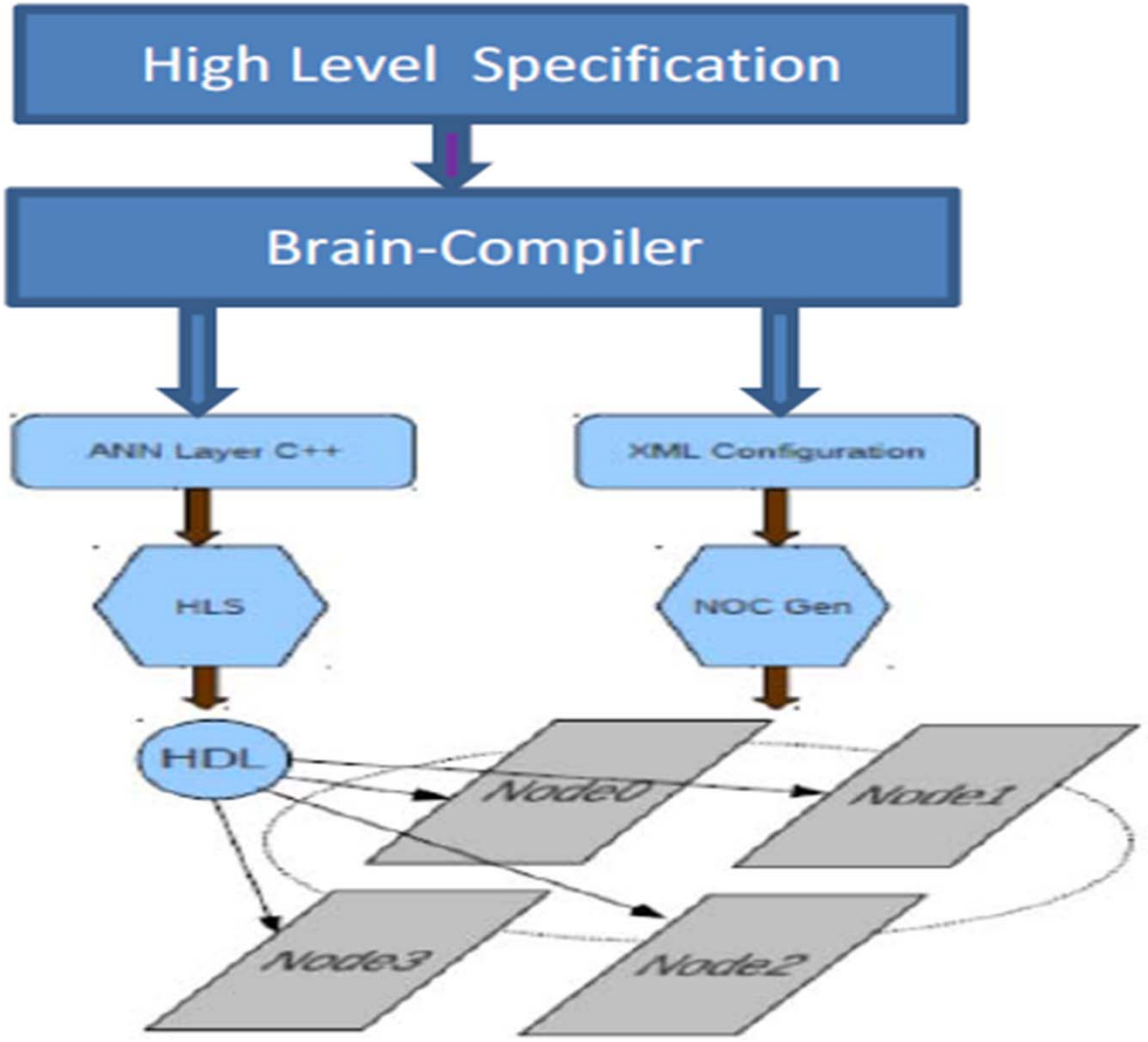
ROYAL INSTITUTE OF TECHNOLOGY

# Automated Tool for Experimental Setup

- FPGA base system
- Plasticity through reconfigurability
- Best neural structure
- Experimental platform

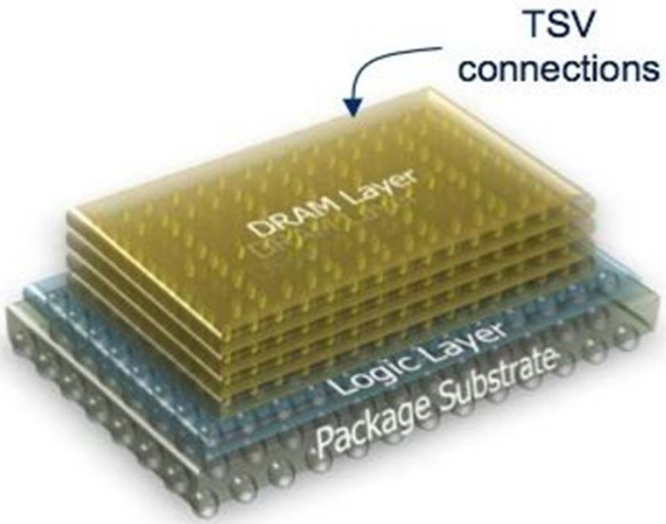
ROYAL INSTITUTE OF TECHNOLOGY

# Our Vision



ROYAL INSTITUTE OF TECHNOLOGY

# Architecture




- BPUs on chip
- BPU Board contaning BPU Chips
- Packet-switch NOC for BPUs
- Off-chip protocol for BPU Boards
- For sufficient bandwidth chip-stack of DRAMs on BPU chips

*Figure from Micron Technologies- IBM 3D Chips

ROYAL INSTITUTE OF TECHNOLOGY

# Introduction-i

- ANN modelled as
  - Software,Hardware or Hybrid
- Software realisation
  - Sequential and not suitable for real time apps
- Hardware realization
  - Ideally suited for real time
  - Conventional HW does not scale up
  - Non linear inter-neuron communication
  - Plasticity, fan-in/fan-out & power consumption

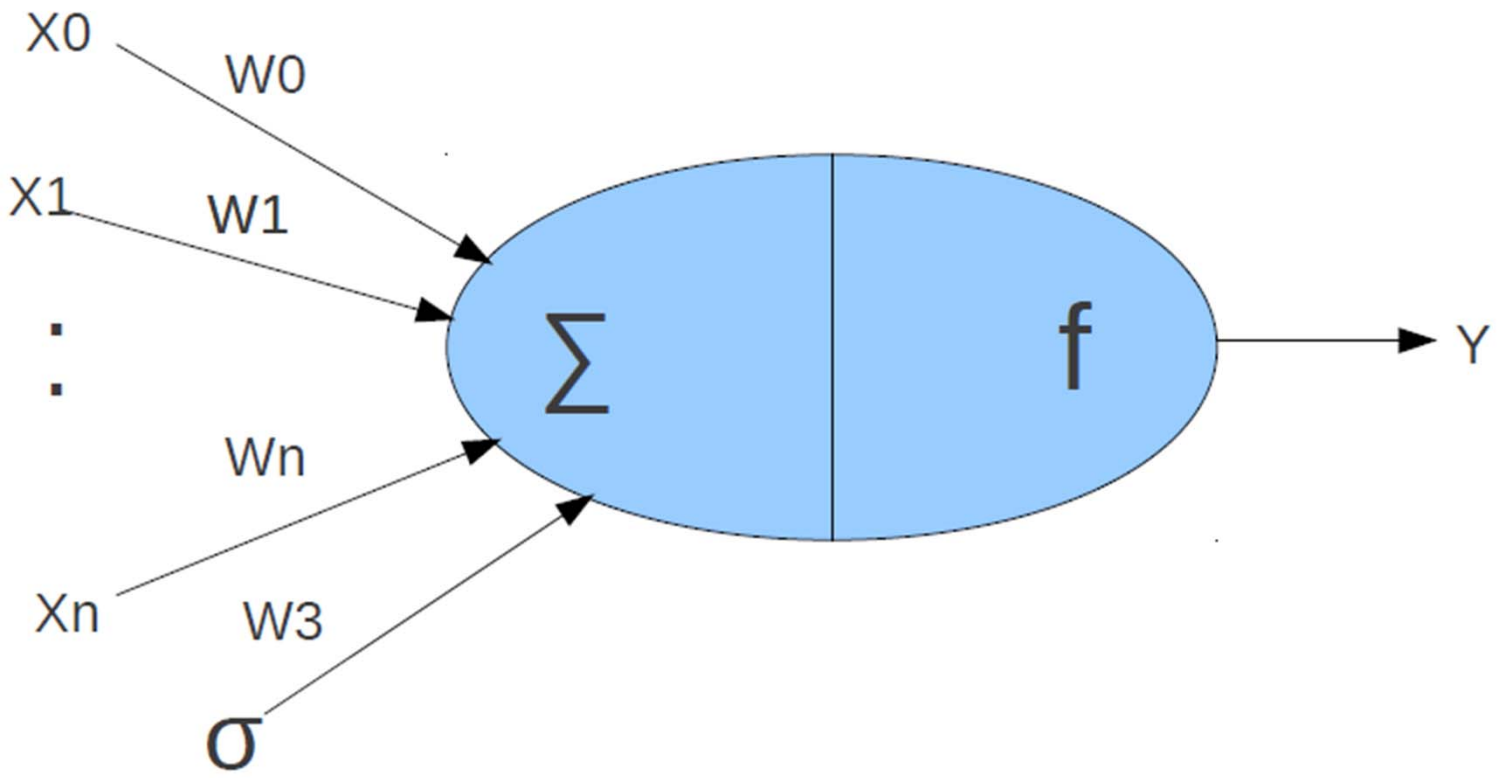
ROYAL INSTITUTE OF TECHNOLOGY

# How many FPGAs to Build a Brain ?

- High level neural net model
- Design-space exploration
- Synthesized model using HLS tools
- Mapping of synthesize model on NOC based FPGA platform

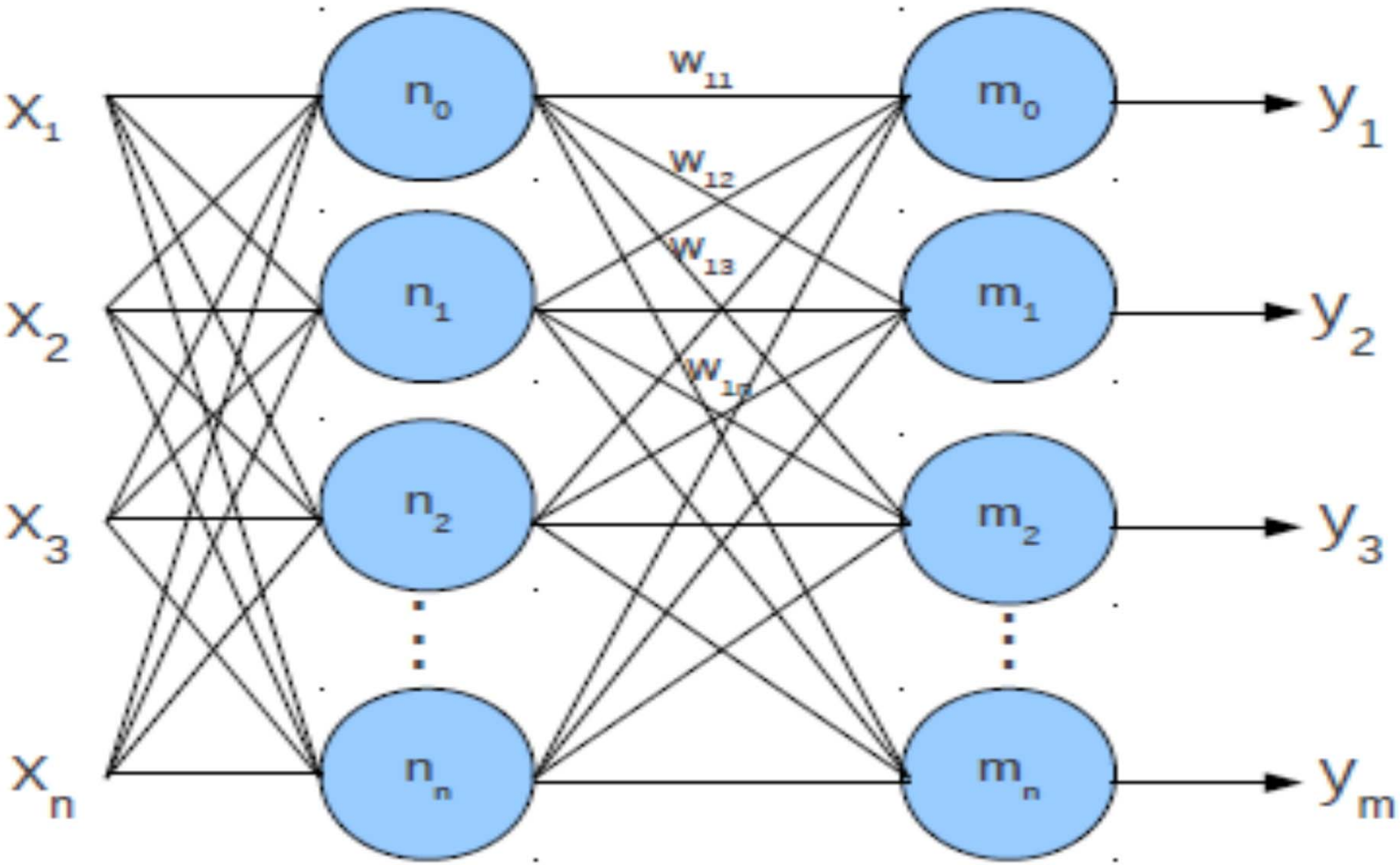
ROYAL INSTITUTE OF TECHNOLOGY

# Basic Neuron



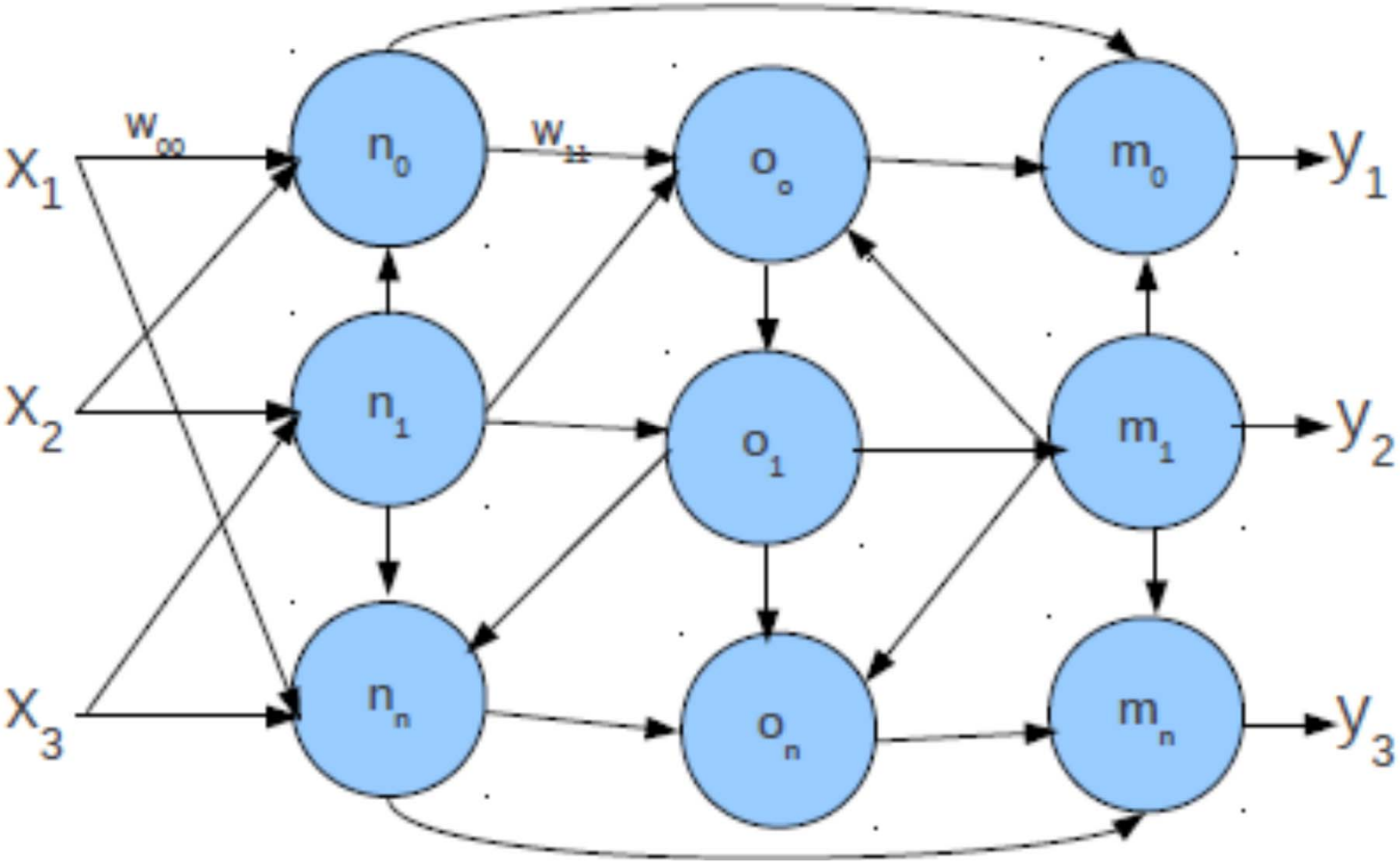
ROYAL INSTITUTE OF TECHNOLOGY

# Neural Network-i



ROYAL INSTITUTE OF TECHNOLOGY

# Neural Network-ii



ROYAL INSTITUTE OF TECHNOLOGY

# Requirements

- On Avg 1000 weights & multiply with 1000 for every neuron firing
- $10^{11}$ neurons & 1% firing every 10ms
- Need $10^3 \times 10^{11}$ weighted memory
- Need to fetch $10^3 \times 10^{11} \times 1\% / 10_{ms} = 10^{14}$ input weights per sec

ROYAL INSTITUTE OF TECHNOLOGY

# Design-Space Exploration

- No of BPUs depends on
  - Speed of the DRAM-don't shrink with technology
  - Amount of DRAM on chip-stack-more TSVs means simple logic (row/column addr), pipelining to speedup increase area so speed won't be an issue
  - No of TSVs possible
  - Storage limited by no of dies in stack(Technology) & no of TSVs
  - Finally no of TSVs limiting factor

ROYAL INSTITUTE OF TECHNOLOGY

# Design-Space Exploration

| Global Level, W2W, D2W or D2D 3D-stacking | 2009-2012 | 2012-2015 |
|---|---|---|
| Minimum TSV diameter | 4-8 µm | 2-4µm |
| Minimum TSV pitch | 8-16 µm | 4-8 µm |

-Stacking TSVs diameter as per ITRS Roadmap,2012

ROYAL INSTITUTE OF TECHNOLOGY

# Design-Space Exploration

- Conservative approach with TSV pitch 8µm
  - TSV density $1cm^2/64\mu m^2 = 1{,}562{,}500$ TSV/$cm^2$
  - Never need $\log_2(10^{15}) = 50$ bit to address whole memory
  - One calculation need 1000 input weight
  - Address bits and result bits are tiny and can be ignored

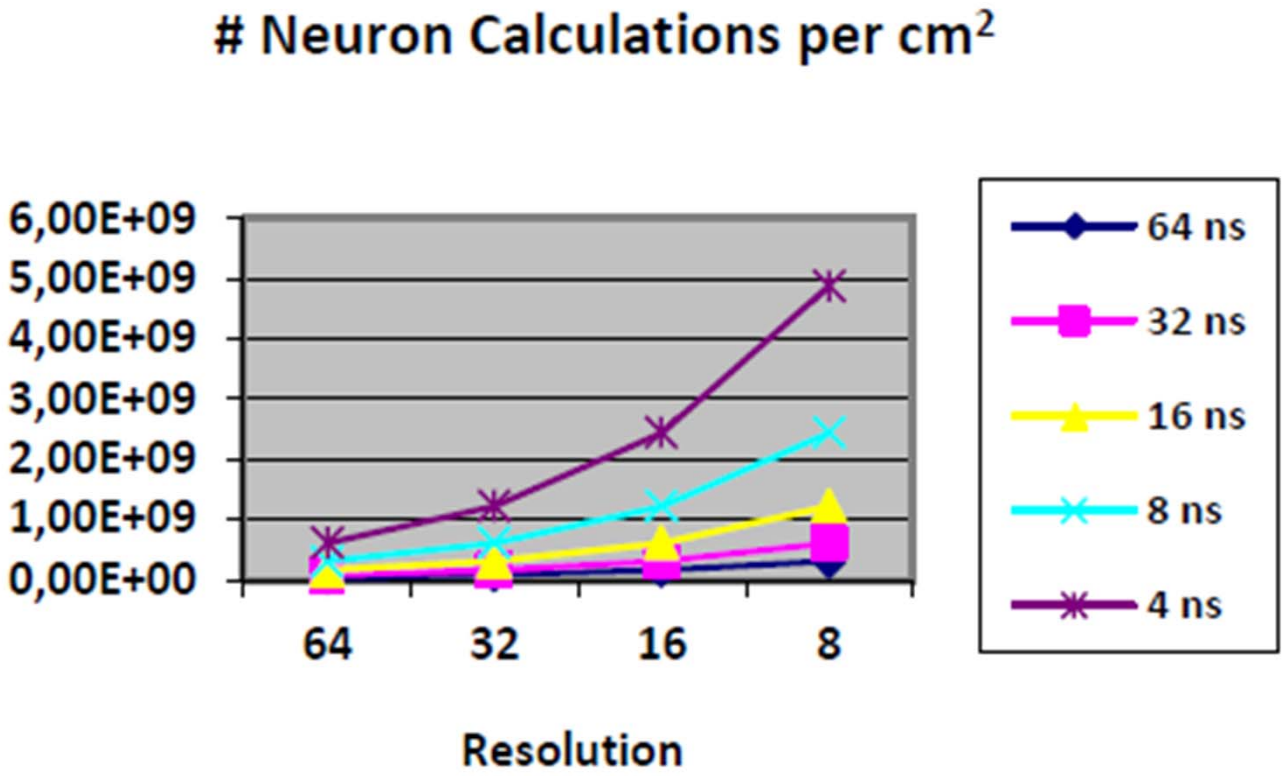
ROYAL INSTITUTE OF TECHNOLOGY

# Design-Space Exploration

$$NrNeuronCalcsPerCycleCm2 = \frac{TSV_{density}}{resolution * NrInputWeightPairsPerNeuron}$$

ROYAL INSTITUTE OF TECHNOLOGY

# Design-Space Exploration



ROYAL INSTITUTE OF TECHNOLOGY

# Design-Space Exploration

| Input-Weight Resolution [bits] | | 64 | 32 | 16 | 8 | 4 |
|---|---|---|---|---|---|---|
| DRAM Size [Byte/brain] | | $8*10^{15}$ | $4*10^{15}$ | $2*10^{15}$ | $10^{15}$ | $5*10^{14}$ |
| #Chips | 100 | $8*10^{13}$ | $4*10^{13}$ | $2*10^{13}$ | $10^{13}$ | $5*10^{12}$ |
| | 1,000 | $8*10^{12}$ | $4*10^{12}$ | $2*10^{12}$ | $10^{12}$ | $5*10^{11}$ |
| | 10,000 | $8*10^{11}$ | $4*10^{11}$ | $2*10^{11}$ | $10^{11}$ | $5*10^{10}$ |
| | 100,000 | $8*10^{10}$ | $4*10^{10}$ | $2*10^{10}$ | $10^{10}$ | $5*10^{9}$ |

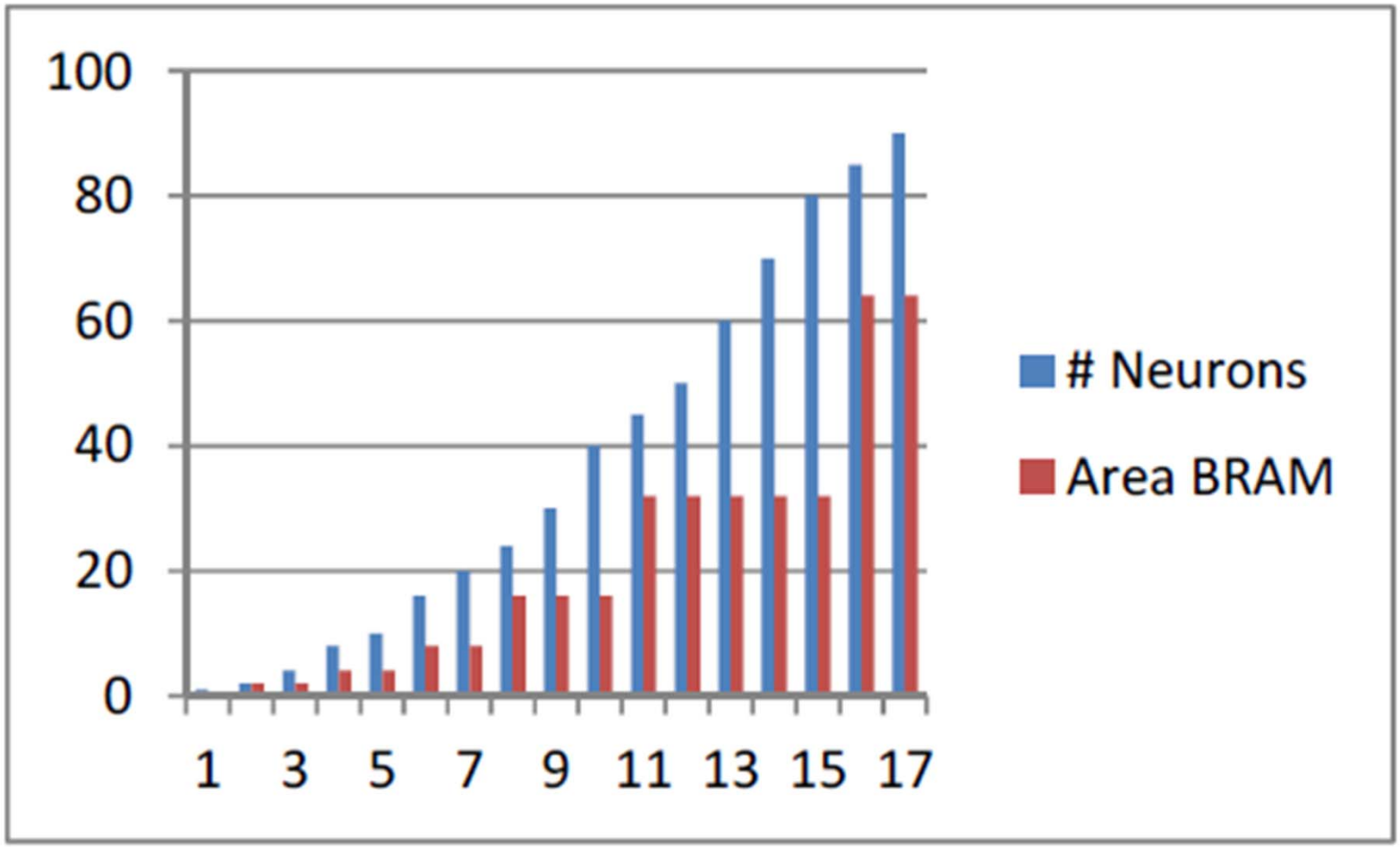
ROYAL INSTITUTE OF TECHNOLOGY

# Experimental Results

- For Vivado HLS two cases are considered
  - Case-1 weights are modelled as beig read from input ports
  - Case-2 weights are modelled as being synthesized as on-chip BRAM
- NOC synthesized separately
  - NOC node consumed around 750LUTs
  - One BPU around 18000LUTs

ROYAL INSTITUTE OF TECHNOLOGY

# No. of Neurons vs FPGA Area



ROYAL INSTITUTE OF TECHNOLOGY

# Weight Data Type Dependencies

| Weight Data Type | Area [LUT] | Delay [clk-cc] |
|---|---|---|
| 64-bit Double Serial | 13807 | 5563 |
| 64-bit Integer Serial | 15862 | 5005 |
| 64-bit Integer fully parallel | 64000[1] | 1 |
| 32-bit Single Serial | 7201 | 5563 |
| 32-bit Integer Serial | 8957 | 5004 |
| 16-bit Integer Serial | 5555 | 5004 |
| 8-bit Integer Serial | 3504 | 5004 |
| 4-bit Integer Serial | 2652 | 5003 |

[1]Vivado could not handle fully parallel case, assumed manual calc

ROYAL INSTITUTE OF TECHNOLOGY

# Throughput Considerations

| Input-Weight Resolution [bits] | 64 | 32 | 16 | 8 |
|---|---|---|---|---|
| Production rate #Neuron Calculations/cycle/chip | 54.9 | 109 | 220 | 439,4 |
| Consumption rate #Neuron Calculations/cycle/chip | 15 | 31 | 62 | 125 |
| Fraction of required rate | 27% | 28% | 28% | 28% |

ROYAL INSTITUTE OF TECHNOLOGY

# Most Feasible Design

| DRAM Access Time [ns] | Resolution (#bits per Input-weight pair) | #Chips (2.25 $cm^2$) | DRAM Size (Bytes per chip) |
|---|---|---|---|
| 8 | 64 | $5.5*10^5$ | $1,5*10^9$ |
| | 32 | $2.7*10^5$ | |
| | 16 | $1.3*10^5$ | |
| | 8 | $6.6*10^4$ | |

ROYAL INSTITUTE OF TECHNOLOGY

# Conclusion

- Design will be area limited for FPGAs
- Using DRAMs ANN models can be mapped to modern FPGAs
- Equivalent of full brain will need ~66,000 chips each with 1.5Gb DRAM stack

ROYAL INSTITUTE OF TECHNOLOGY

# Future work

- More types of neuron models
- Other types of activation functions
- Build specialized NOC for ANN
- Build complete model with limited neurons
- HLS tool for massive parallelism

ROYAL INSTITUTE
OF TECHNOLOGY

# Questions??

ROYAL INSTITUTE OF TECHNOLOGY

# Design-Space exploration

| DRAM Access Time [ns] | Resolution (#bits per Input-weight pair) | #Neuron Calculations per second per $cm^2$ | Area needed for entire brain [$cm^2$] | #Chips (2.25 $cm^2$) |
|---|---|---|---|---|
| 64 | 64 | $3.8*10^8$ | $26*10^5$ | $12*10^5$ |
| | 8 | $30.5*10^8$ | $3.3*10^5$ | $1.5*10^5$ |
| 32 | 64 | $7.6*10^8$ | $13*10^5$ | $5.8*10^5$ |
| | 8 | $61*10^8$ | $1.6*10^5$ | $7.3*10^4$ |
| 16 | 64 | $15.3*10^8$ | $6.6*10^5$ | $2.9*10^5$ |
| | 8 | $122*10^8$ | $8.2*10^5$ | $3.6*10^4$ |
| 8 | 64 | $30.5*10^8$ | $3.3*10^5$ | $1.5*10^5$ |
| | 8 | $244*10^8$ | $4.1*10^4$ | $1.8*10^4$ |
| 4 | 64 | $61*10^8$ | $1.6*10^5$ | $7.3*10^4$ |
| | 8 | $488*10^8$ | $2.0*10^4$ | $9.1*10^3$ |
| 2 | 64 | $122*10^8$ | $8.2*10^4$ | $3.6*10^4$ |
| | 8 | $977*10^8$ | $1.0*10^4$ | $4.6*10^3$ |
| 1 | 64 | $244*10^8$ | $4.1*10^4$ | $1.8*10^4$ |
| | 8 | $1{,}950*10^8$ | 5,120 | 2,276 |

ROYAL INSTITUTE OF TECHNOLOGY

# Design-space exploration

- Cycle time is determined by DRAM access time
- More agressive pitch 4um
  - We get values 4 times larger and 4 times smaller in terms of area
- Assume DRAM access time of 1ns for the next table

ROYAL INSTITUTE OF TECHNOLOGY

# Design-Space exploration

| **Input-Weight Resolution [bits]** | 64 | 32 | 16 | 8 |
|---|---|---|---|---|
| **# Neuron Calculations/cycle/cm$^2$** | 97.7 | 195.3 | 390.6 | 781.3 |
| **# Chips** | 4,551 | 2,276 | 1,138 | 569 |
| **# BPU Boards (100 chips/board)** | 45.6 | 22.8 | 11.4 | 5.7 |

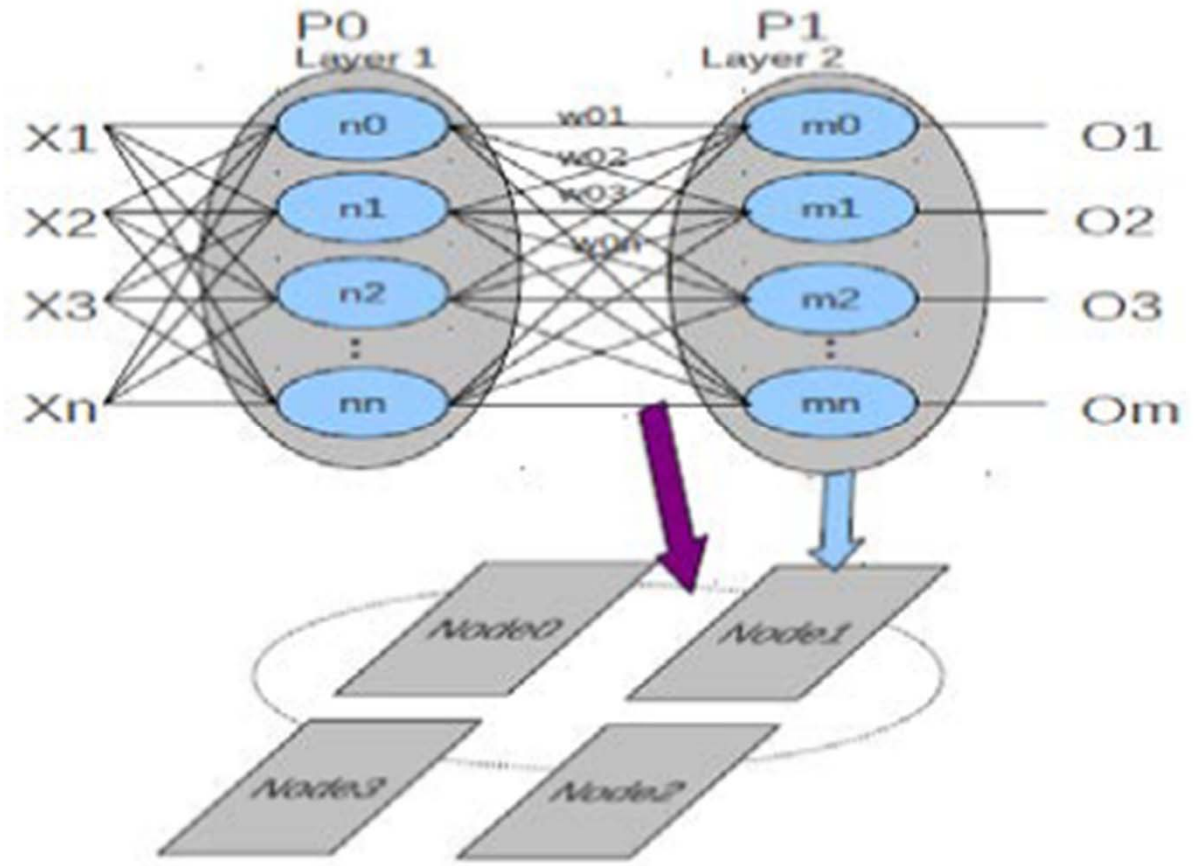
ROYAL INSTITUTE OF TECHNOLOGY

# BPU Implementation

- 1000xprec x res[bits] need to be fetched for single Neuron calc
- For actual BPUs we selected multi-layer perceptron model
- Minicolumn & macrocolumns
- Packet switch NOC to connect BPUs

ROYAL INSTITUTE OF TECHNOLOGY

# Mapping BPUs to NOC platform



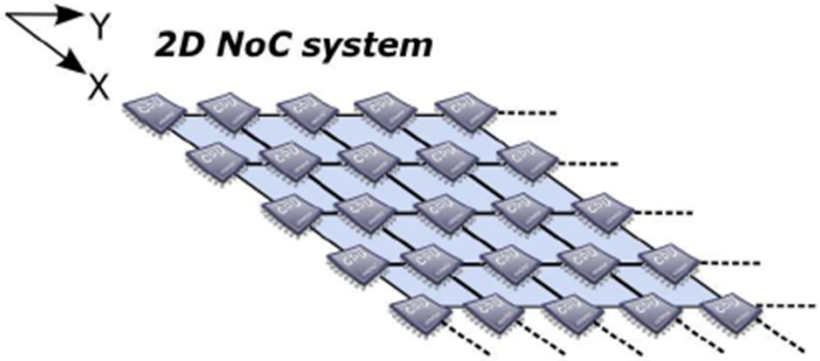
ROYAL INSTITUTE OF TECHNOLOGY

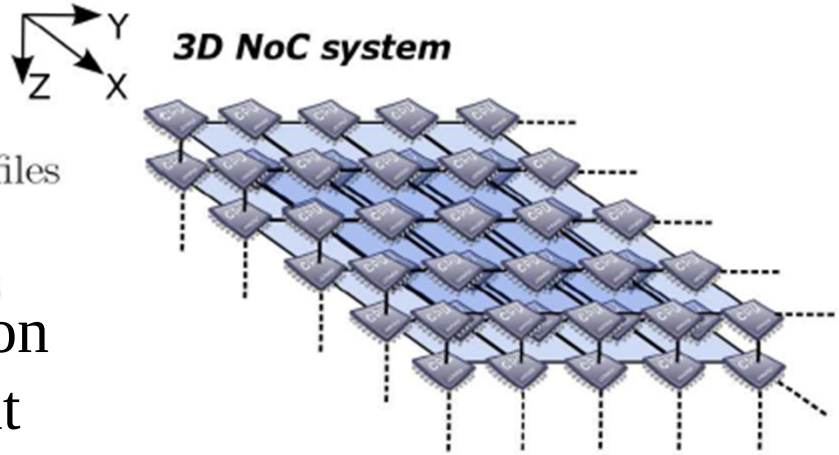
# NOC Generator


NOC Gen

files

t




Sopc/mhs files Creation
Partial Pin Assignment

ALTERA
Nios II
QUARTUS II
XILINX
MicroBlaze
AEROFLEX GAISLER
LEON3 Synthesizable processor

ROYAL INSTITUTE OF TECHNOLOGY

# Vivado HLS tool

- Xilinx state of the art HLS tool
- ANN's layer behavrial model in c++
- Single layer is synthsized
- Leaving low-level optimizations to tool
- For new point in desig-space just change any parameter and re-run

ROYAL INSTITUTE OF TECHNOLOGY

# Nr of Neuron vs Resource requirement & latency

| # Neu | BRAM (18K) | DSP48E | FF | LUT | Lat (clk-cc) |
|---|---|---|---|---|---|
| 1 | 0 | 14 | 13591 | 15871 | 616 |
| 2 | 0 | 28 | 13610 | 16251 | 668 |
| 4 | 0 | 28 | 13619 | 16336 | 768 |
| 8 | 0 | 28 | 13628 | 16421 | 968 |
| 10 | 0 | 28 | 13628 | 16770 | 1068 |
| 20 | 0 | 28 | 13637 | 16892 | 1568 |
| 30 | 0 | 28 | 13642 | 17043 | 2068 |
| 40 | 0 | 28 | 13646 | 17013 | 2568 |
| 50 | 0 | 28 | 13652 | 17180 | 3068 |
| 60 | 0 | 28 | 13652 | 17170 | 3568 |
| 70 | 0 | 28 | 13655 | 17254 | 4068 |
| 80 | 0 | 28 | 13655 | 17135 | 4568 |
| 90 | 0 | 28 | 13655 | 17319 | 5068 |
| 100 | 0 | 28 | 13662 | 17311 | 5568 |