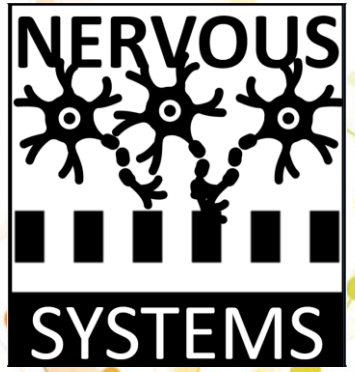
UNIVERSITY of York

Ulster University

UKRI Engineering and Physical Sciences Research Council



Applications of SNNs - Networks for Fault-tolerance

Martin Trefzer, Andy Tyrrell, Andrew Walter & Shimeng Wu
School of Physics, Engineering & Technology
University of York

Jim Harkin, Liam McDaid, Malachy McElholm & Thandassery Nidhin
School of Computing, Engineering & Intelligent Systems
Ulster University

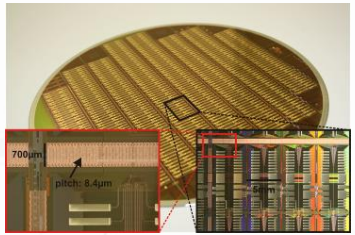
jg.harkin@ulster.ac.uk

XILINX

ThalesAlenia Space
a Thales / Leonardo company

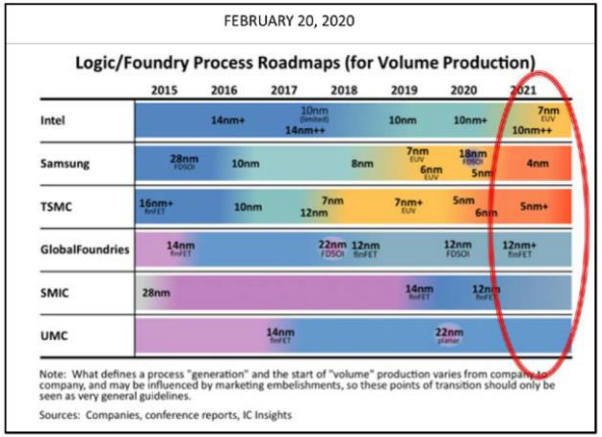
arm

Motivation



- **Reliability** is a significant challenge for modern **electronic systems**.
- Increased physical defects in advanced silicon manufacturing processes; wear-out faults etc. Permanent, Temp. (SEU, Electromagnetic Interference)

(2021) Apple M1 Pro (33 billion) and M1 Max (57 billion)



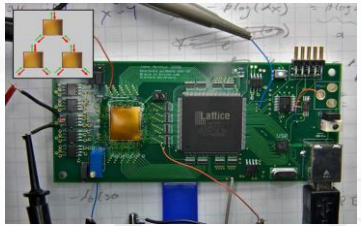
IEEE International Roadmap for Devices and Systems (IRDS), 2020

*"With device feature sizes projected to decrease to **less than 5 nm** within the next 10 years, scaling as we know it is expected to soon reach its physical limits or get to a point where cost and **reliability issues** far outweigh the benefits."*

Now 2nm and below

IEEE International Roadmap for Devices and Systems (IRDS), 2020

Self-X



Self-X
- self-monitor
- self-detect
- self-repair

- Traditional approaches: redundancy/replication models, error correction techniques, radiation hardening, Evolutionary/reconfigurable.

- Limited levels of reliability – constraints on:
  - **number of faults** that can be tolerated (degree)
  - **level of granularity** with which repairs can be implemented
  - Often a **central repair mechanism** not distributed, therefore fault-prone

Operation of Fault Tolerant Systems

• There are several key areas in the development of fault tolerant hardware computing systems:

• Fault monitoring/detecting

- be able to detect a fault has occurred

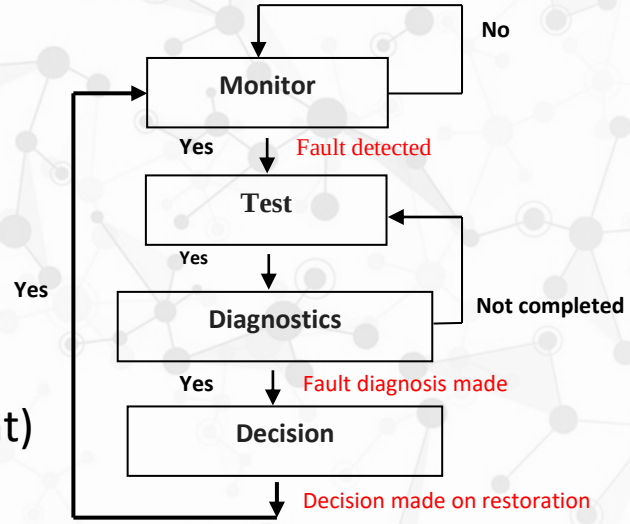
• Fault test and diagnostics

- where the fault is located

- the type of fault (transient, temporary or permanent)

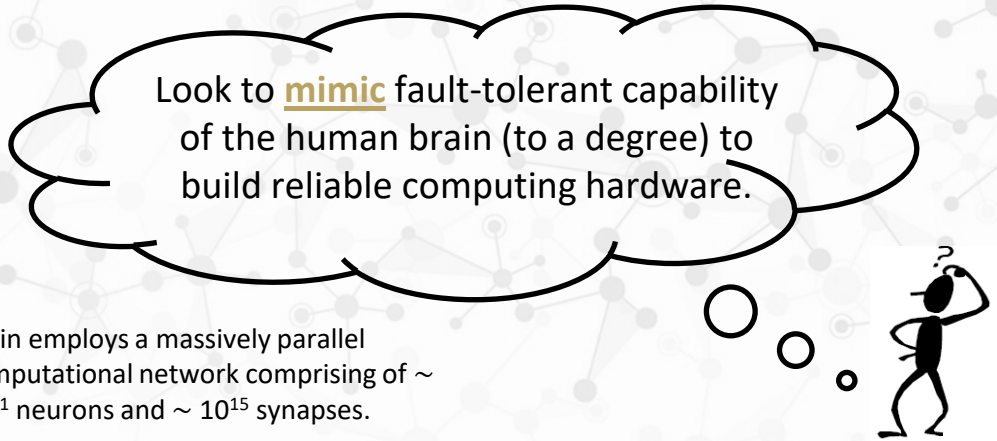
• Fault tolerant decision

- what can be done to restore the system's operation, i.e. how can the fault be tolerated.

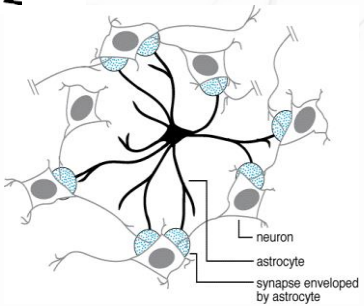


‘Reliability’: Unmet Need

- We can learn a lot from biology, in particular neuroscience!
- Brain processing : robust and power-efficient information computing.



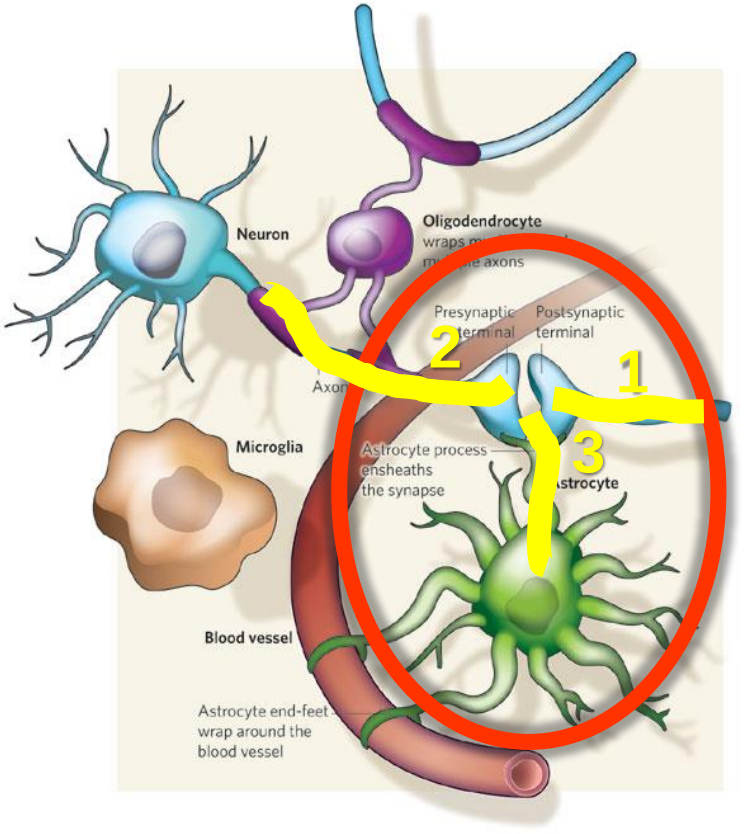
- Exploit the brain’s self-repair mechanism (**astrocyte** cells)
- Aim to develop **astrocyte-neuron networks**…….
  “Self-rePAiring spiking Neuron NEtwoRk” (SPANNER)



More Knowledge on Brain Repair

- Astrocyte enwraps many (~10^5) synapses and can connect to multiple (~6-8) neighbouring neurons.
- The connection between the astrocyte and neurons is named the ***tripartite synapse***.
- When an action potential (AP) arrives at the presynaptic axon how do we describe the interactions between the neurons, synapses and astrocyte?

Astrocyte-Neuron (AN) Model



Nicola J. Allen & Ben A. Barres, "***Glia — more than just brain glue***," Nature, vol. 457, 675-677, 2009. doi:10.1038/457675a

Software Model (AN) - A tripartite synapse story

Action potential (Spike) - presynaptic axon

Neuro-transmitter (Glutamate) is released across the cleft and binds to the receptors of the postsynaptic terminal.

After depolarization of postsynaptic neuron, a type of endocannabinoid (2-AG) is synthesized and released from dendrite. 2-AG feeds back to the *pre-synaptic terminal* in **two ways**:

Directly

2-AG binds directly to the type 1 Cannabinoid Receptors (CB1Rs) of the presynaptic terminal.

This causes a decrease of transmission probability rate (**PR**) of the synapses, and is termed Depolarization-induced **Suppression** of Excitation (DSE).

Indirectly

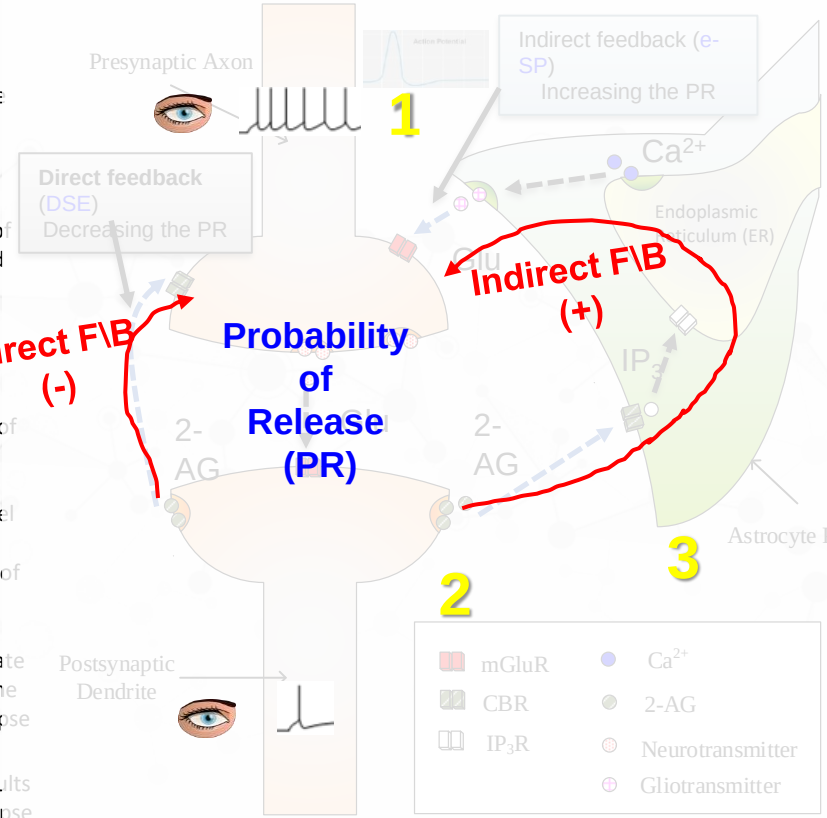
2-AG binds to CB1Rs of the astrocyte cell.

Increasing the IP_3 level

Triggers the release of calcium (Ca^{2+})

Releases the glutamate (Glu) and binds to the receptors in the synapse

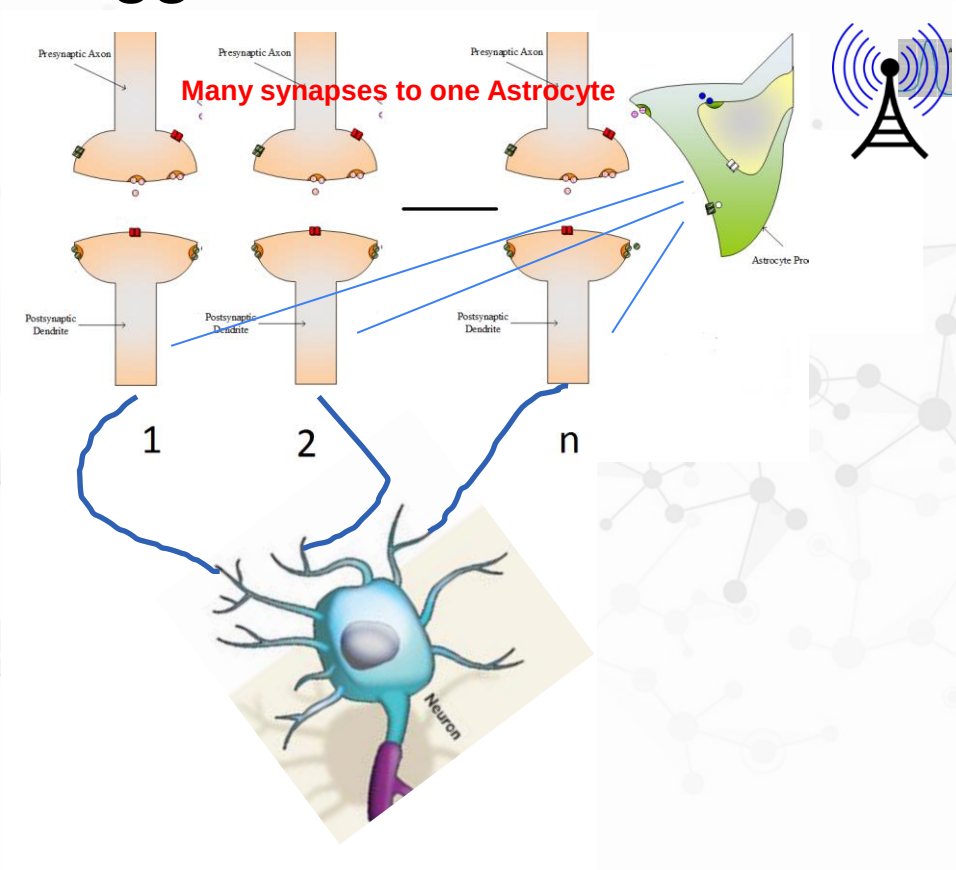
Termed e-SP, this results in an **increase** of synapse probability of release (PR).



New Probability of Release!

J. Wade, L. McDaid, J. Harkin, *et. Al* Frontiers in Computational Neuroscience, 6(76), pp. 1–12, 2012.

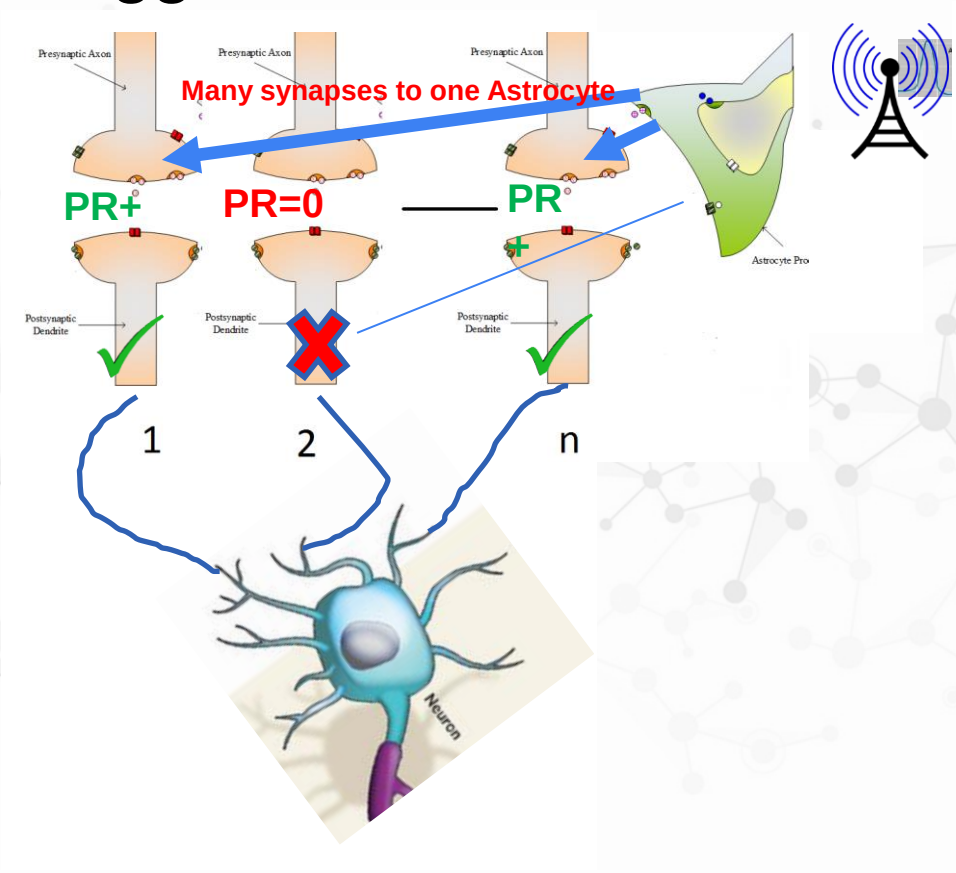
The Bigger Picture - "*Astro-Neuron Network*"



(Calcium signal broadcast) IP_3

J. Wade, L. McDaid, J. Harkin, *et. Al* Frontiers in Computational Neuroscience, 6(76), pp. 1–12, 2012.

The Bigger Picture - "*Astro-Neuron Network*"



(Calcium signal broadcast) IP_3

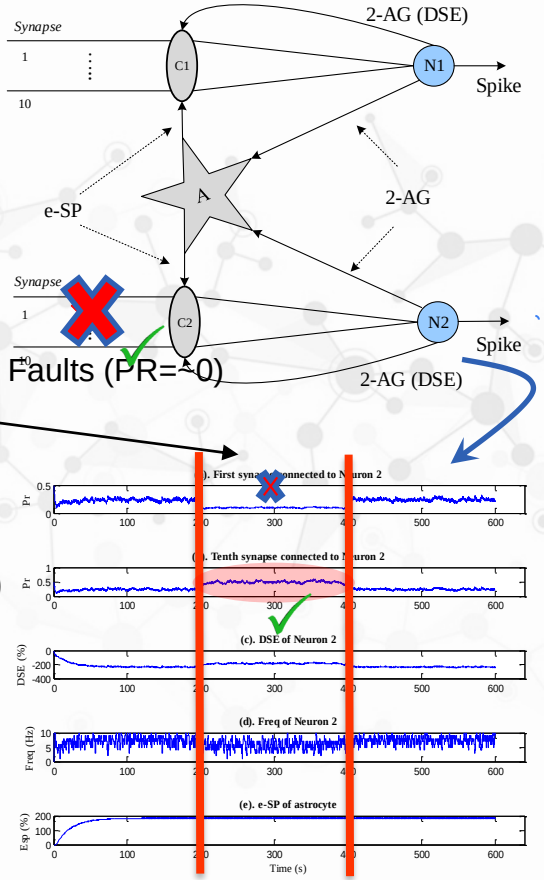
J. Wade, L. McDaid, J. Harkin, *et. Al* Frontiers in Computational Neuroscience, 6(76), pp. 1–12, 2012.

Self-repair of a ‘Small’ Astrocyte-Neuron Network

Temporary Faults injected

PR of Synapse #1

PR of Synapse #10

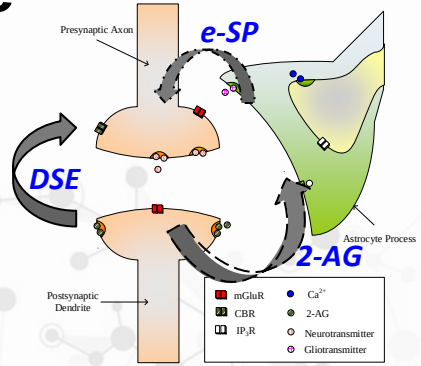


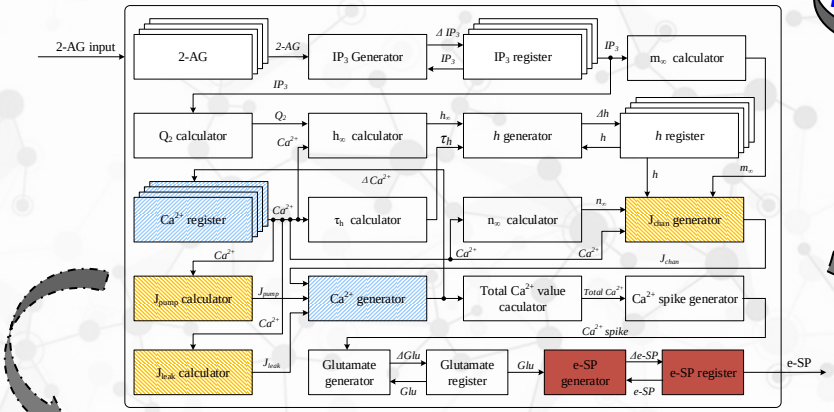
Neuron #2 under **80% fault rate** with temporary faults. (*80% - severely damaged*)

Naeem M, McDaid L, Harkin J, Wade J, et al. IEEE Transactions Neural Networks & Learning Systems, 2015.

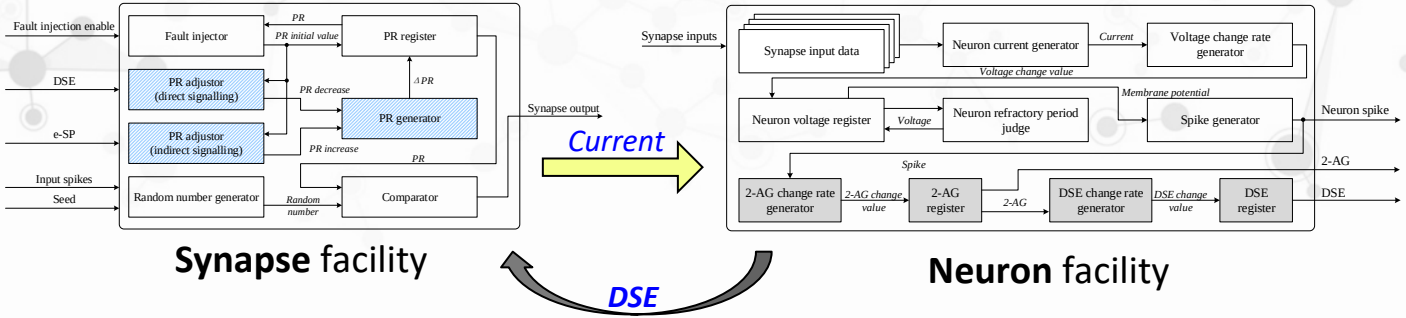
Moving to Hardware

- Explored the mapping to hardware as it enables repairs to achieved when the underlying hardware is unreliable.





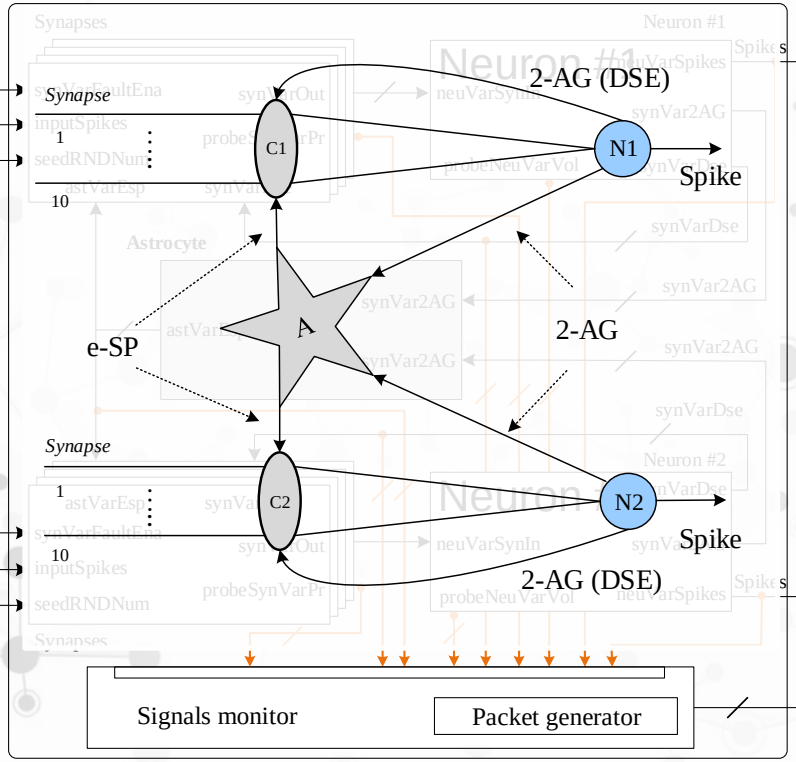
Astrocyte facility



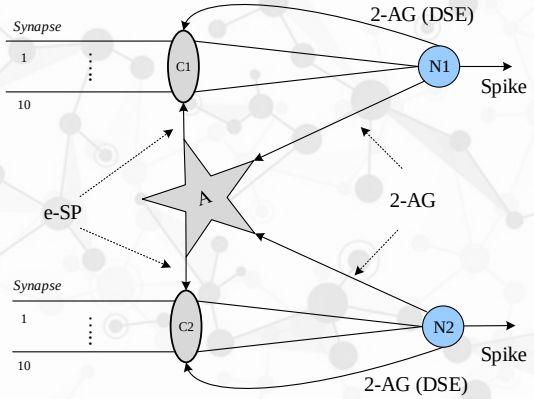
Synapse facility

Neuron facility

FPGA Hardware Implementation



- Astrocyte-neuron network with 2 neurons (N1, N2), 20 synapses (C1, C2), and 1 astrocyte cell (A).



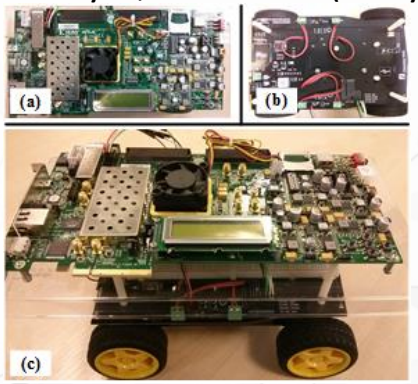
Xilinx VC707 development board

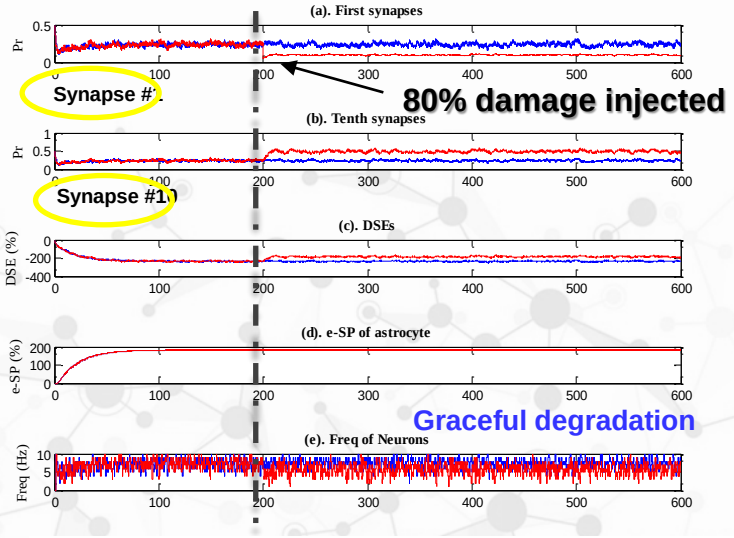
- Xilinx Virtex-7 XC7VX485T
- Vivado High Level Synthesis tool to generate the IP blocks

Ulster University

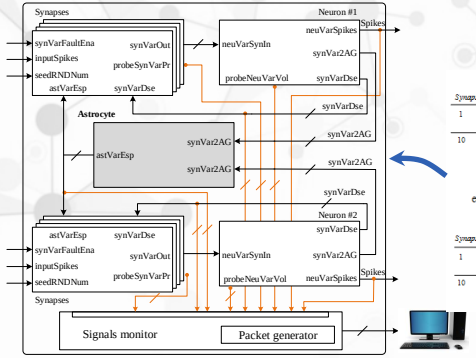
‘Small’ Astrocyte-Neuron Network to FPGAs

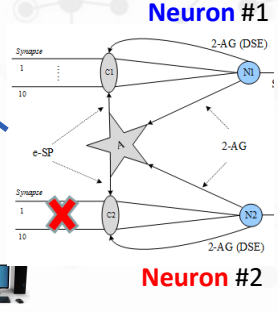
- 1 astrocyte, 2 neurons (20 synapse)



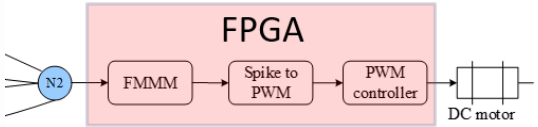


Xilinx Virtex-7 XC7VX485T



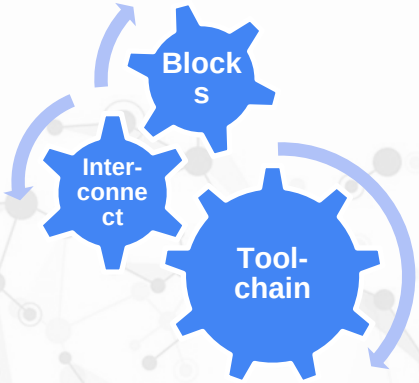


Fault density	Platform	Average output frequency	
		Neuron 1	Neuron 2
0%	Simulation	7.19	7.20
	Hardware	**7.28**	**7.27**
40%	Simulation	7.38	6.81
	Hardware	**7.37**	**6.88**
80%	Simulation	7.38	5.68
	Hardware	**7.37**	**5.75**



Liu J, Harkin J, Maguire LP, et al.; IEEE International Symposium on Circuits and Systems, 2016

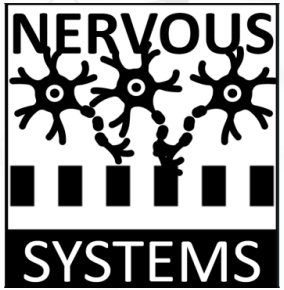
Challenges



- **Model real applications:** Larger networks.
- **Key functional blocks**: Astrocyte, tri-partitie synapse and learning rule (trade-off computational complexity for key principle with area/power efficiency)
- **On-chip interconnect**: High levels of connections, different time scales, different data – spike events, numeric data values (increased interconnect problem due to astrocyte connections – scalability solutions required)
- **Tool-chain**: Tools to map application to network paradigm, program the hardware, fault injection, data analysis and visualisation.

Summary

- Opportunities to fault tolerance not just from the parallelization of partial computations across multiple synapses and neurons.
- Challenges remain in classification of faults and mitigations and **system level** approaches to bio-inspired fault tolerance.



jg.harkin@ulster.ac.uk