Magee Campus

Hierarchical Network-on-Chip and Traffic Compression for Spiking Neural Network Implementations

Snaider Carrillo, Jim Harkin, Liam McDaid
University of Ulster, Magee Campus

Sandeep Pande, Seamus Cawley, Brian McGinley, Fearghal Morgan
National University of Ireland, Galway Campus

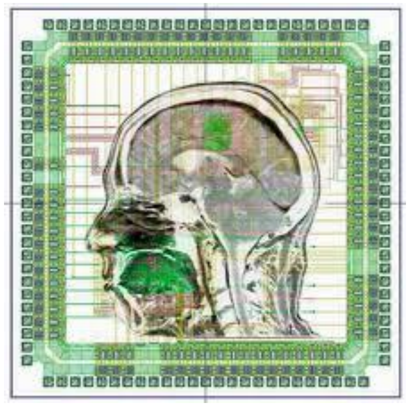
Outline

- **Motivation and Challenges**
- **Hierarchical NoC EMBRACE Architecture**
- **Performance Analysis**
- **Take-home Message & Future Work**

Motivation: Engineer & Neuroscientist

Neural processing systems......Taking inspiration from the biology......to deploy a new computer architecture paradigm !!!

- **An Engineering point of view....**
  - Pattern recognition + Low power consumption
  - Fault-tolerant computers +Self repairing systems

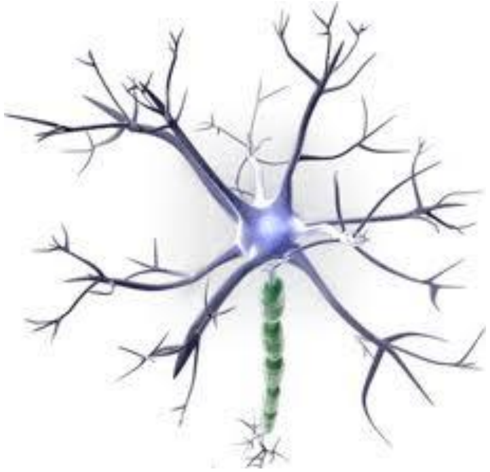


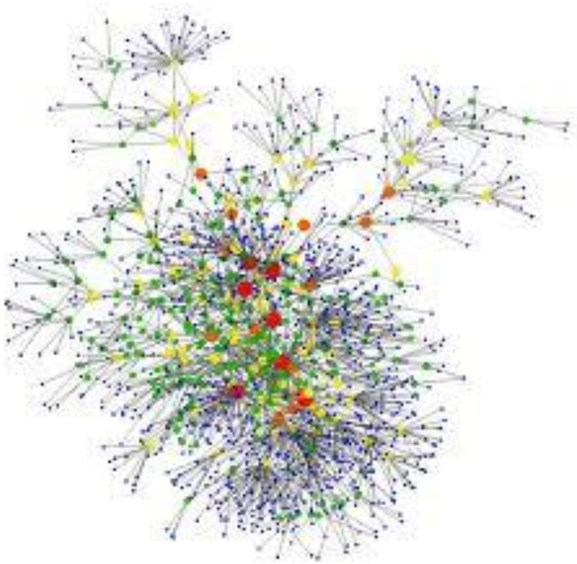
- **A Neuroscientist point of view....**
  - Faster large-scale neural network simulations
  - Ultimately, to learn a bit more about how the human brain works

Neuron Interconnection: The big challenge

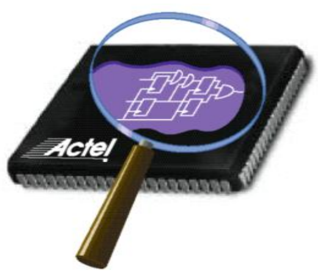
A human brain contains in average...

- *10^{11} neurons*
- *10^{15} synapses*
- *1:1000 Fan in/out connection ration*





We need a highly optimised architecture to overtake this challenge....



Previous Work

Blue Brain Project [Markram'03]

- IBM BlueGene/L supercomputer

SpiNNaker [Furber'06]

- Embedded ARM processors + NoC interconnection

SYNAPSE Project [Modha'11]

- Digital neurons + Crossbar fabric

Neurogrid [Boahen'09]

- Analogue neurons + on-chip routers

FACETS [Schemmel'05]

- Analogue neurons + hierarchical intra-wafers buses

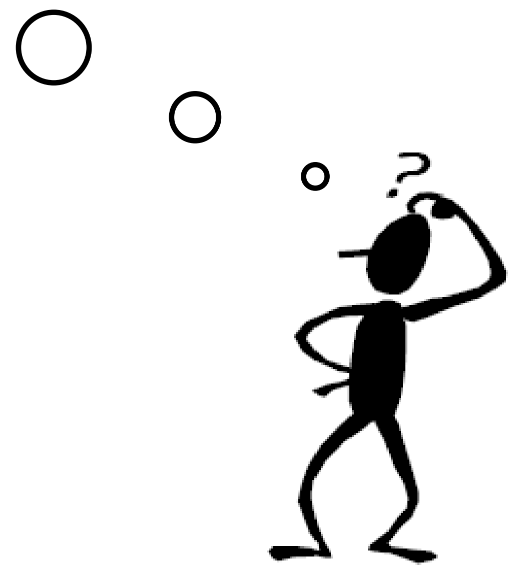
....However, there is still room for improvement ☺

Key Research Problem

*…How to **interconnect** a large number of **spiking neurons** in a **network** fashion **efficiently?**…*

Efficiently?... *a trade-off between*

- *Scalability*
- *Area utilisation*
- *Power consumption*
- *Throughput*
- *Synapse/neuron ratio*

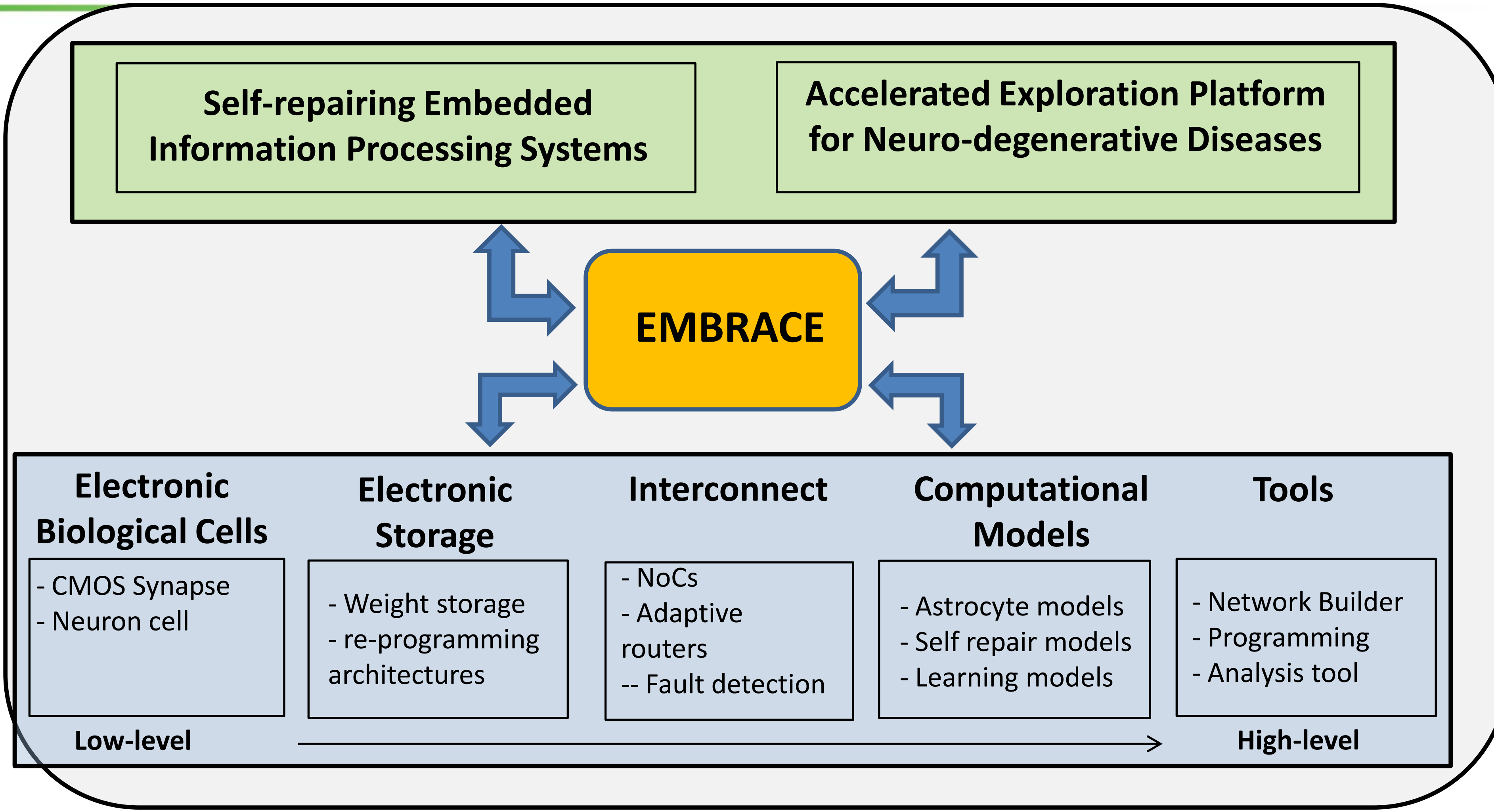


.....And what about hardware acceleration !!

Outline

- **Motivation and Challenges**
- **Hierarchical NoC EMBRACE Architecture**
- **Performance Analysis**
- **Take-home Message & Future Work**

EMulating Biologically-inspiRed ArChitectures in hardwarE (EMBRACE)



- Ulster
- University of Liverpool (S Hall)

- Ulster
- NUI Galway (F Morgan)

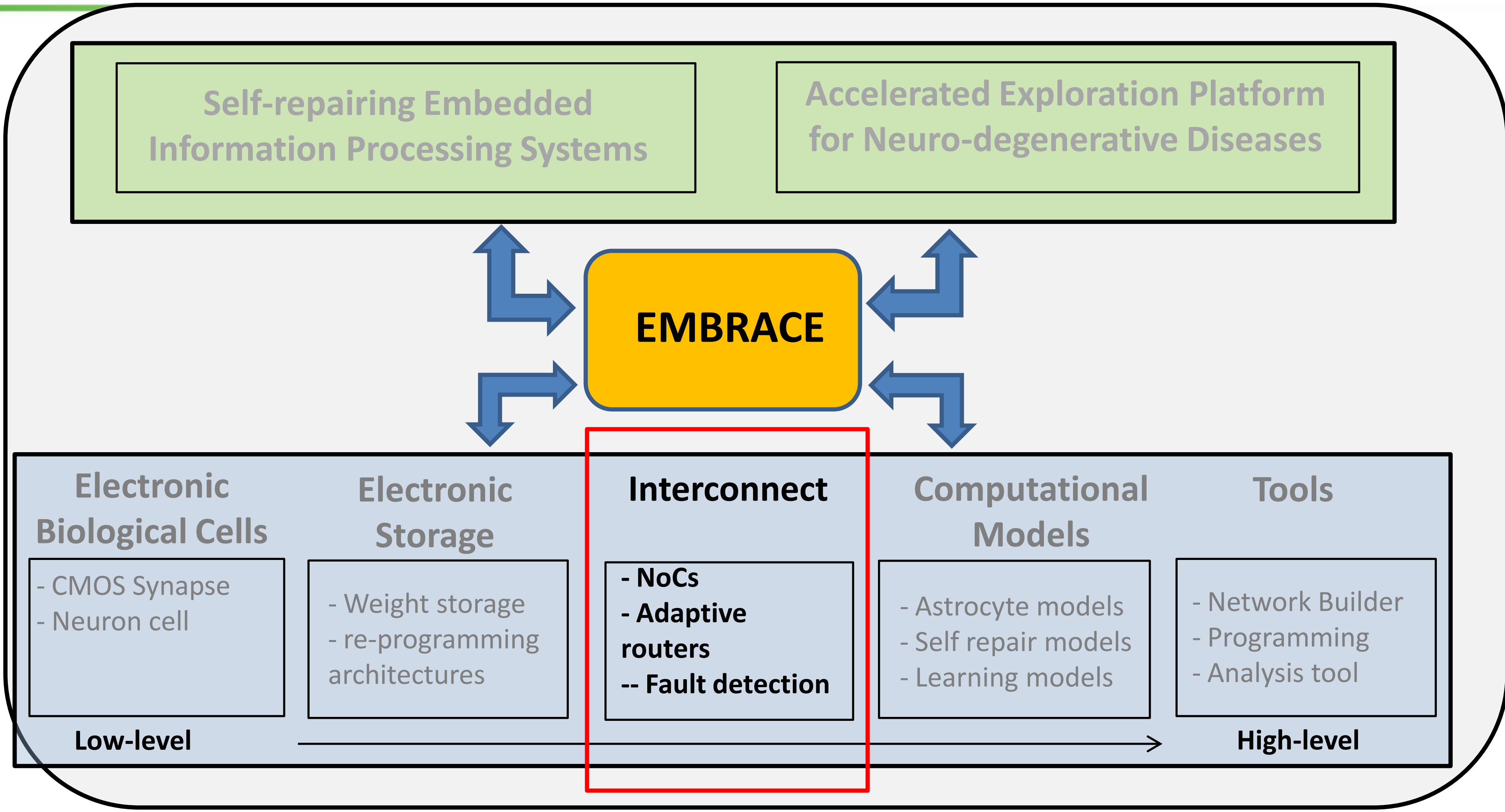
- Ulster
- University of Cardiff (Prof. V Cruneli)

- NUI Galway (F Morgan)

University of Ulster

NUI Galway OÉ Gaillimh

EMulating Biologically-inspiRed ArChitectures in hardwarE (EMBRACE)



- Ulster
- University of Liverpool (S Hall)

- Ulster
- NUI Galway (F Morgan)

- Ulster
- University of Cardiff (Prof. V Cruneli)

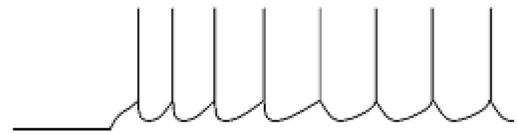
- NUI Galway (F Morgan)

University of Ulster

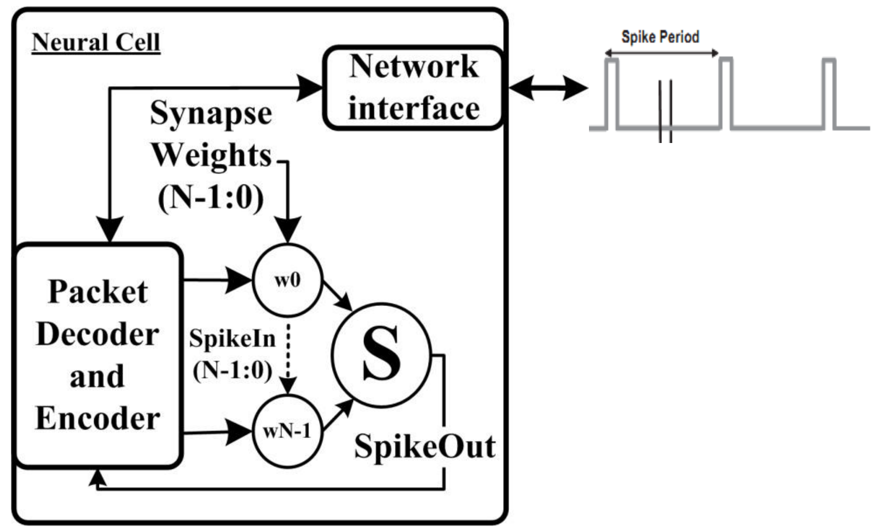
NUI Galway OÉ Gaillimh

EMBRACE Neural Cell

- **Provides:**
  - An analogue **point neuron** (Leaky Integrate & Fire model)
  - Its correspondent **synapse cells** (Dynamic Synapses)
  - A **packet decoder/encoder**
  - A **network interface** to communicate with **digital NoC router**



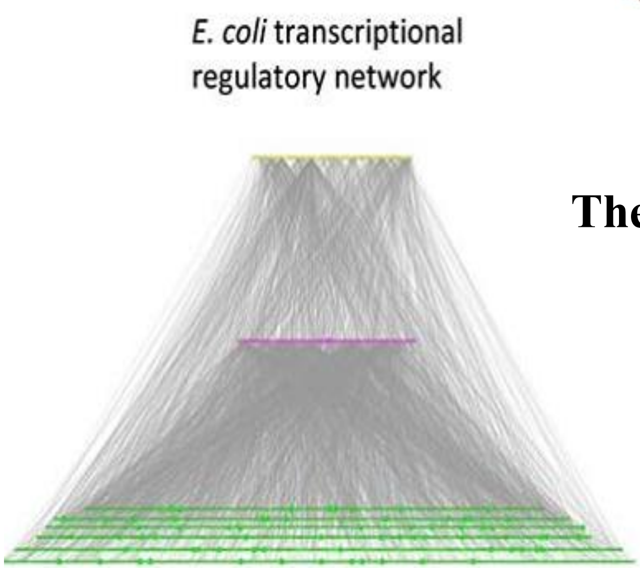
(a) Regular-spiking neurons

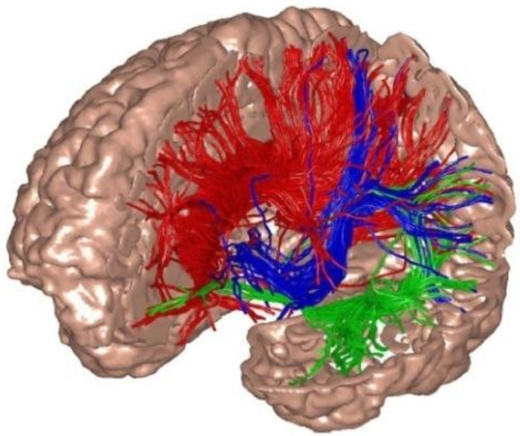


<u>On-going EPSRC project between:</u>
- University of Ulster
- University of Liverpool (S Hall)

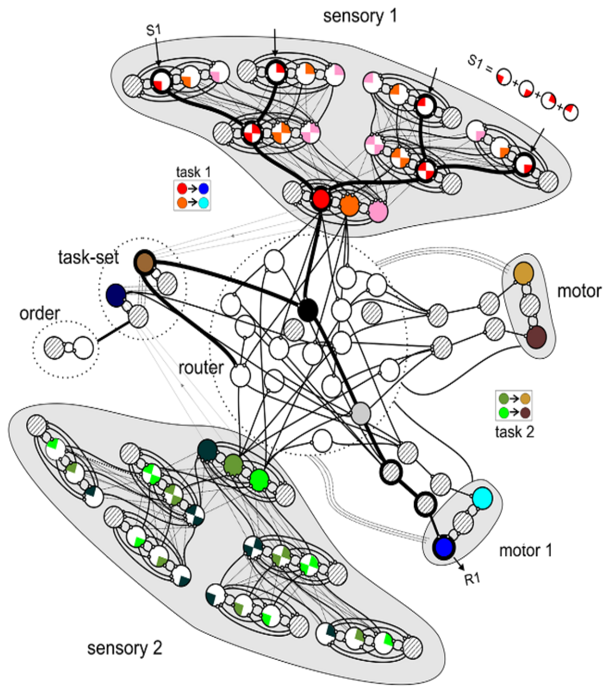
L. McDaid, S. Hall, and P. Kelly, "A programmable facilitating synapse device," in *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, 2008, pp. 1615-1620.

Hierarchical Topology:
Taking inspiration from the biology........





The brain is a 3D structure !!



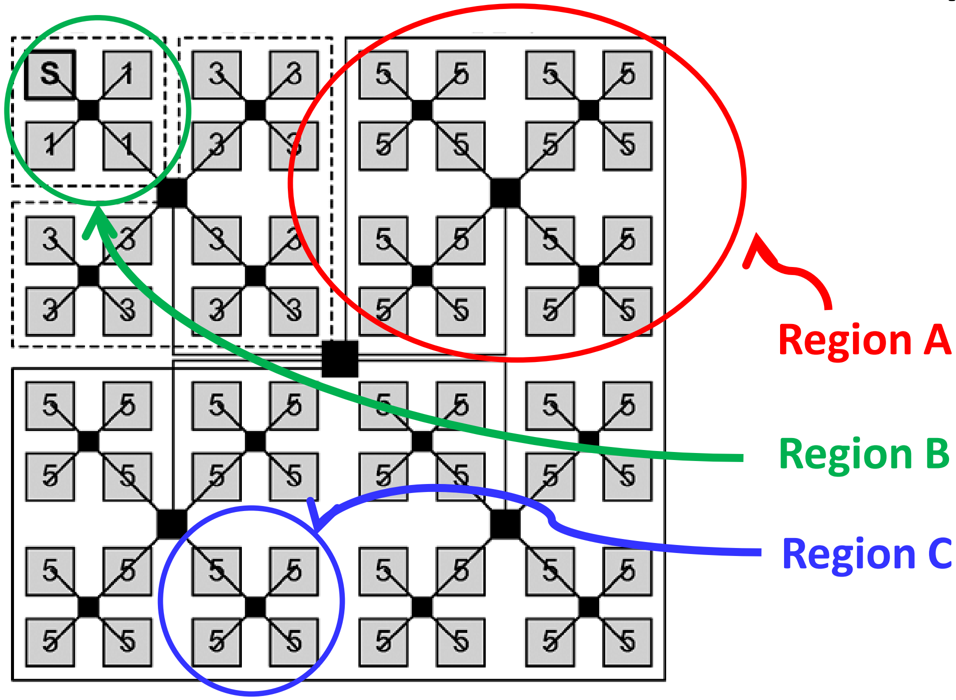
The hierarchical topology of the E. Coli (Yan et al. 2010)

A Schematic representation of a cluster of neurons (Zylberberg et al. 2010)

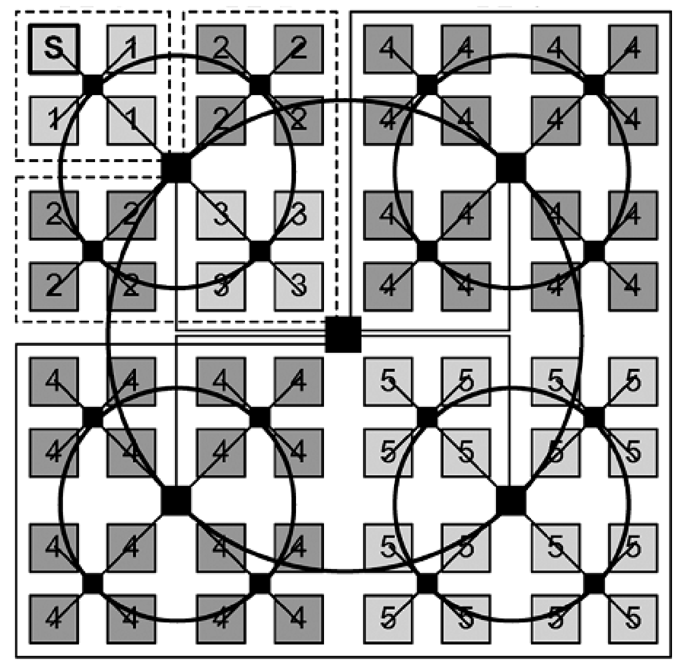
Hierarchical Topology:
.... and also from the NoC community !!

•Hierarchical NoCs **(H-NoCs)** exploit the concept of **region-based routing.**

•**Virtual regions** or facilities are used to allocate resources that process either **local or global traffic.**



Hierarchical star [1]

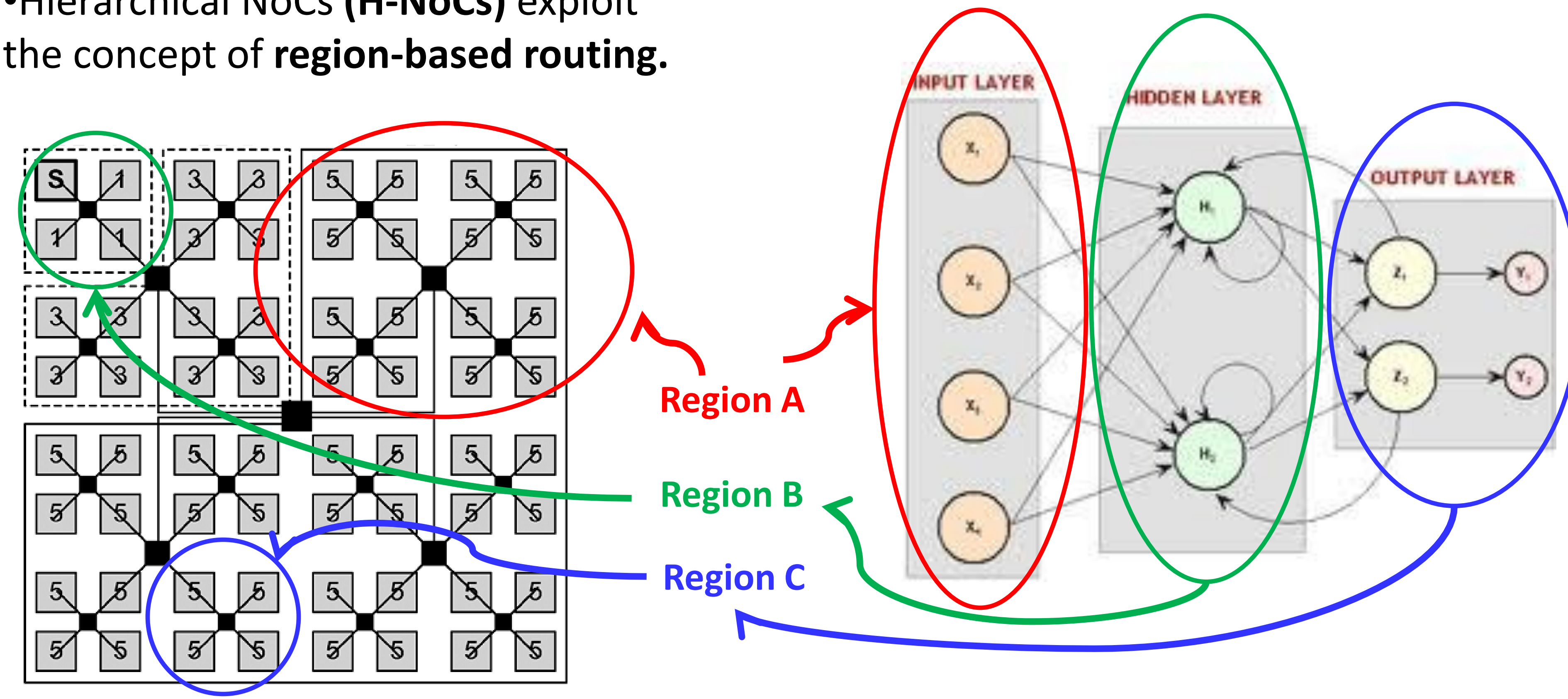


Hierarchical star + ring [1]

[1] J.-Y. Kim, J. Park, S. Lee, M. Kim, J. Oh, and H.-J. Yoo, "A 118.4 GB/s Multi-Casting Network-on-Chip With Hierarchical Star-Ring Combined Topology for Real-Time Object Recognition," *IEEE Journal of Solid-State Circuits*, vol. 45, no. 7, pp. 1399-1409, Jul. 2010

Hierarchical Topology:
…….. and also from the NoC community !!

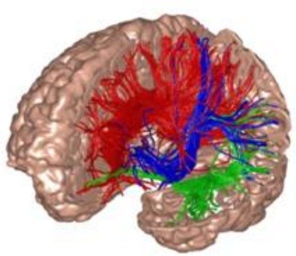
•Hierarchical NoCs **(H-NoCs)** exploit the concept of **region-based routing.**



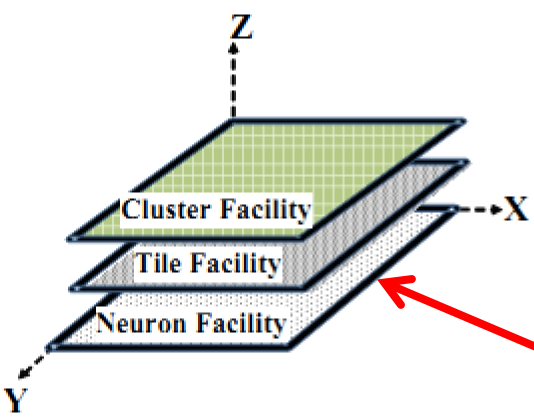
Hierarchical star [1]

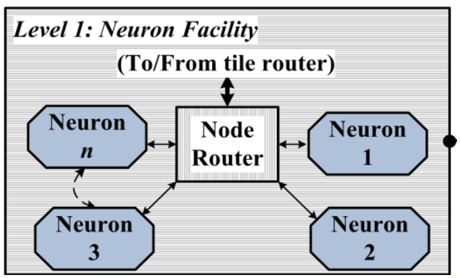
[1] J.-Y. Kim, J. Park, S. Lee, M. Kim, J. Oh, and H.-J. Yoo, "A 118.4 GB/s Multi-Casting Network-on-Chip With Hierarchical Star-Ring Combined Topology for Real-Time Object Recognition," *IEEE Journal of Solid-State Circuits*, vol. 45, no. 7, pp. 1399-1409, Jul. 2010

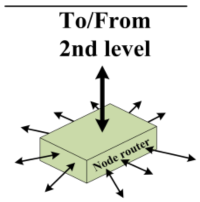
EMulating Biologically-inspiRed ArChitectures in hardwarE (EMBRACE): H-NoC Approach



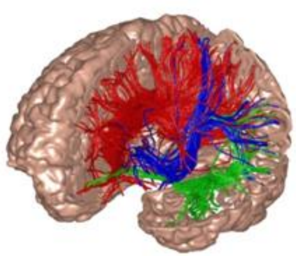
The brain is a 3D structure!!



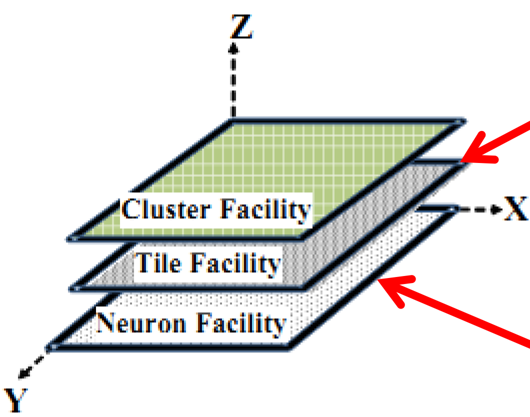


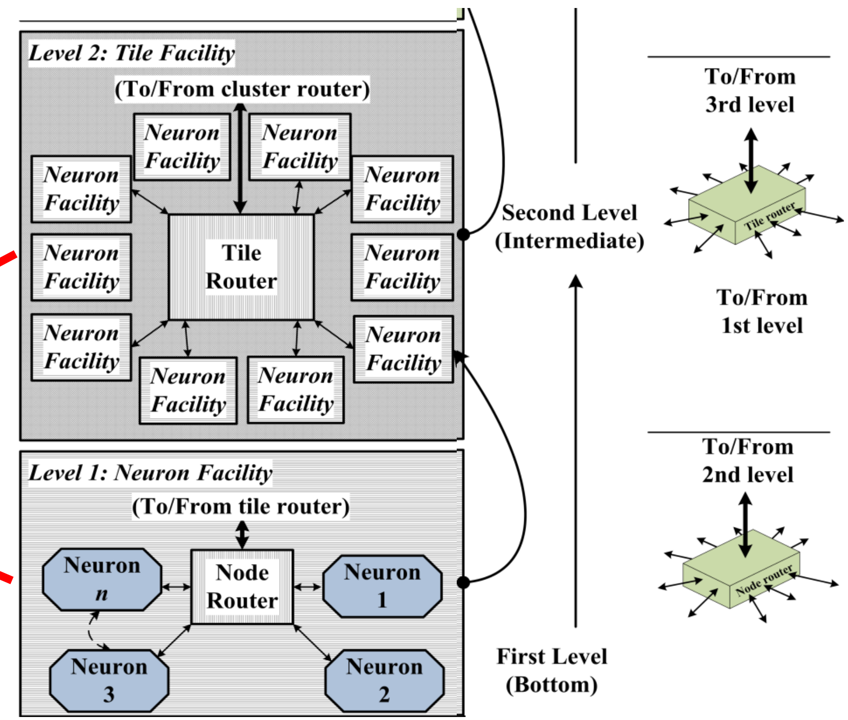


EMulating Biologically-inspiRed ArChitectures in hardwarE (EMBRACE): H-NoC Approach

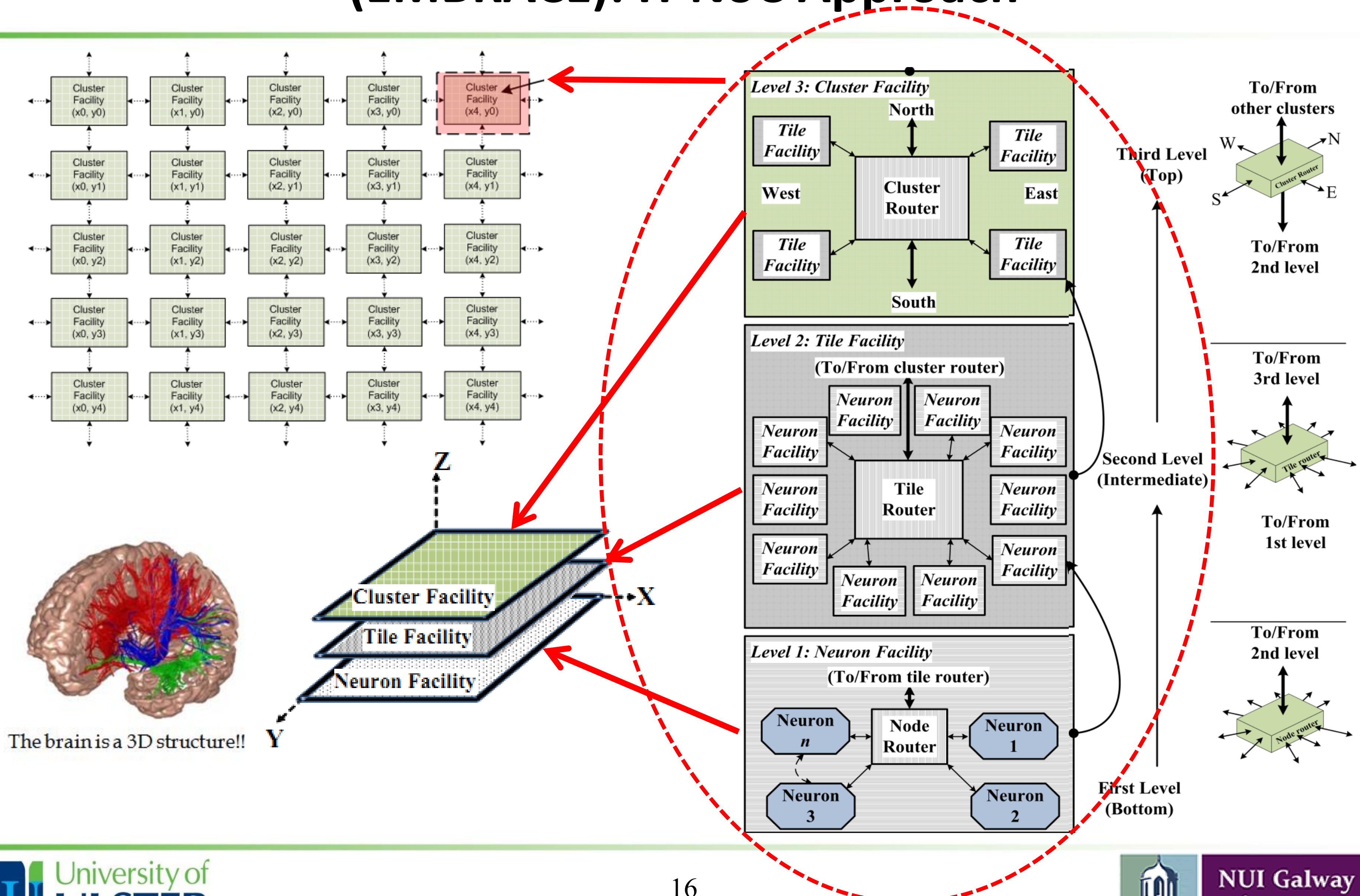


The brain is a 3D structure!!

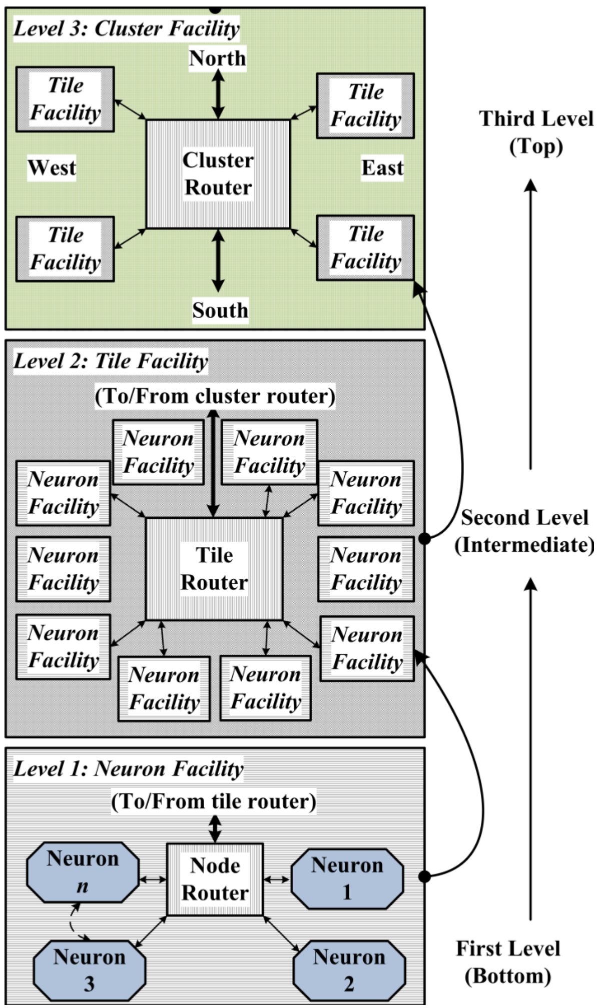




EMulating Biologically-inspiRed ArChitectures in hardwarE (EMBRACE): H-NoC Approach



EMBRACE: H-NoC Architecture



One Cluster Facility contains:

1 Cluster NoC router

4 Tile NoC routers

40 Node NoC Router

A total of 45 NoC Router
to interconnect 400 neurons

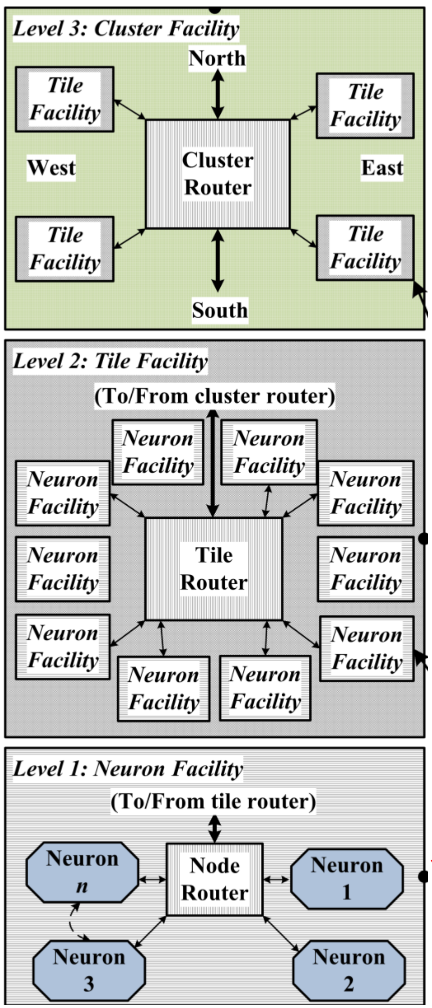
....This is just an initial density ☺

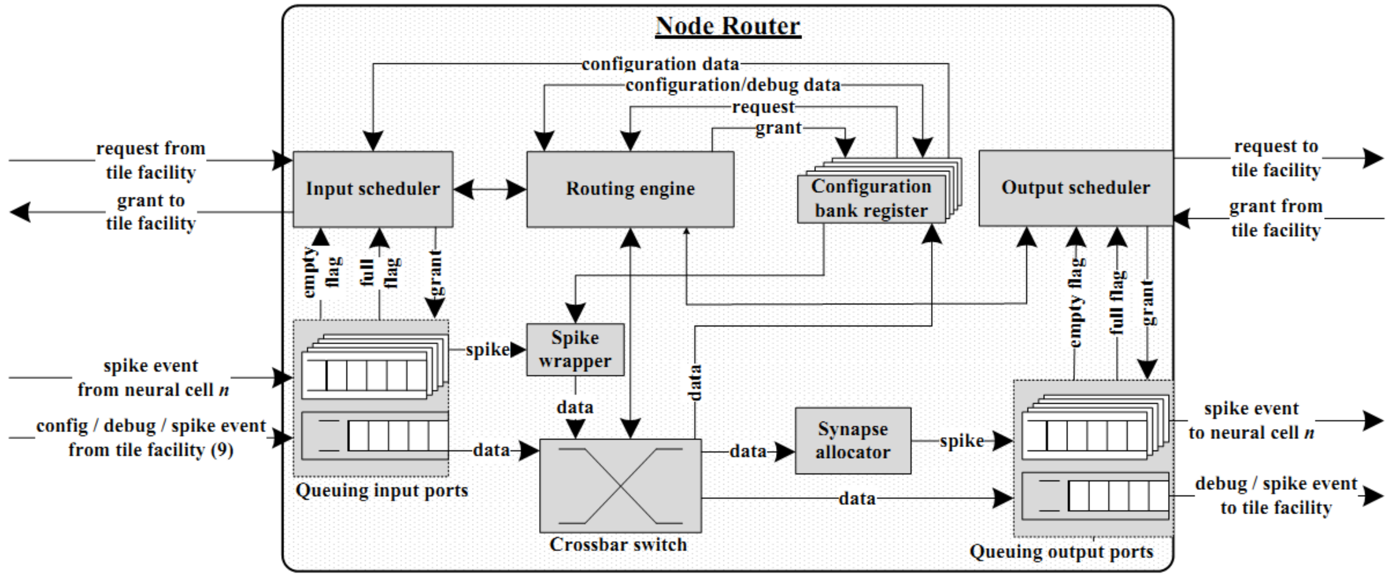
Carrillo, S., et al., "Advancing Interconnect Density for Spiking Neural Network Hardware Implementations using Traffic-Aware Adaptive Network-on-Chip Routers". Neural Networks, Vol 33, pp. 42-57, September 2012.

University of ULSTER

NUI Galway OÉ Gaillimh

Neuron Facility – @Bottom-Level





Layout for the configuration bank register

Neuron facility address register	Spike event timer register
Address range [3:0]	Scaling factor range [13:0]

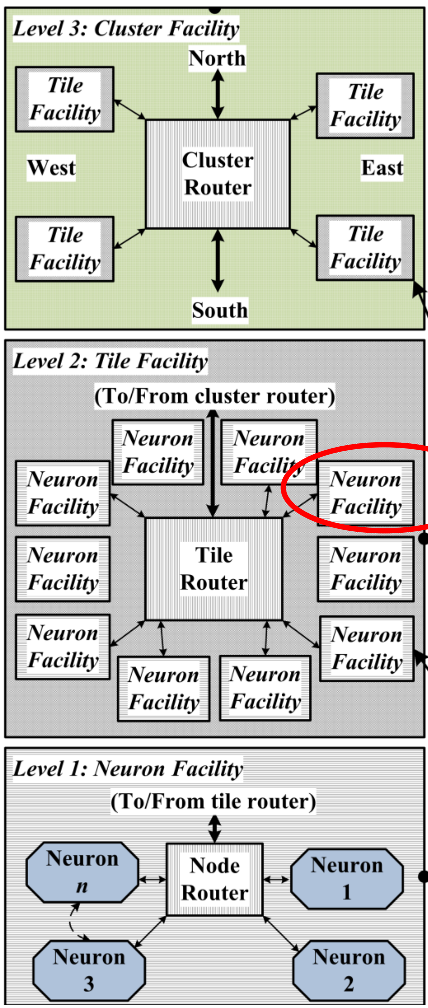
Communication protocol register	Spike event reception register
IT_CM/NB_MC/NB_BC/DB [3:0]	Neurons position within node [9:0]

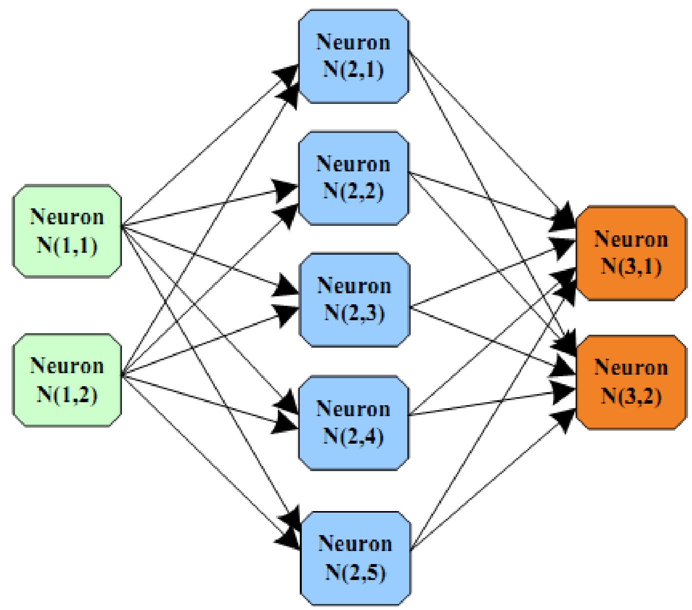
Target neuron population register

Target X-cluster addr [29:22]	Target Y-cluster addr [21:14]
Target tile addr [13:10]	Target node addr [9:0]

NB_MC: neighbour multicast DB: debug
NB_BC: neighbour broadcast IT_CM: internal communication

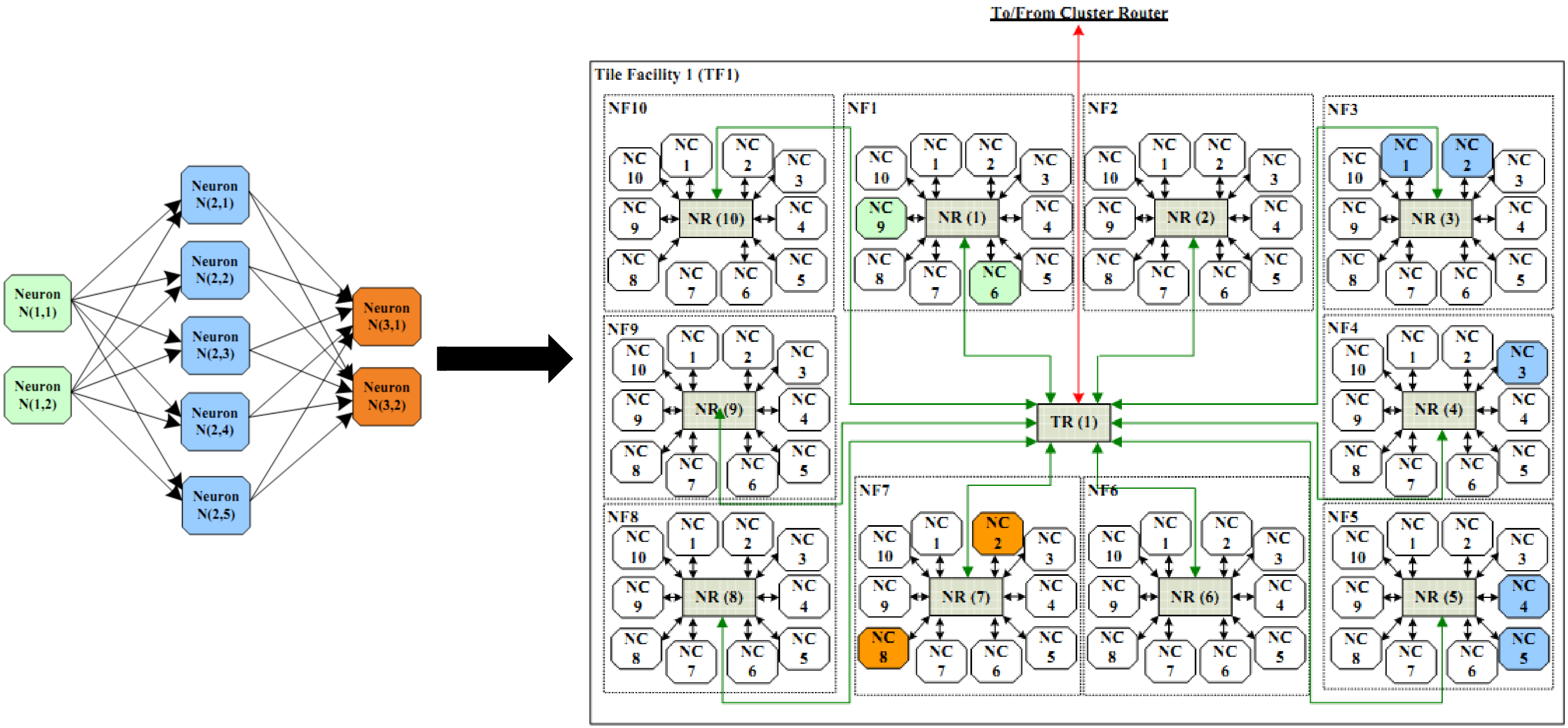
H-NoC Architecture: Example Scenario



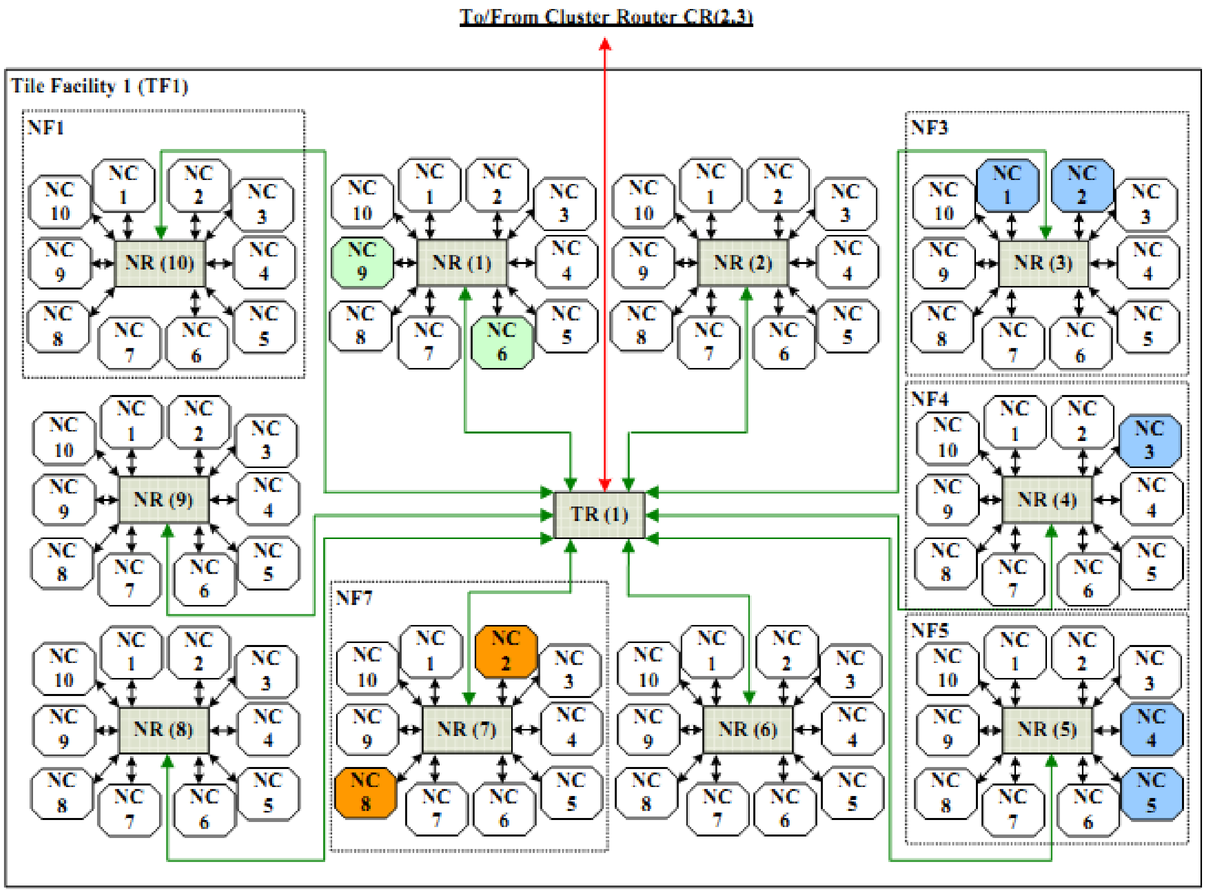


2x5x3 Feed-forward neural network

H-NoC Architecture: Example Scenario



H-NoC Architecture: Example Scenario



Node Router NR(1)

Address Register	01h	Comm. Protocol	0Fh	
Reception Register	000h	Timer Register	0001h	
Target Register	02h	03h	1h	01Ch

Node Router NR(3)

Address Register	03h	Comm. Protocol	0Fh	
Reception Register	003h	Timer Register	0001h	
Target Register	02h	03h	1h	040h

Node Router NR(4)

Address Register	04h	Comm. Protocol	0Fh	
Reception Register	004h	Timer Register	0001h	
Target Register	02h	03h	1h	040h

Node Router NR(5)

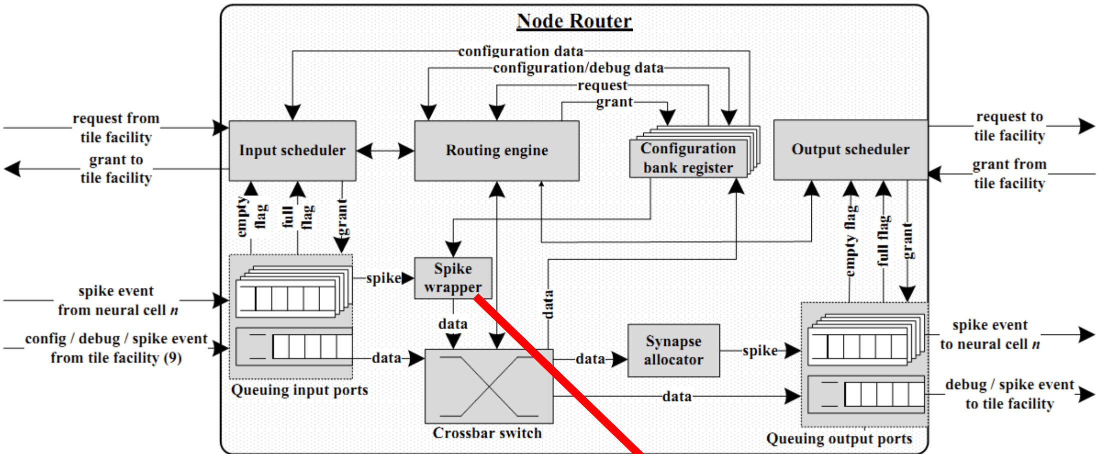
Address Register	05h	Comm. Protocol	0Fh	
Reception Register	018h	Timer Register	0001h	
Target Register	02h	03h	1h	040h

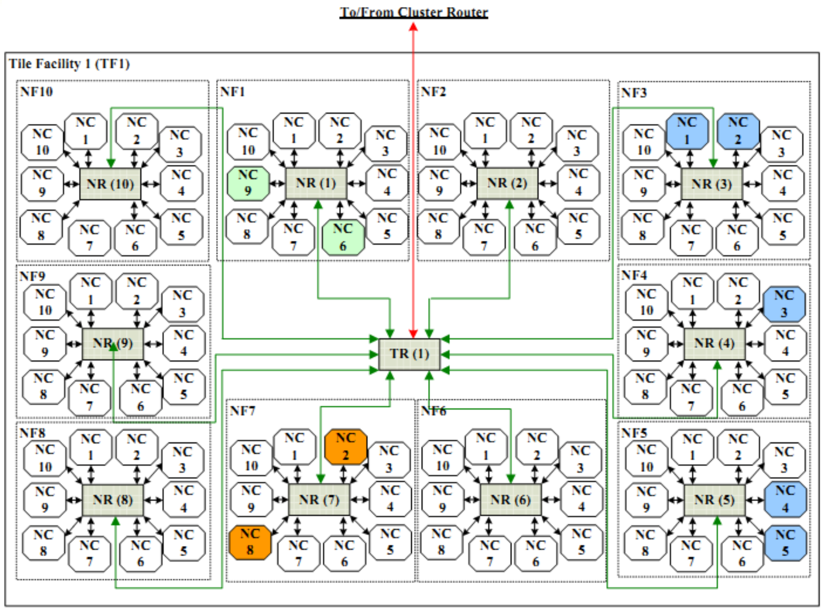
Node Router NR(7)

Address Register	07h	Comm. Protocol	0Fh	
Reception Register	082h	Timer Register	0001h	
Target Register	00h	00h	0h	000h

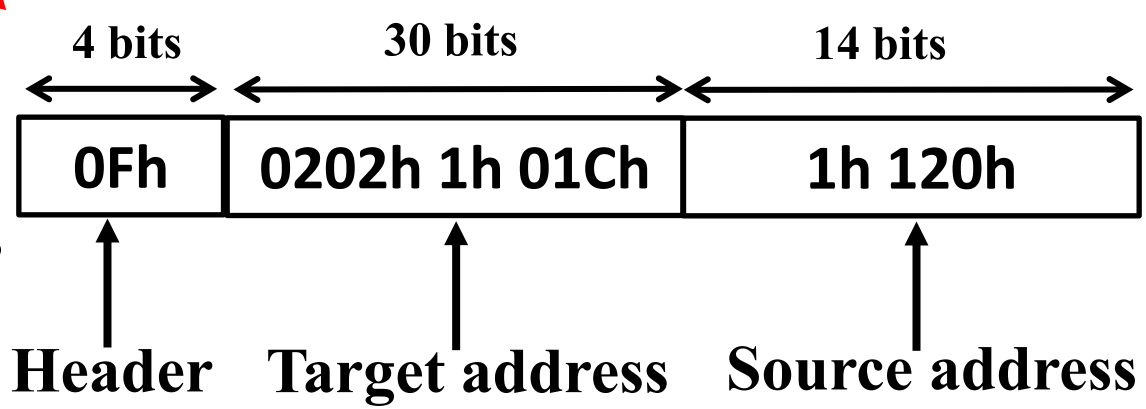
H-NoC Architecture: Example Scenario

On-chip Comm: Spike event generation

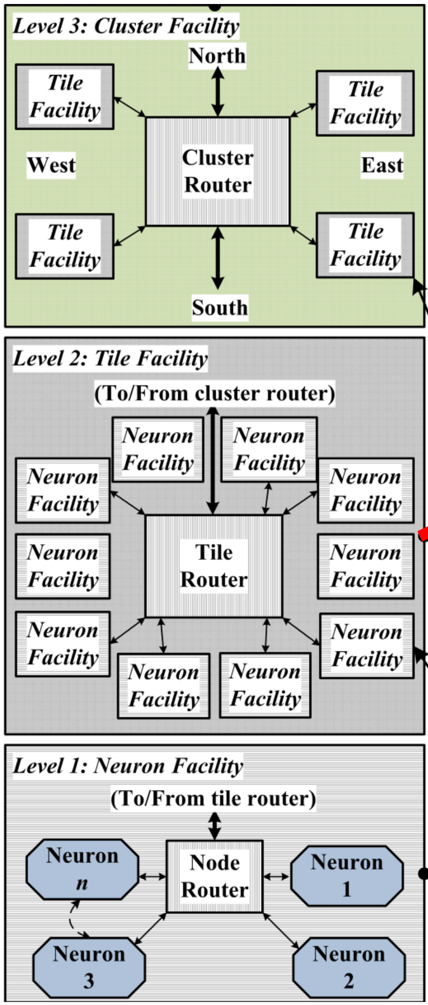


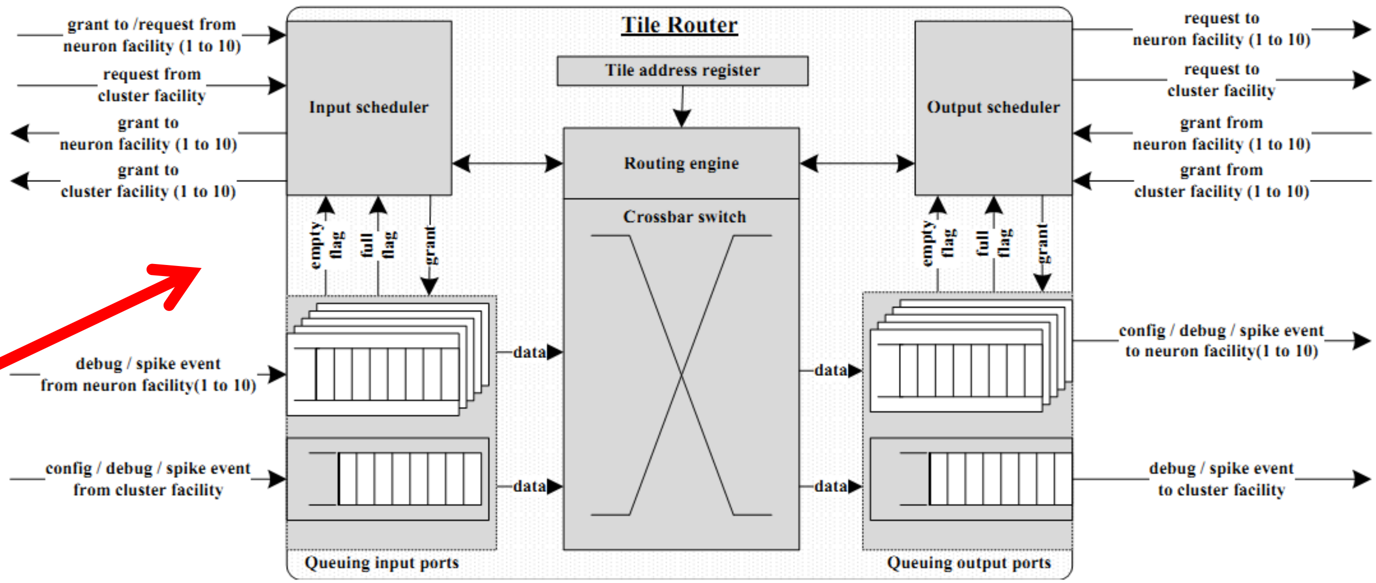


Packet generated when Input neurons #6 and # 9 are generating spike events



Tile Facility – @Mid-Level



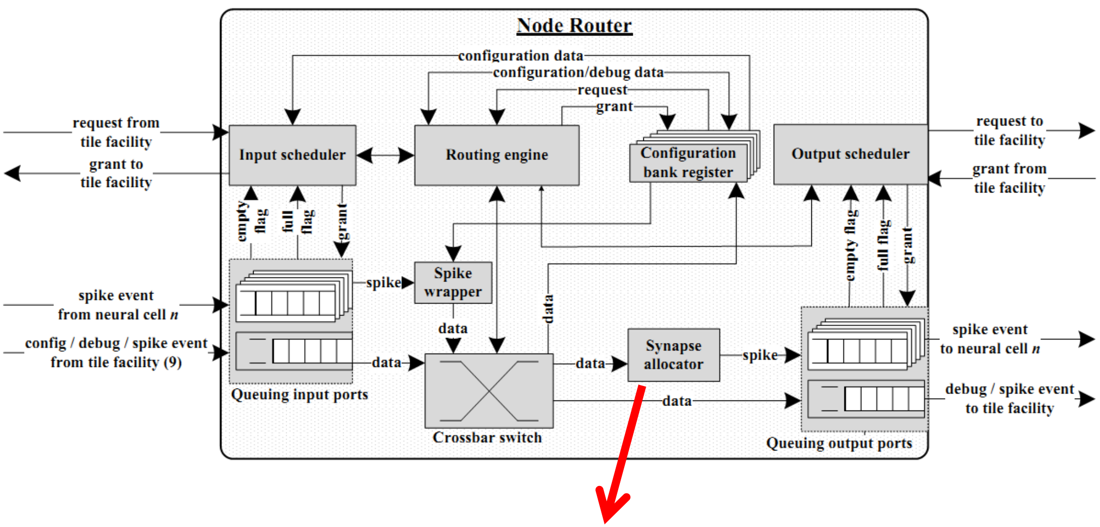


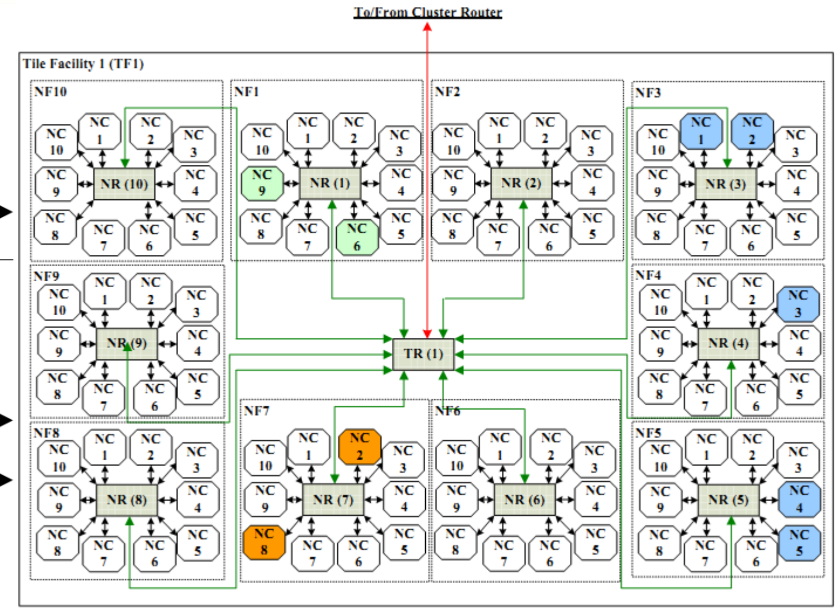
It's the arbitration point for NoC packets coming from the Bottom & Top Levels !!

Distributed parallel datapath to handle multiple incoming spike events !!

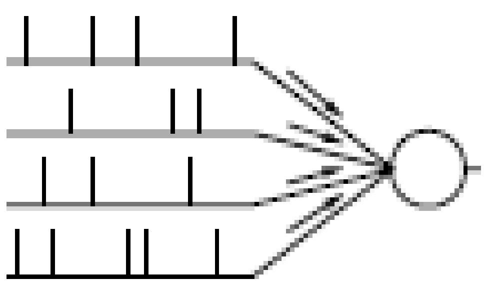
H-NoC Architecture: Example Scenario

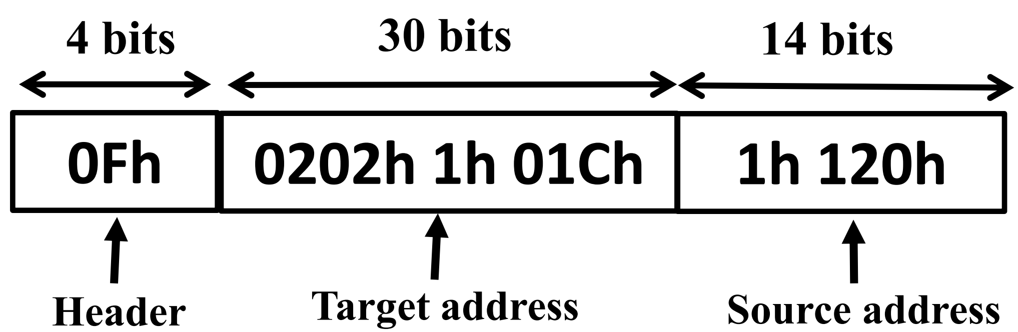
On-chip Comm: Spike event absorption





$$As = \eta - \left[A_{neuron} + \left(A_{neuron\_facility} \times \eta\right)\right]$$

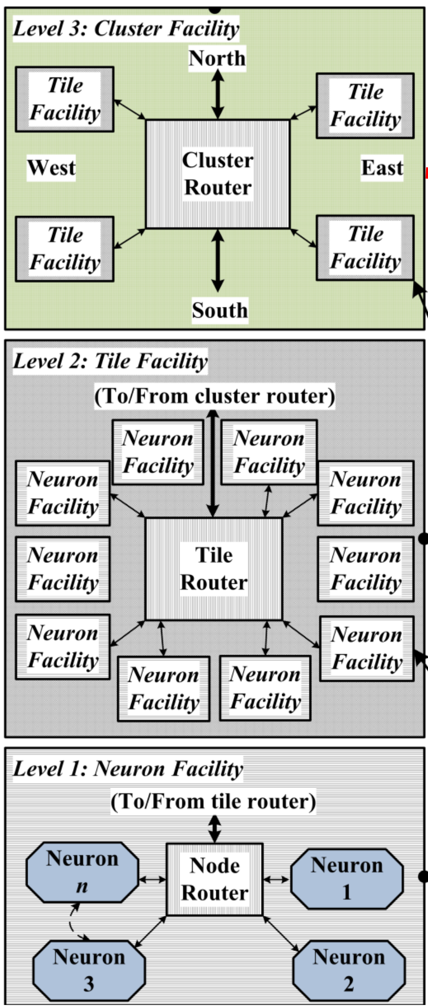


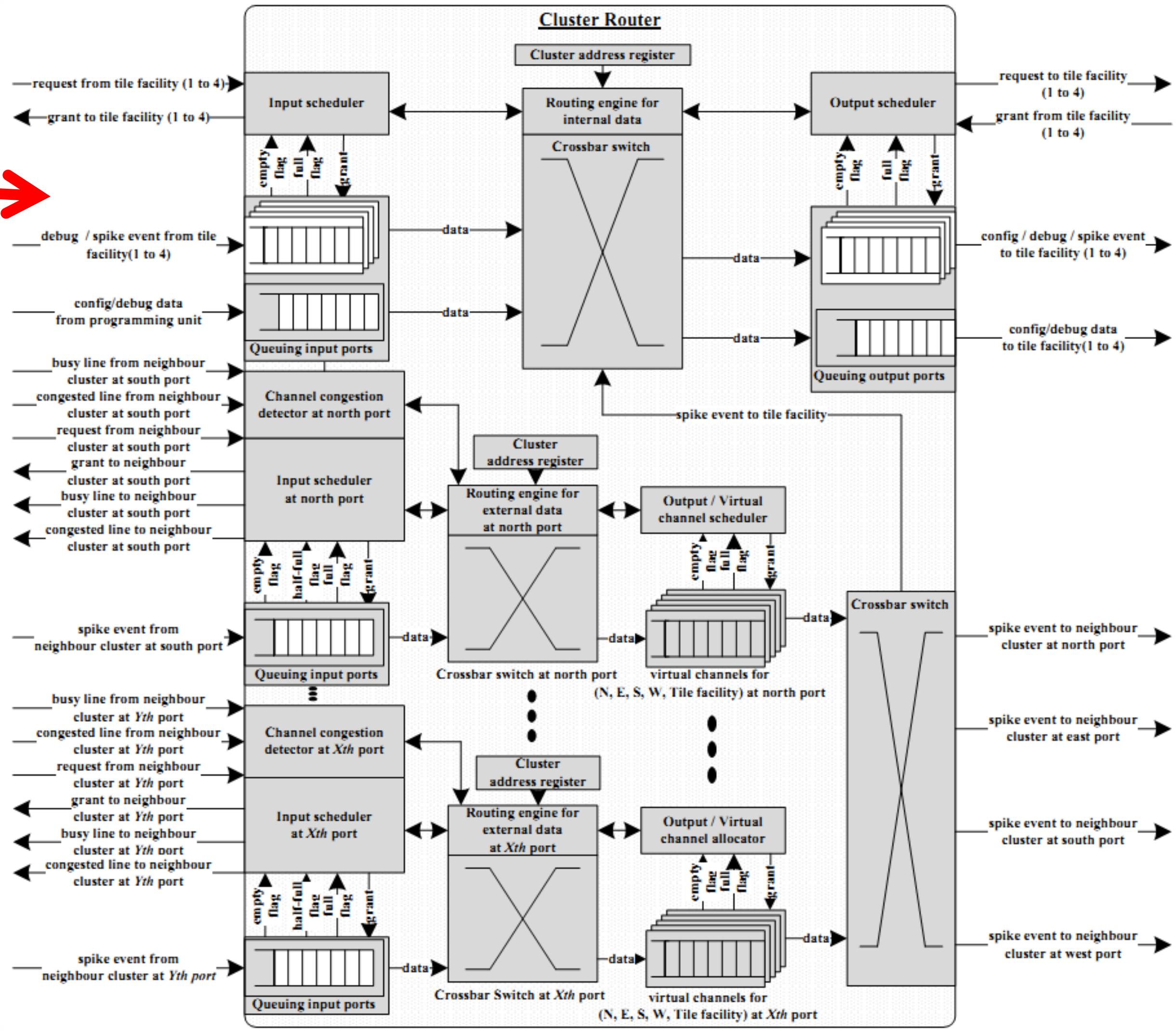


University of ULSTER

NUI Galway OÉ Gaillimh

Cluster Facility – @Top Level





On-chip Communication Protocols & Free Look-up Table Approach

Layout for the configuration bank register

Neuron facility address register

Address range [3:0]

Spike event timer register

Scaling factor range [13:0]

Communication protocol register

IT_CM/NB_MC/NB_BC/DB [3:0]

Spike event reception register

Neurons position within node [9:0]

Target neuron population register

Target X-cluster addr [29:22]	Target Y-cluster addr [21:14]
Target tile addr [13:10]	Target node addr [9:0]

NB_MC: neighbour multicast DB: debug
NB_BC: neighbour broadcast IT_CM: internal communication

Packet layout

Packet layout for the configuration mode

Header [47:44]	Target address* [43:30]	Configuration data [29:0]

Packet layout for the debug mode

Header [47:44]	Debug information [43:14]	Source address [13:0]

Packet layout for the execution mode

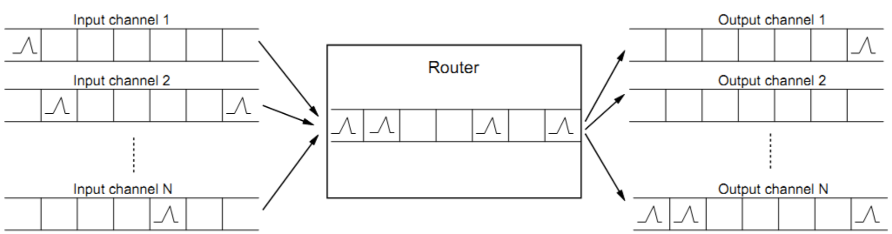
Header [47:44]	Target address [43:14]	Source address [13:0]

Target address field

Target X-cluster addr [43:36]	Target Y-cluster addr [35:28]
Target tile addr [27:24]	Target node addr [23:14]

Source address field

Node address [13:10]	Neurons position within node [9:0]



Header Information	Value
Cluster address register	0x00h
Tile address register	0x01h
Node address register	0x02h
Spike event timer	0x03h
Communication protocol register	0x04h
Spike event reception register	0x05h
Target neuron population register	0x06h
Broadcast between all neighbour clusters	0x07h
Multicast between selected neighbour clusters op1	0x08h
Multicast between selected neighbour clusters op2	0x09h - 0x0Ch
Internal communication	0X0F
Debug mode	0x0Dh - 0x0Eh

University of ULSTER

NUI Galway OÉ Gaillimh

On-chip Communication Protocols & Free Look-up Table Approach

ON-CHIP COMMUNICATION REGISTERS

Component	Memory Register Requirements [bit]					
	address	comm	Rx	Tx	timer	**total**
Cluster router	16	NA	NA	NA	NA	**16**
Tile router	20	NA	NA	NA	NA	**20**
Node router	4	4	10	30	14	**62**

(1 cluster router) **X** (16 bits)
+ (4 tile routers) **X** (20 bits)
(40 node routers) **X** (62 bits)

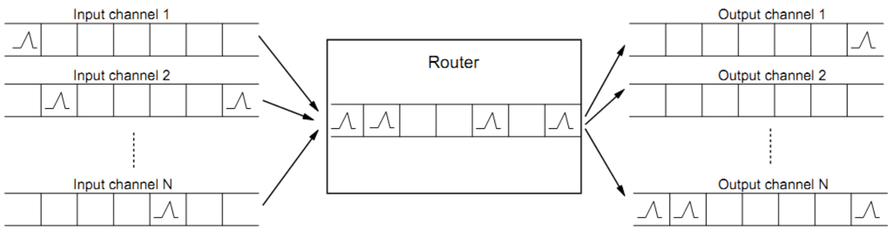
= **2.576Kbit (400 neurons)**

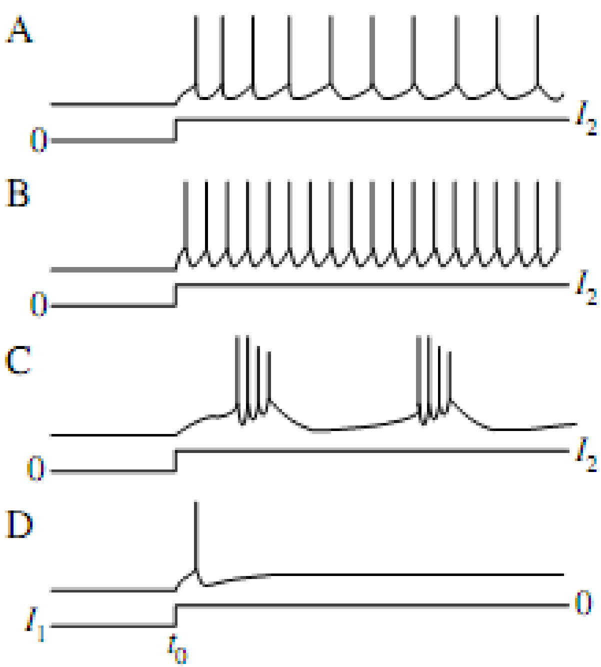
- **The implemented approach shows a very significant reduction in memory size.**
- **Previous work shows memory requirements in the order of Mbits !!**

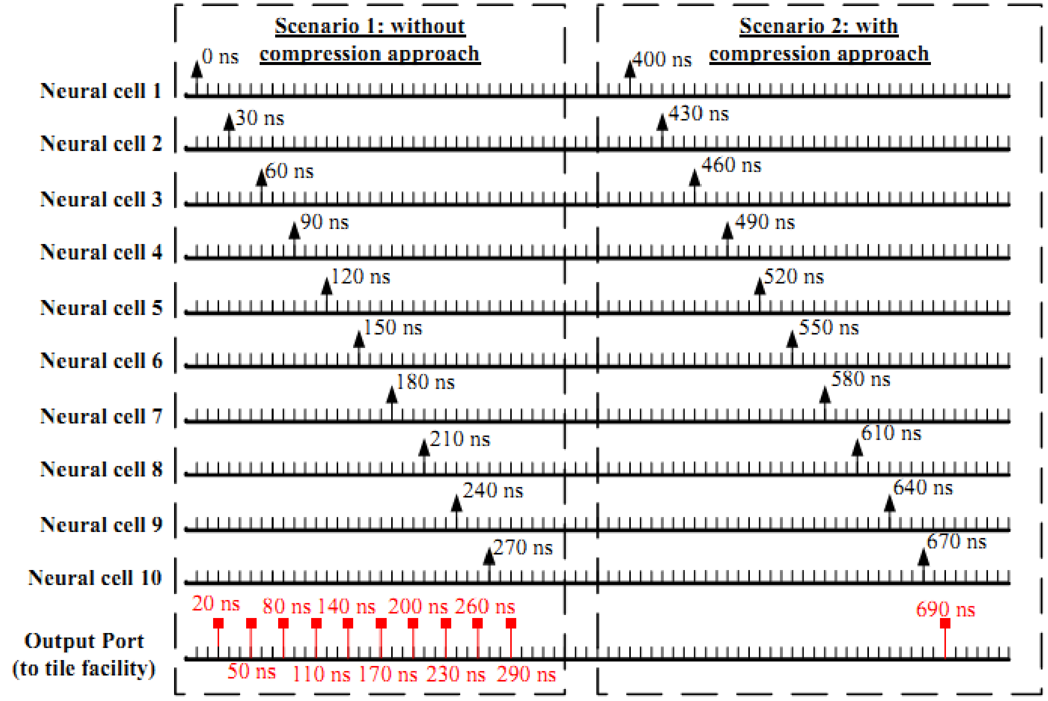
Spike Event Compression Technique

Motivation:

- **SNN traffic is slow (ISI > 1ms)**
- **Irregular pattern**
- **Polychronous Phenomena [Izhikevich'09]**
- **(i.e. More than 1 spike arriving at the same time)**







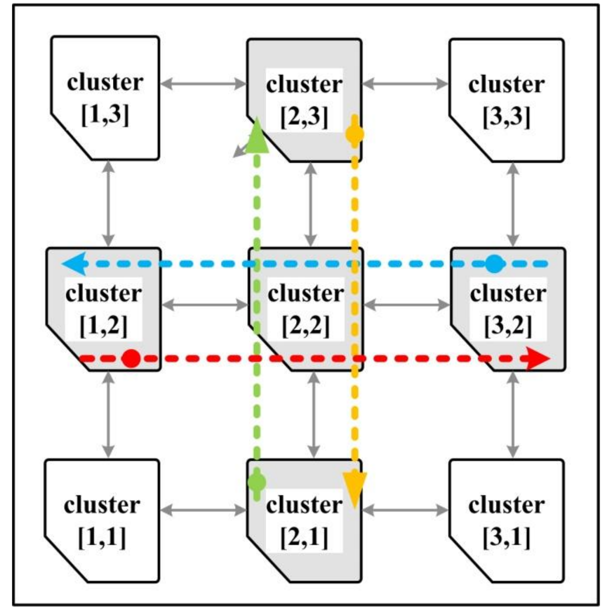
Outline

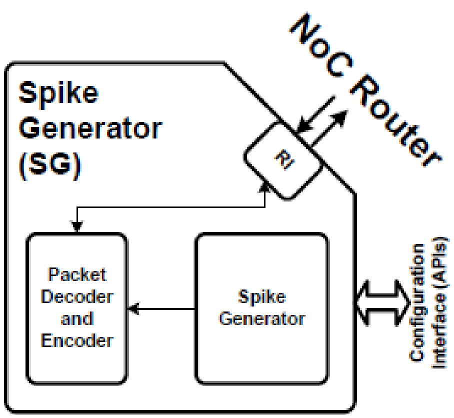
- Motivation and Challenges
- Hierarchical NoC EMBRACE Architecture
- **Performance Analysis**
- Take-home Message & Future Work

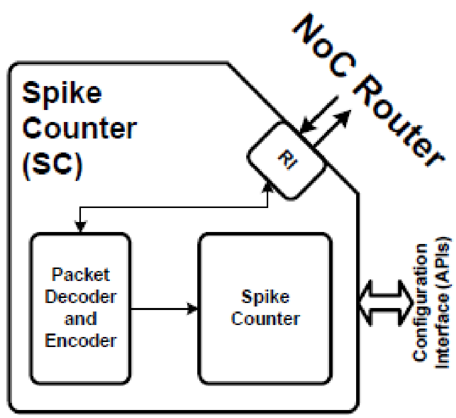
Experimental Setup

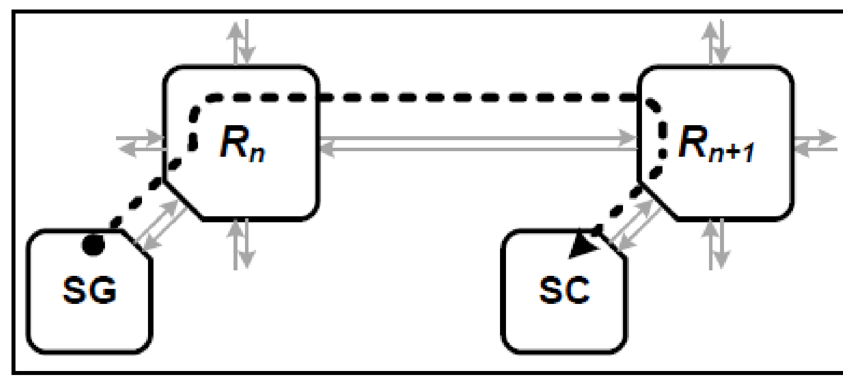
Methodology:

- **VHDL Simulation** of up to **50 x 50** array of clusters
- **FPGA implementation** of a 3x3 proof of concept array of clusters
- **100MHz** clock frequency per cluster & a **48-bits packet**
- **65-nm CMOS technology (estimated)**

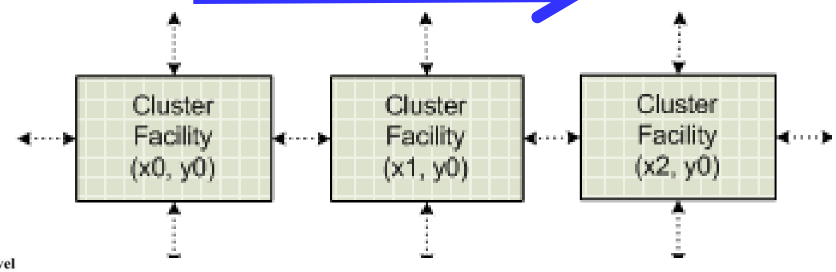


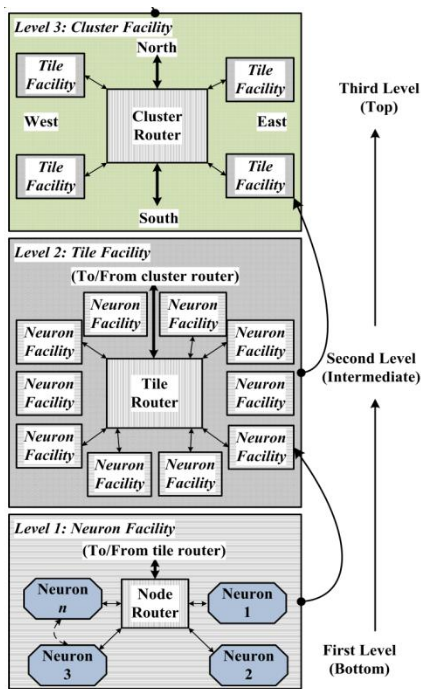






Traffic Load Analysis





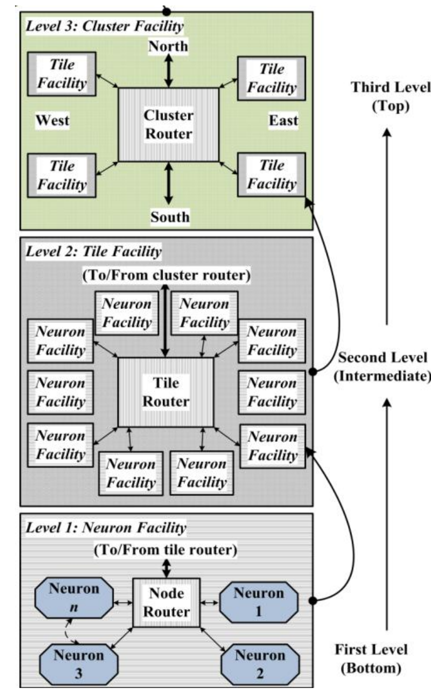
$$(N_c \times t_h)$$

$$N_c = n_{pg} = n_{pa}$$

t_h: 12 clock cycles

The total packet propagation delay:

$$\delta = T_{up} + T_{down} + (N_c \times t_h)$$



$$T_{up} = \tau_u + [t_u \times (n_{pg} - 1)]$$

τ_u : 24 clock cycles
t_u : 12 clock cycles

$$T_{down} = \tau_d + [t_d \times (n_{pa} - 1)]$$

τ_d : 30 cc
t_d : 1 cc

Traffic Load Analysis for Large Scale Scenarios

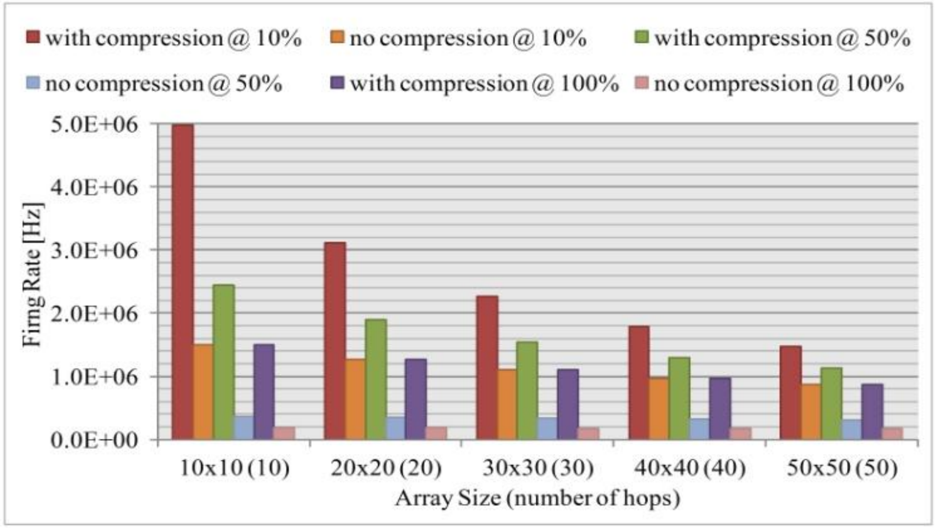
Typical biological spiking neurons show a **firing rate around 100 Hz**, but some others can show a firing rate **up to 1KHz**.

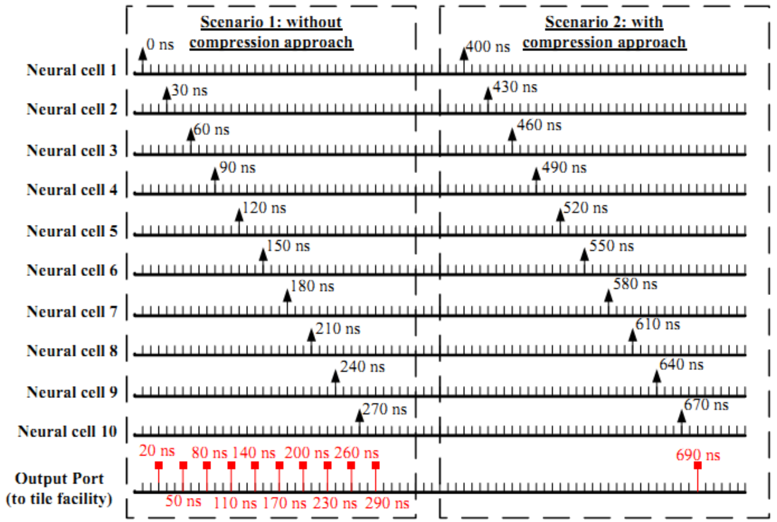
A maximum firing rate of ~5 MHz for a 10 hop scenario is highlighted **using the compression approach.**

This offers a ~3.3x improvement compared to the same scenario
without the compression technique.

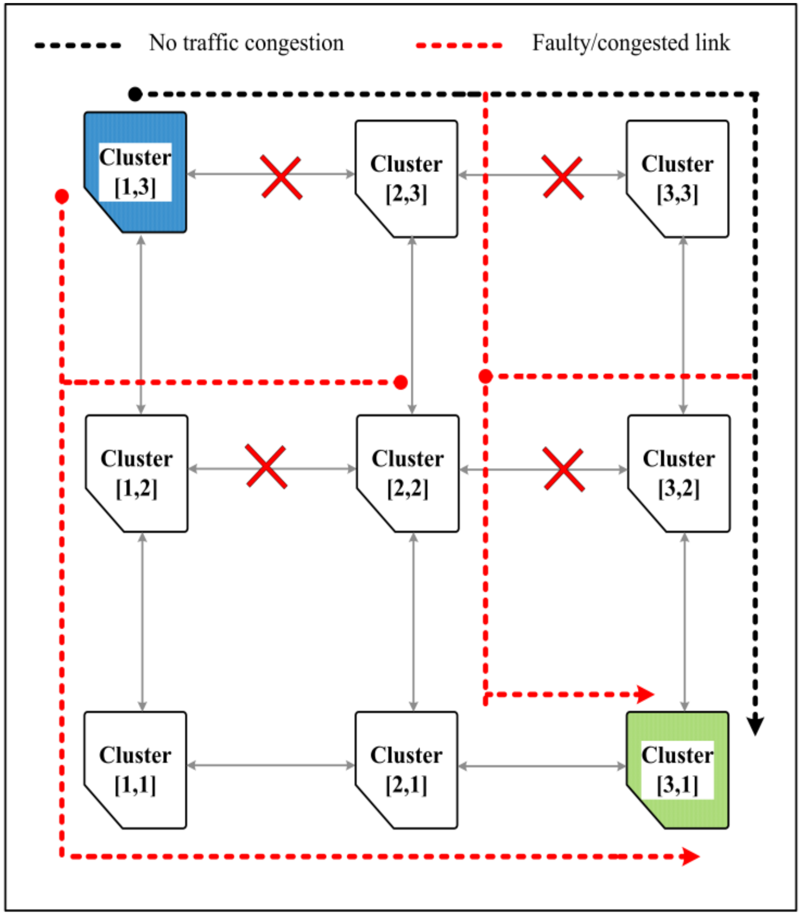
In the 50 hop scenario, although the firing rate can decrease to 172 KHz when the compression technique is not used,

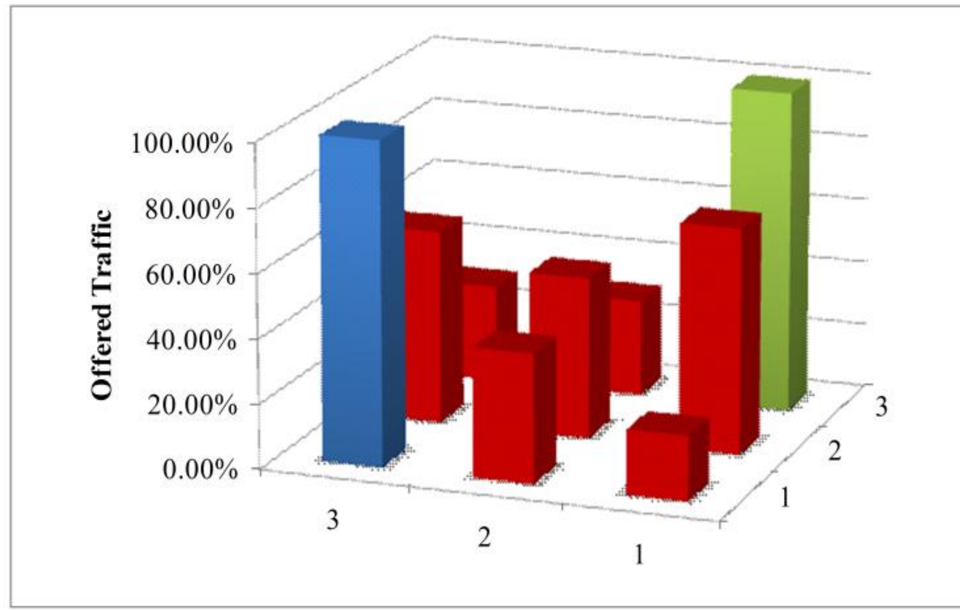
From a hardware point of view, if higher firing frequencies can be achieved, **the platform can be used as a neural network hardware accelerator.**





Adaptive Router Validation on FPGA



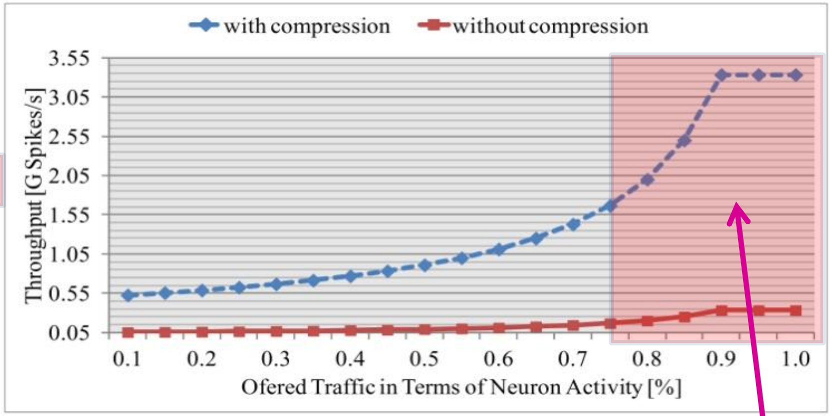


XY routing algorithm is used as a default routing mechanism when there is no traffic congestion

Throughput and Synthesis Results

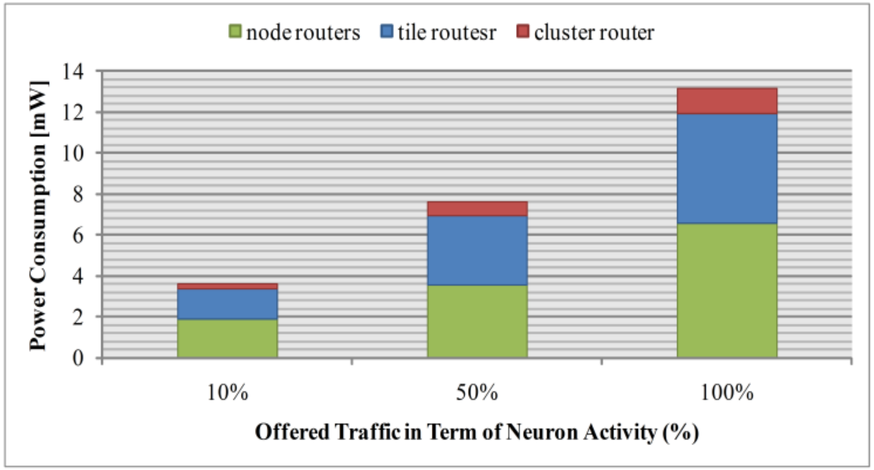
Area/power performance of router (65nm)

Component	Area [mm²]			Power [mW]		
	Comb	seq	total	dynamic	static	total
Cluster	0.139	0.448	0.587	10.62	2.54	13.16
Cluster router	0.005	0.017	0.022	1.10	0.09	1.19
Tile router	0.055	0.015	0.070	3.86	0.27	4.13
Node router	0.002	0.005	0.007	0.41	0.03	0.44



Increased throughput under load testing

Power Consumption vs. Offered Traffic (65nm)



Project Reference	Quality of Service (QoS)	Congestion Mechanism	Throughput [Events/s]	Improvement
EMBRACE	Best Effort	Yes	3.33×10^9	--
FACETS	Best/Guaranteed Effort	No	1.50×10^9	2.2x
SpiNNaker	Best Effort	No	0.200×10^9	16.5x

Proposed router outperforms existing approaches

University of ULSTER

NUI Galway OÉ Gaillimh

Outline

- **Motivation and Challenges**
- **Hierarchical NoC EMBRACE Architecture**
- **Performance Analysis**
- **Take-home Message & Future Work**

Take-home Message & Future Work

- There are **many problems** associated with the development of **efficient large scale SNN platform in hardware**.

- A **H-NoC approach** is proposed as a way to **overcome the intercommunication constrains** currently experienced in the efficient realisation **of SNNs in hardware**.

- **Future Work: Real-life SNN applications & Self-repair Mechanism** based on the information received from the adaptive routing algorithm.

Acknowledgments

Snaider Carrillo Lindado is supported by
a Vice-Chancellor's Research Scholarship **(VCRS)**
from the University of Ulster

Thanks for your attention
Any question/feedback welcome

Snaider Carrillo

Email: carrillo_lindado-s@email.ulster.ac.uk

http://isrc.ulster.ac.uk/Staff/SCarrillo/Contact.html