Ulster University

INTELLIGENT SYSTEMS RESEARCH CENTRE

Computational Neuroscience, Neurotechnology and Neuro-inspired Artificial Intelligence (ISRC-CN3), 24th-28th Oct, 2022

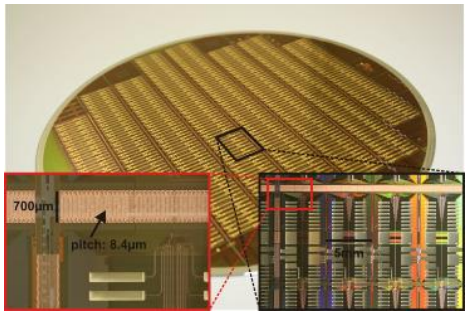
Building Reliable Embedded Systems with Neuromorphic Computing

Jim Harkin

Intelligent Systems Research Centre
Ulster University, Magee Campus, Co. Derry

Content

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)

Logic/Foundry Process Roadmaps (for Volume Production)

	2016	2017	2018	2019	2020	2021	2022
Intel	14nm+	10nm (limited) 14nm++		10nm	10nm+	10nm++	7nm EUV
Samsung	10nm		8nm	7nm EUV 6nm EUV	18nm FDSOI 5nm	4nm	3nm GAA
TSMC	10nm	7nm 12nm		7nm+ EUV	5nm 6nm	5nm+	4nm 3nm
GlobalFoundries			22nm FDSOI 12nm finFET		12nm FDSOI	22nm+ FDSOI 12nm+ finFET	
SMIC				14nm finFET	12nm finFET		8-10nm finFET
UMC		14nm finFET			22nm planar		

Note: What defines a process "generation" and the start of "volume" production varies from company to company, and may be influenced by marketing embelishments, so these points of transition should only be seen as very general guidelines.

Sources: Companies, conference reports and IC Insights

IEEE International Roadmap for Devices and Systems (IRDS), 2022

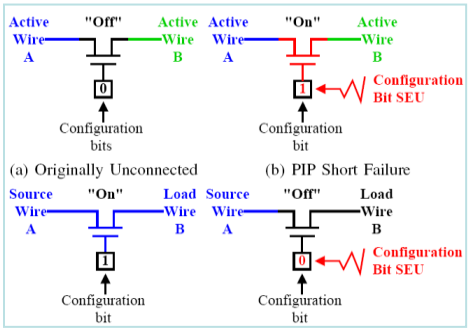
*"With device feature sizes projected to decrease to **less than 5 nm** within the next 10 years, scaling as we know it is expected to soon reach its physical limits or get to a point where cost and **reliability issues** far outweigh the benefits."*

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

2022: Apple M1 Ultra (114 billion) @5nm
2021: Apple M1 Max (57 billion)

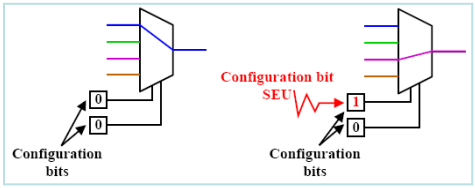
Electronic Faults

- Faults can be **transient** (disappear after a relatively short time), **intermittent** (cycle between events) or **permanent** (always present).



Transistor charge perturbed

MEAN time between failures (MTBF)



Multiplexer select lines can get corrupted

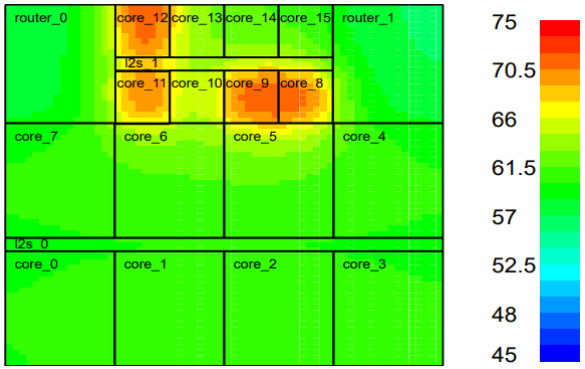
- Shrinking feature size
- Increasing number of faults manifesting post fabrication
- Wear-out faults
- Failure due to Aging
- Need for fault-tolerance or **self-correcting strategies.**

Thermal overstress

Excess heat— melts materials, warps and breaks semiconductor dies.

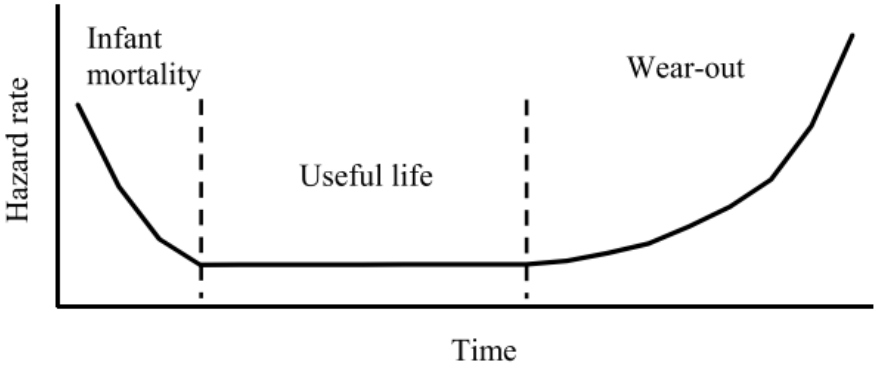
Electrical overstress

High voltage transient at base can damage transistor creating a base-emitter short circuit.

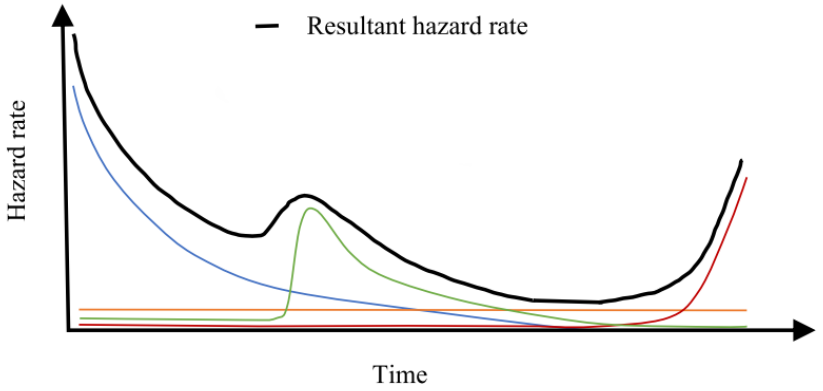


M. Ansari *et al.*, *IEEE Transactions on Emerging Topics in Computing* 2021

Electronic Faults

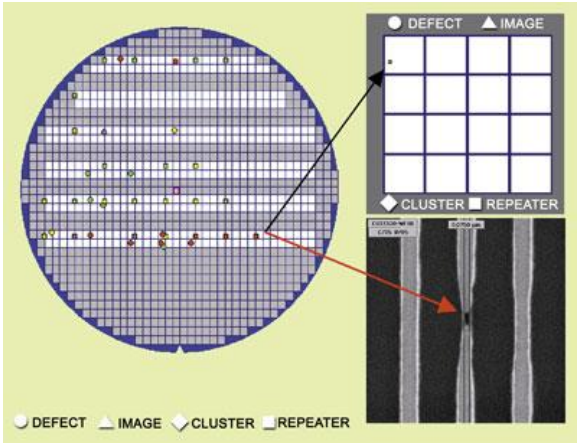


- Bathtub model of hazard rate



- Roller-coaster curve of the hazard rate

A. Gaonkar, et al., "*An Assessment of Validity of the Bathtub Model Hazard Rate Trends in Electronics*," in IEEE Access, vol. 9, pp. 10282-10290, 2021

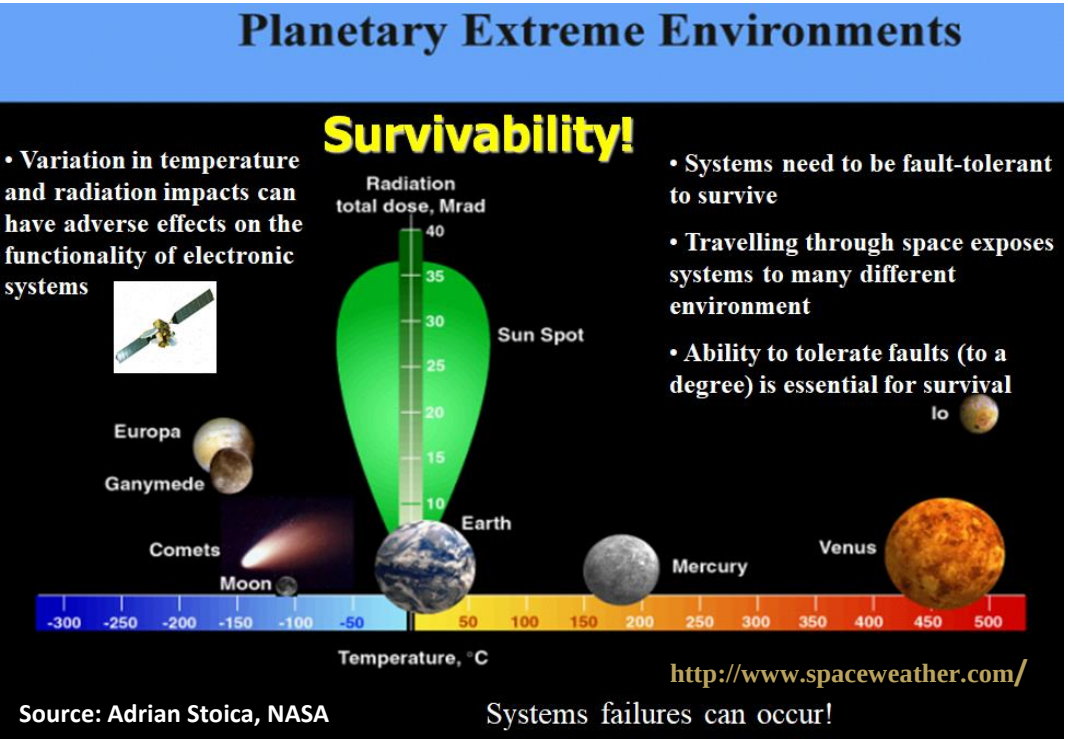


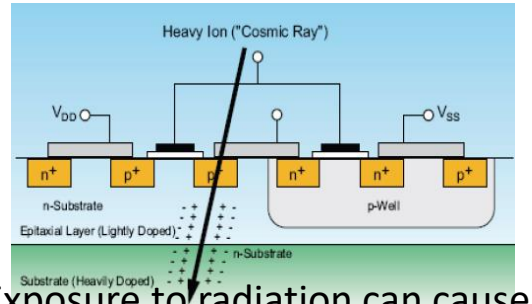
Electronic Faults

- Bathtub model of hazard rate

- Roller-coaster curve of the hazard rate

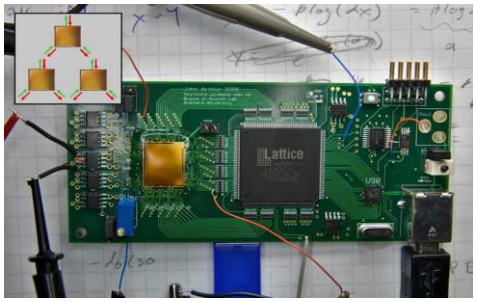
A. Gaonkar, et al., "*An Assessment of Validity of the Bathtub Model Hazard Rate Trends in Electronics*," in IEEE Access, vol. 9, pp. 10282-10290, 2021.





Exposure to radiation can cause SEU

Self-X



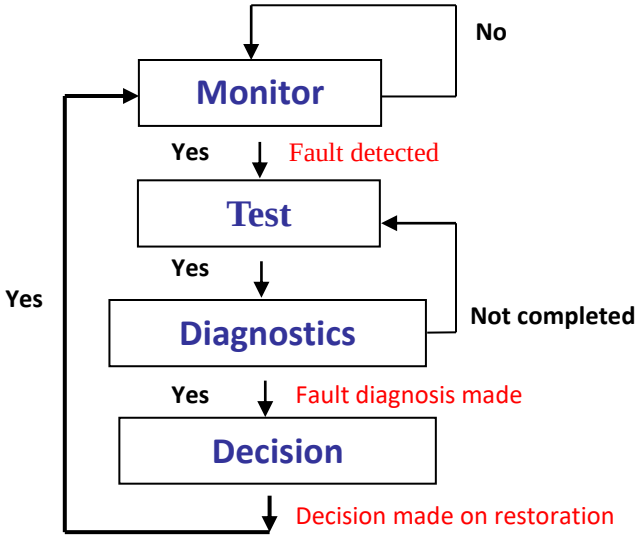
Self-X
- self-monitor
- self-detect
- self-repair

- 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)

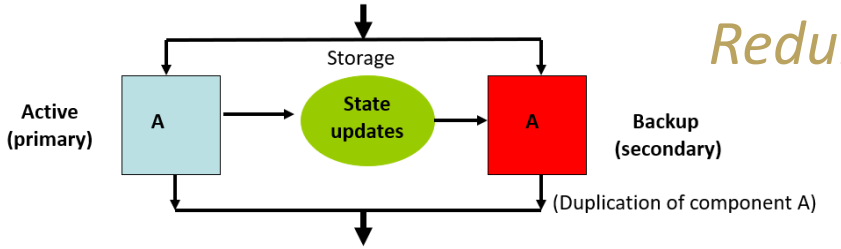
• Fault tolerant decision

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

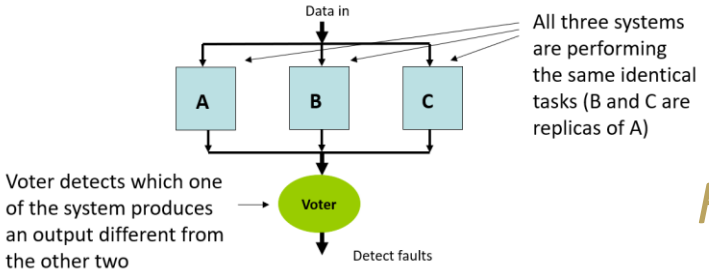
Mitigating Hardware Faults

- There are several further strategies to mitigating faults:
  - Fault tolerant *hardening*
  - Fault tolerant *redundancy*
  - Fault tolerant *replication*

- Three main *redundancy* techniques used to creating fault-tolerant (FT) systems:
  - Hot Standby
  - Cold Standby
  - Warm Standby



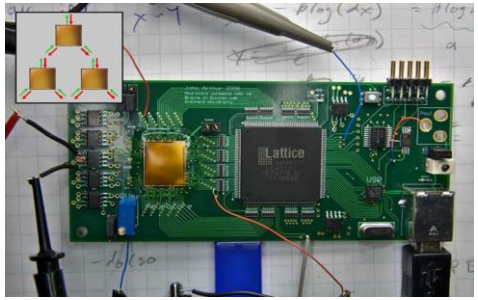
Redundancy



Replication

Barriers

- **Low number of faults** that can be tolerated
- **Coarse level of granularity**
- **Central repair mechanism** (fault-prone)



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

Reliability Challenge

Brain-inspired Hardware

Brain-inspired Information Processing

Reliability via Self-repair (the concept)

Building Self-repairing Hardware

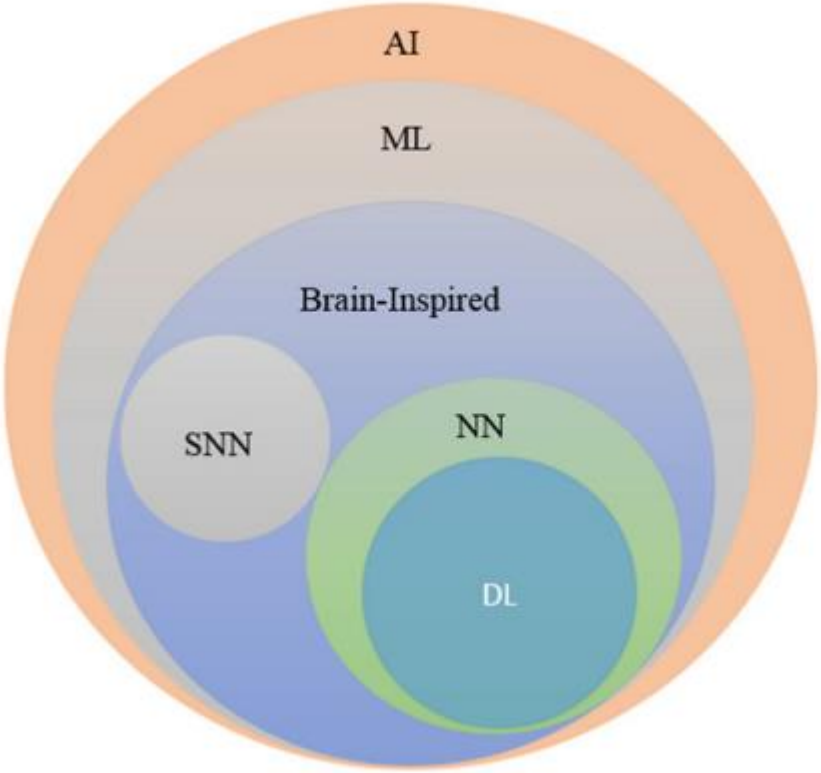
On-chip Communication Challenge and NoC Solution

The Opportunities

Artificial Intelligence

Next Computing and Engineering Wave

- Taxonomy of AI



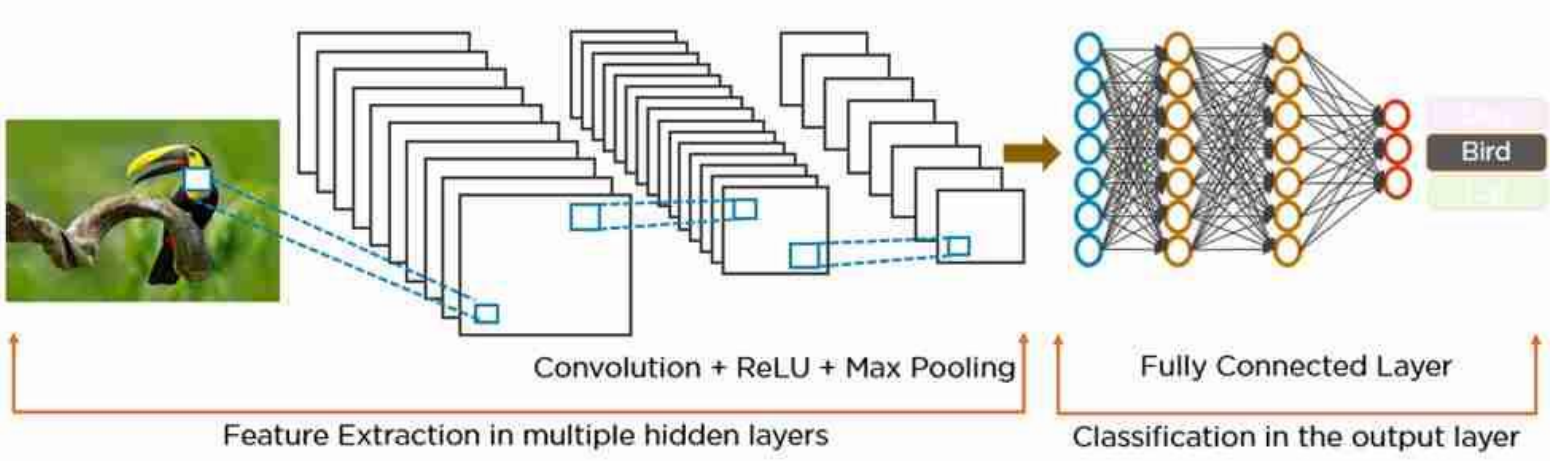
ML: Machine Learning
NN: Neural Networks
DL: Deep Learning
SNN: Spiking Neural Networks

Ulster University

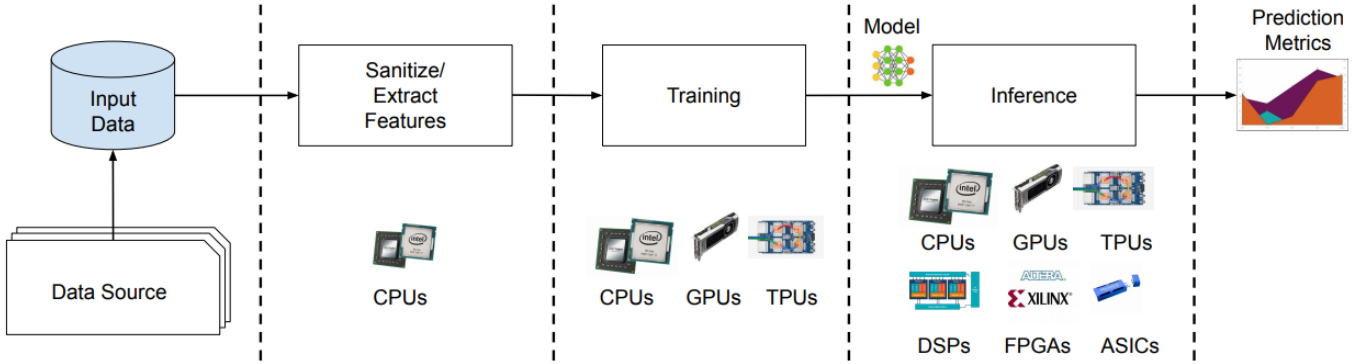
"A State-of-the-Art Survey on Deep Learning Theory and Architectures", Electronics, 8, 292, 2019

Artificial Intelligence

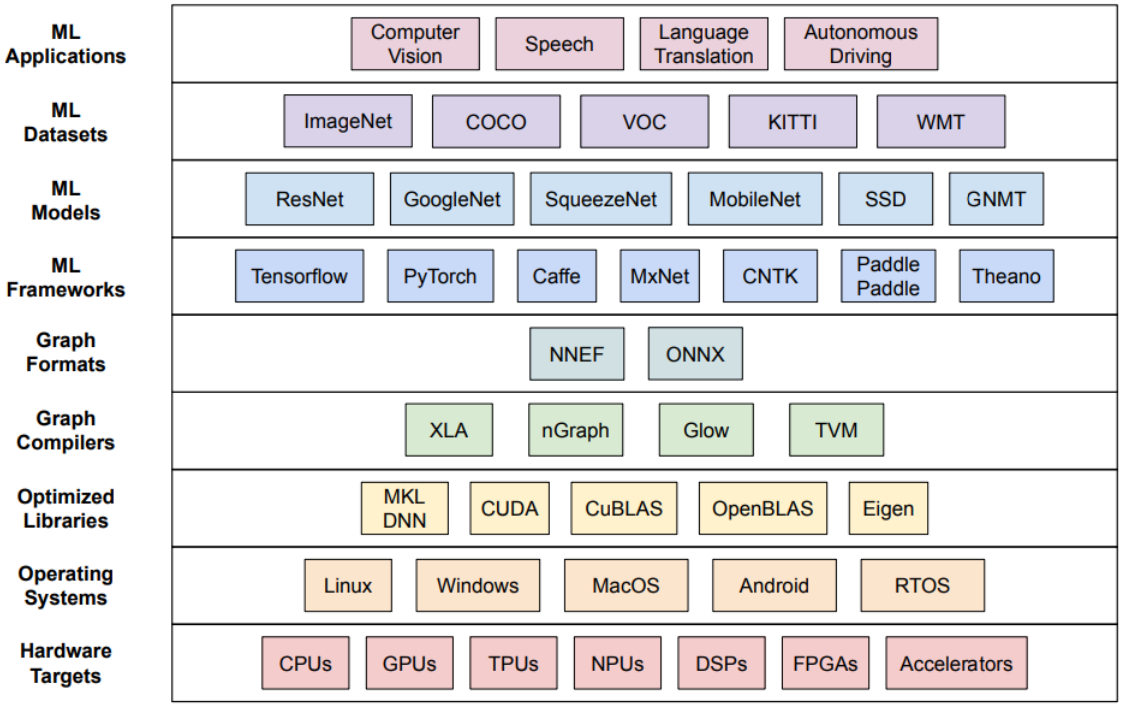
- CNN/DNN – popular in image recognition/speech processing, machine translation.



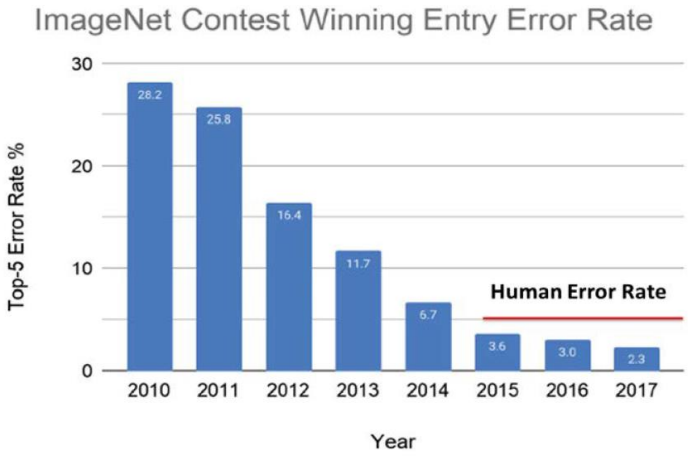
Source: "*Understanding Deep Learning: DNN, RNN, LSTM, CNN and R-CNN*" (2019)



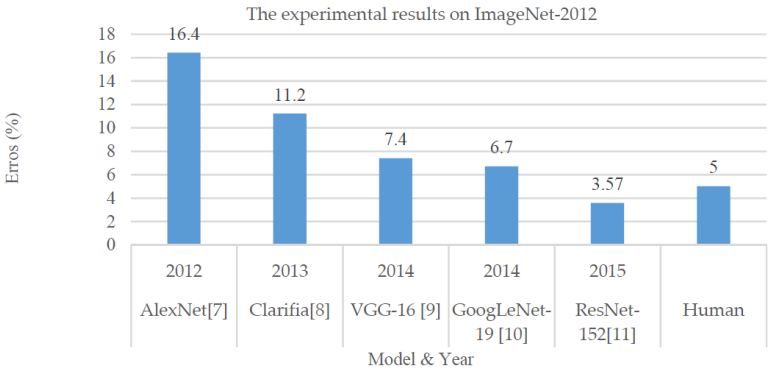
Artificial Intelligence



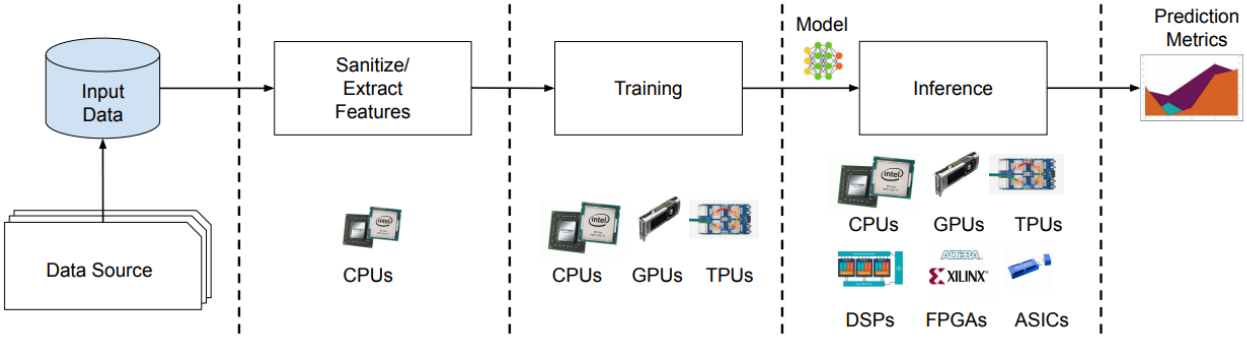
Source: "MLPerf Inference Benchmark", https://mlperf.org/ (2020)



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



Md Zahangir Alom, Electronics, 2019 (doi:10.3390/electronics8030292)



"The demands of AI neural networks, and of deep learning techniques… require thousands of petaflop-days to train…strain on energy consumption, and increasing carbon dioxide emissions."

Source: eFutures 2.0 Report, August 2021

Artificial Intelligence

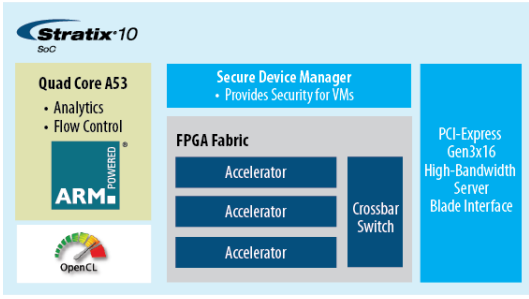
Embedded AI (Hardware)

IC Vendors	Intel, Qualcomm, Nvidia, Samsung, AMD, Xilinx, IBM, STMicroelectronics, NXP, MediaTek, HiSilicon, Rockchip	12
Tech Giants & HPC Vendors	Google, Amazon_AWS, Microsoft, Apple, Aliyun, Alibaba Group, Tencent Cloud, Baidu, Baidu Cloud, HUAWEI Cloud, Fujitsu, Nokia, Facebook	11
IP Vendors	ARM, Synopsys, Imagination, CEVA, Cadence, VeriSilicon, Videantis	7
Startups in China	Cambricon, Horizon Robotics, DeePhi, Bitmain, Chipintelli, Thinkforce	6
Startups Worldwide	Cerebras, Wave Computing, Graphcore, PEZY, KnuEdge, Tenstorrent, ThinCI, Koniku, Adapteva, Knowm, Mythic, Kalray, BrainChip, Almotive, DeepScale, Leepmind, Krtkl, NovuMind, REM, TERADEEP, DEEP VISION, Groq, KAIST DNPU, Kneron, Esperanto Technologies, Gyrfalcon Technology, SambaNova Systems, GreenWaves Technology	28

(2017) https://medium.com/@shan.tang.g/a-list-of-chip-ip-for-deep-learning-48d05f1759ae

- ❑ **Acceleration of network inference and/or training**
- ❑ **Low power**
  - ✓ **Movidius** Myriad 2 - Deep learning
  - ✓ **Huawei's** Neural Processing Unit (NPU) (Kirin 970)
  - ✓ **ARM** Ethos-U55 NPU (Edge AI - CortexM55)
  - ✓ **Google** Pixel Neural Core (Google Pixel)
  - ✓ **Qualcomm** Hexagon DSP (AI Engine\Snapdragon 865)
  - ✓ **Habana Labs** Gaudi and Goya neural processors (replaces **Nervana**)

Xilinx/Altera FPGAs



Artificial Intelligence

Embedded AI (Hardware)

IC Vendors	Intel, Qualcomm, Nvidia, Samsung, AMD, Xilinx, IBM, STMicroelectronics, NXP, MediaTek, HiSilicon, Rockchip	12
Tech Giants & HPC Vendors	Google, Amazon_AWS, Microsoft, Apple, Aliyun, Alibaba Group, Tencent Cloud, Baidu, Baidu Cloud, HUAWEI Cloud, Fujitsu, Nokia, Facebook	11
IP Vendors	ARM, Synopsys, Imagination, CEVA, Cadence, VeriSilicon, Videantis	7
Startups in China	Cambricon, Horizon Robotics, DeePhi, Bitmain, Chipintelli, Thinkforce	6
Startups Worldwide	Cerebras, Wave Computing, Graphcore, PEZY, KnuEdge, Tenstorrent, ThinCI, Koniku, Adapteva, Knowm, Mythic, Kalray, BrainChip, Almotive, DeepScale, Leepmind, Krtkl, NovuMind, REM, TERADEEP, DEEP VISION, Groq, KAIST DNPU, Kneron, Esperanto Technologies, Gyrfalcon Technology, SambaNova Systems, GreenWaves Technology	28

- ❑ **Acceleration of network inference and/or training**
- ❑ **Low power**
  - ✓ Movidius Myriad 2 - Deep learning
  - ✓ Huawei's Neural Processing Unit (NPU) (Kirin 970)
  - ✓ ARM Ethos-U55 NPU (Edge AI - CortexM55)
  - ✓ Pixel Visual Core (Google Pixel)
  - ✓ Qualcomm Hexagon DSP (AI Engine\Snapdragon 865)
  - ✓ Habana Labs' Gaudi and Goya neural processors (replaces Nervana)

(2017) https://medium.com/@shan.tang.g/a-list-of-chip-ip-for-deep-learning-48d05f1759ae

Xilinx/Altera FPGAs

- ❑ **New kids on the blocks.**
  - ✓ **Cerebras** (CS-1) process ~1.2 terabytes of data per second
  - ✓ **Graphcore** (Colossus GC2) x100 over CPU/GPU
  - ✓ **Groq** (Tensor Stream Processor) 250 trillion FLOPS
  - ✓ **Hailo** (Hailo-8) penny-sized + outperformed Nvidia's Xavier AGX
  - ✓ **Syntiant** (NDP100\NDP101) 140 µW x20 throughput over low-power MCUs
  - ✓ **Tenstorrent** (Grayskull)

Artificial Intelligence

Neuromorphic Computing?

- Carver Mead “Neuromorphic electronic systems” (1990)
- Much lower power

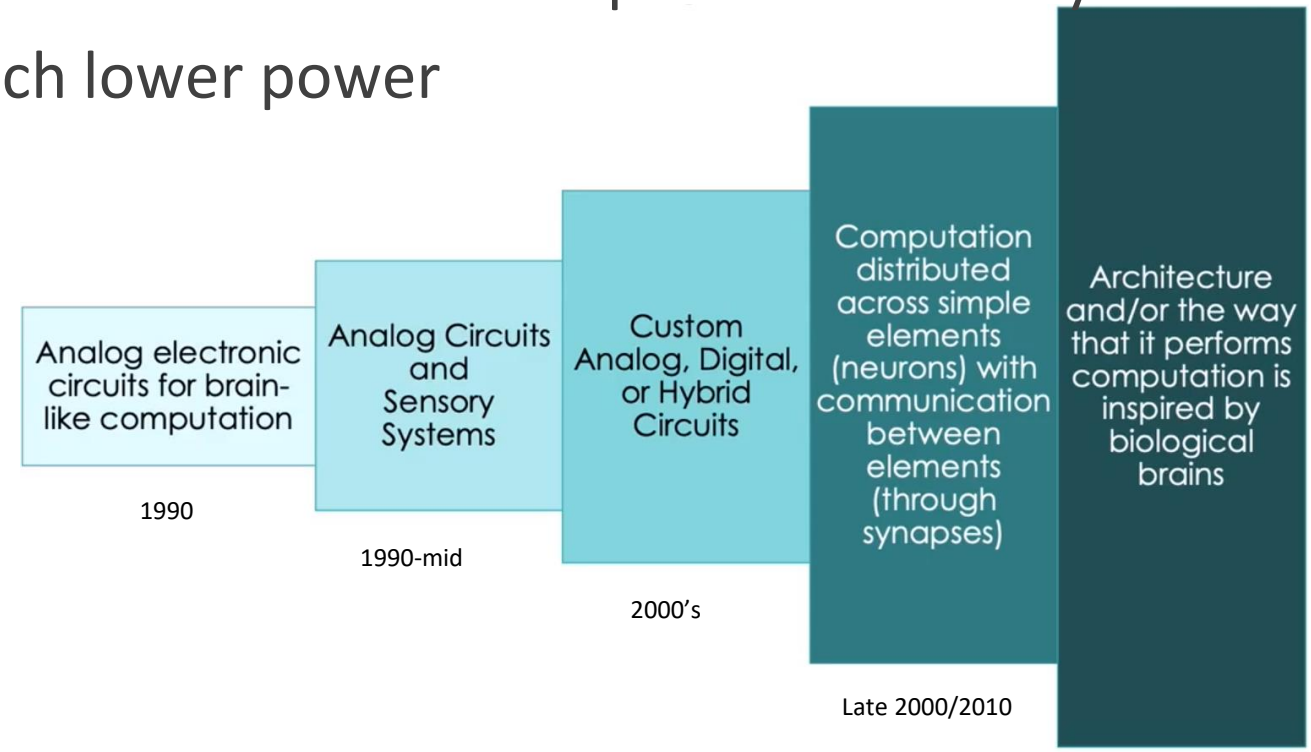
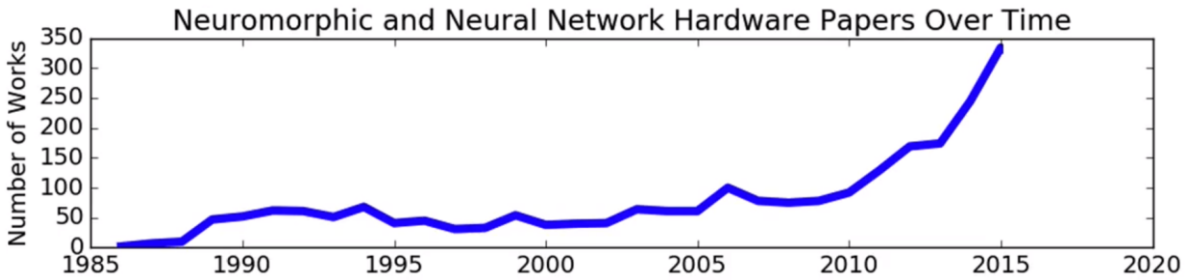


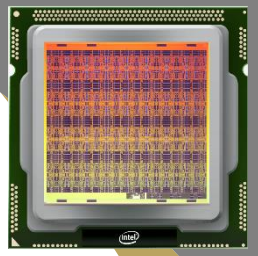
Image: Oak Ridge Labs (2020)



Catherine Schuman, "*NEUROMORPHIC COMPUTING: PAST, PRESENT, AND FUTURE*", DATE 2020

Artificial Intelligence

Neuromorphic Computing?



- Intel's neuromorphic computing (**Loihi**) 130 million synapses (x1,000 speedup/CPU) On-chip learning.

- IBM Neural network processors (**TrueNorth**) 4,096 core, *256 million synapses*
- University of Manchester (**SpiNNaker**)

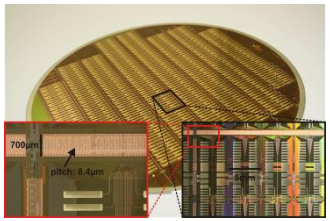
Human Brain Project

~ 10^{11} neurons

~ 10^{15} synapses

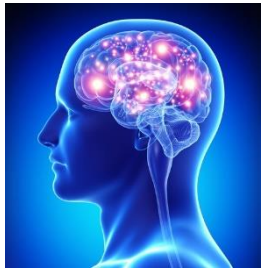
- Heidelberg University (**BrainScaleS**)

Speedup >1,000 biological time



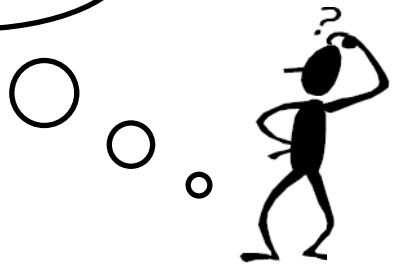
Ulster University

‘Reliability’: Unmet Need



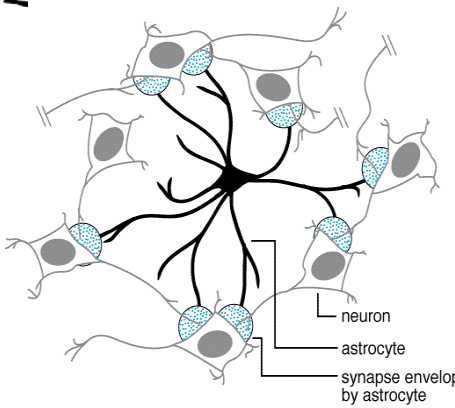
- We can learn a lot from biology, in particular neuroscience!
- Brain processing : robust and power-efficient information computing.

Look to **mimic** fault-tolerant capability of the human brain (to a degree) to build reliable computing hardware.

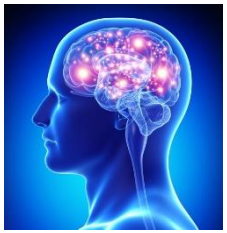


Brain employs a massively parallel computational network comprising of $\sim 10^{11}$ neurons and $\sim 10^{15}$ synapses.

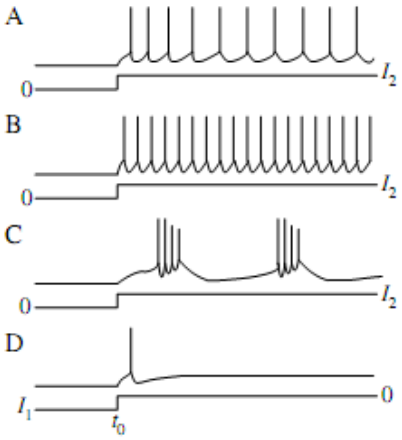
- Exploit the brain’s self-repair mechanism (**astrocyte** cells)
- Aim to develop **astrocyte-neuron networks**…….
  “Self-rePAiring spiking Neuron NEtwoRk” (SPANNER)



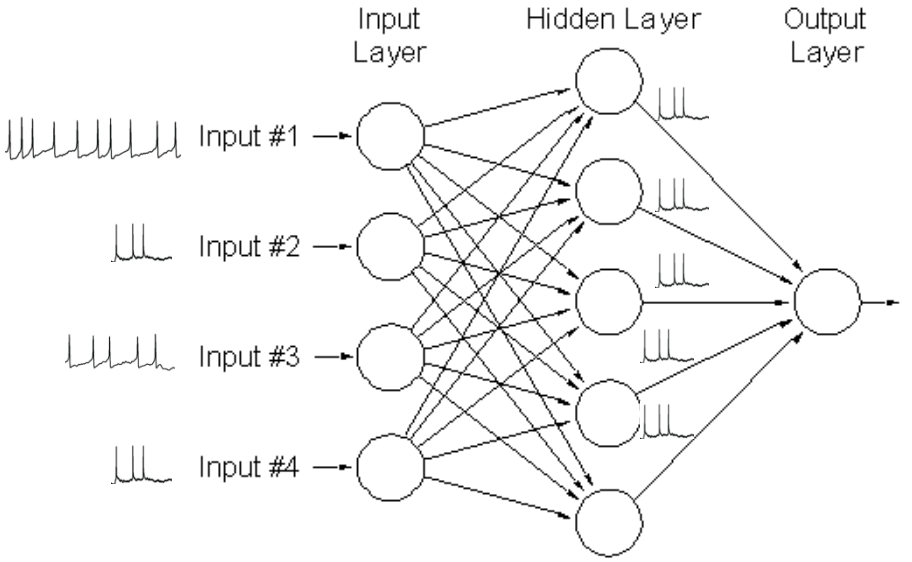
Brain-Inspired Data Processing



- **Spiking neural network (SNN)** a more biologically plausible model of the brain – a neural computing paradigm.



Data representation



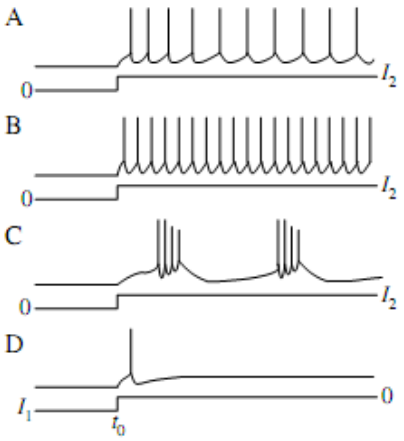
Input

- Reads I/P patterns and classifies with an O/P pattern

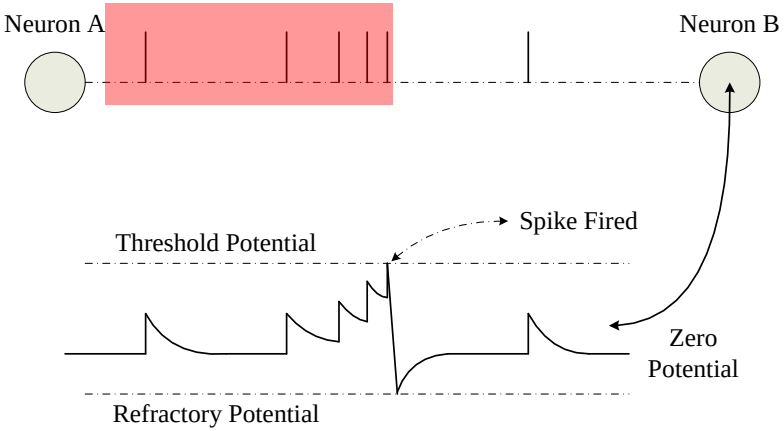
Output

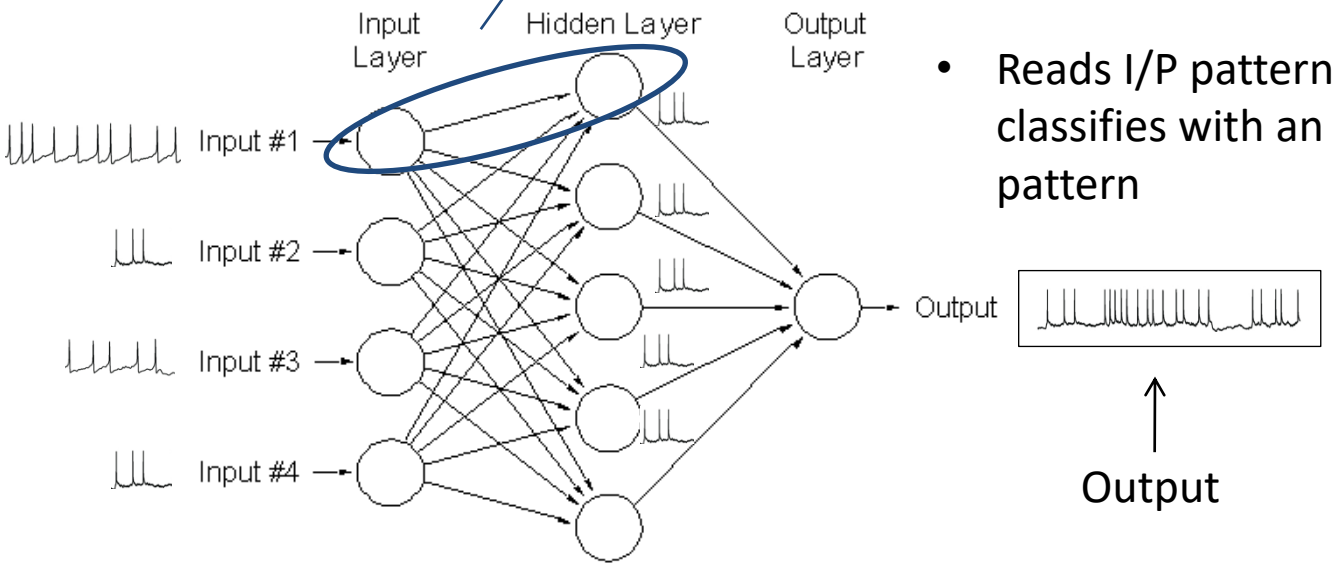
Brain-Inspired Data Processing

- **Spiking neural network (SNN)** a more biologically plausible model of the brain – a neural computing paradigm.



Data representation





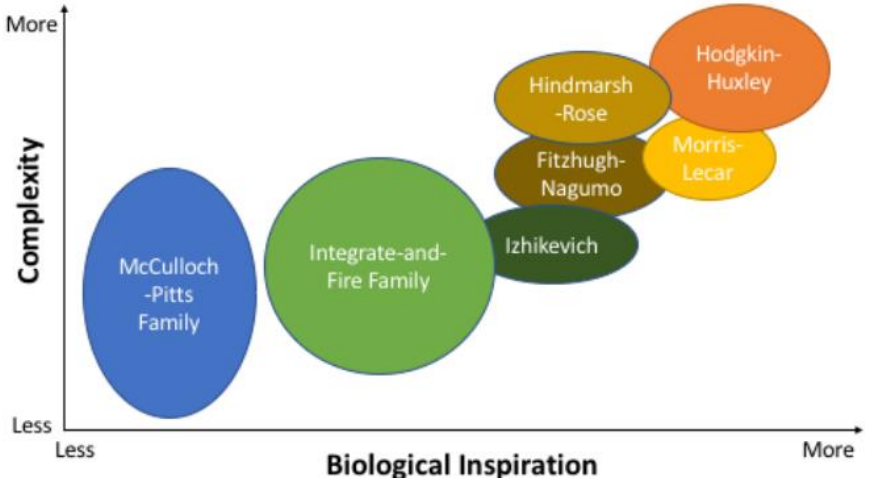
Input

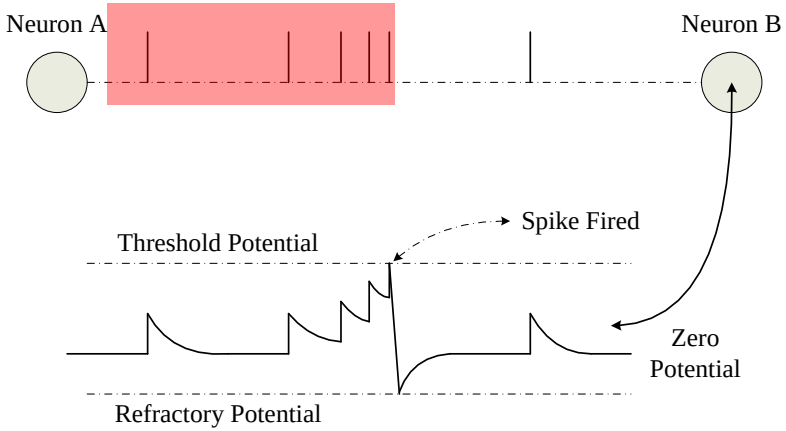
- Reads I/P patterns and classifies with an O/P pattern

Output

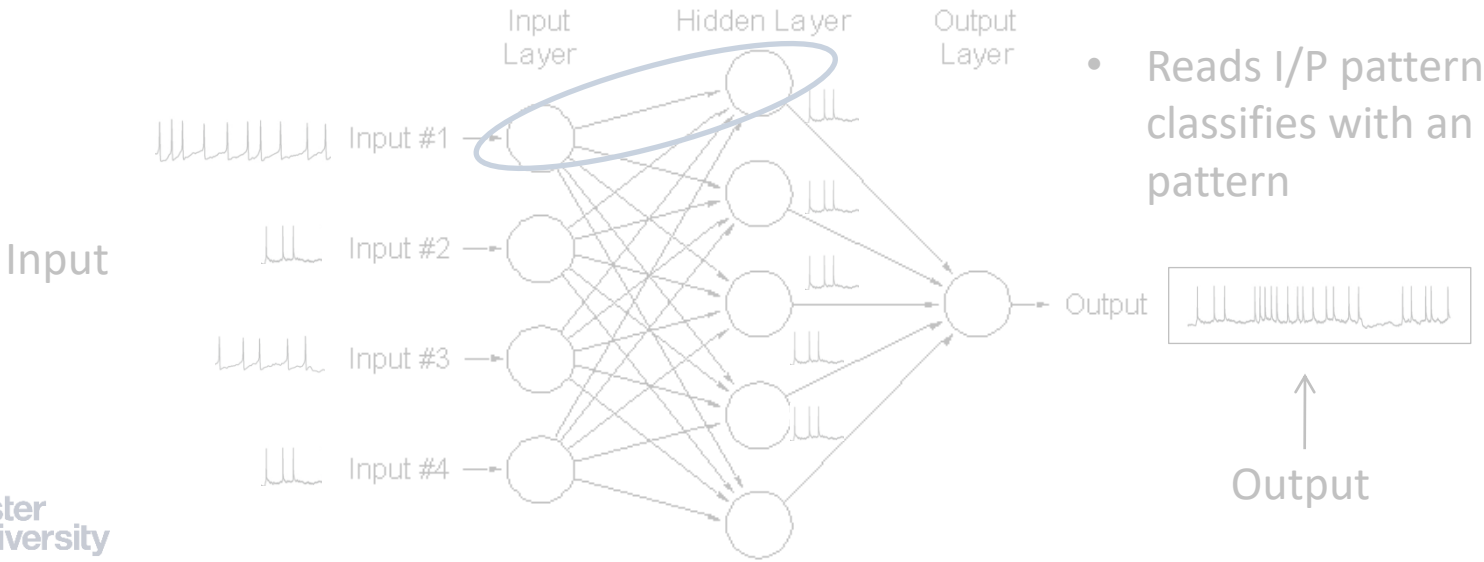
Brain-Inspired Data Processing

• **Spiking neural network (SNN)** a more biologically plausible model of the brain – a neural computing paradigm.





CD Schuman, "*A survey of neuromorphic computing and neural networks in hardware*", 2017



- Reads I/P patterns and classifies with an O/P pattern

Reliability Challenge

Brain-inspired Hardware

Brain-inspired Information Processing

Reliability via Self-repair (the concept)

Building Self-repairing Hardware

On-chip Communication Challenge and NoC Solution

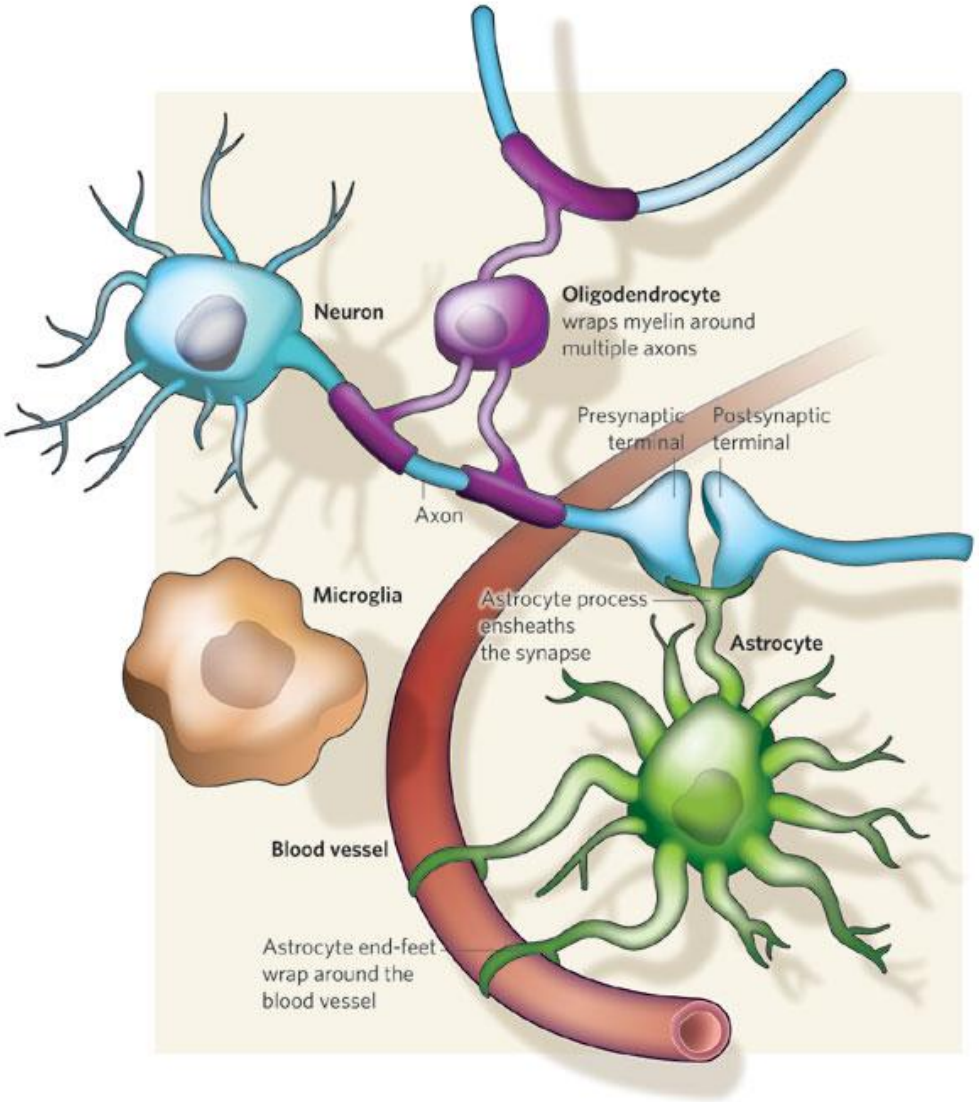
The Opportunities

Ulster University

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

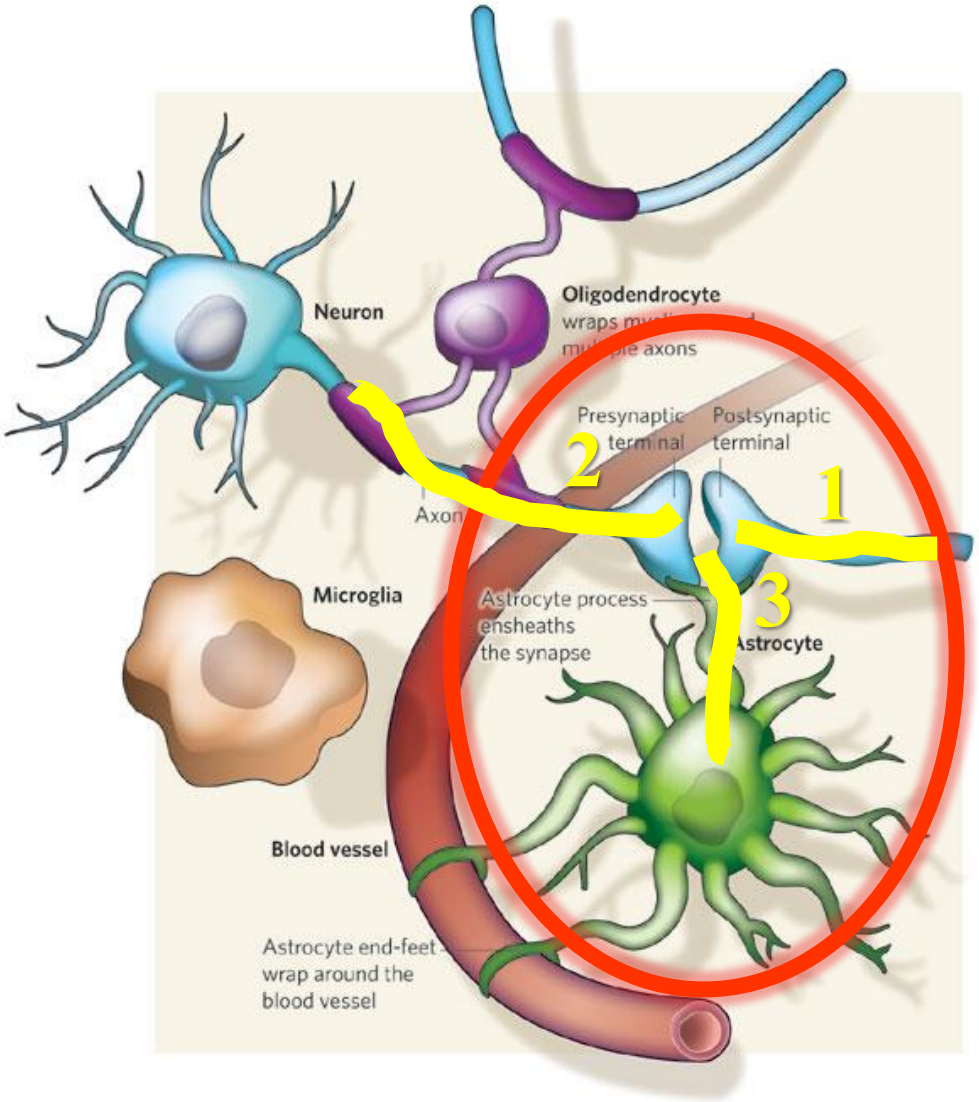


Nicola J. Allen & Ben A. Barres, "***Glia — more than just brain glue***," Nature, vol. 457, 675-677, 2009. doi:10.1038/457675a

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

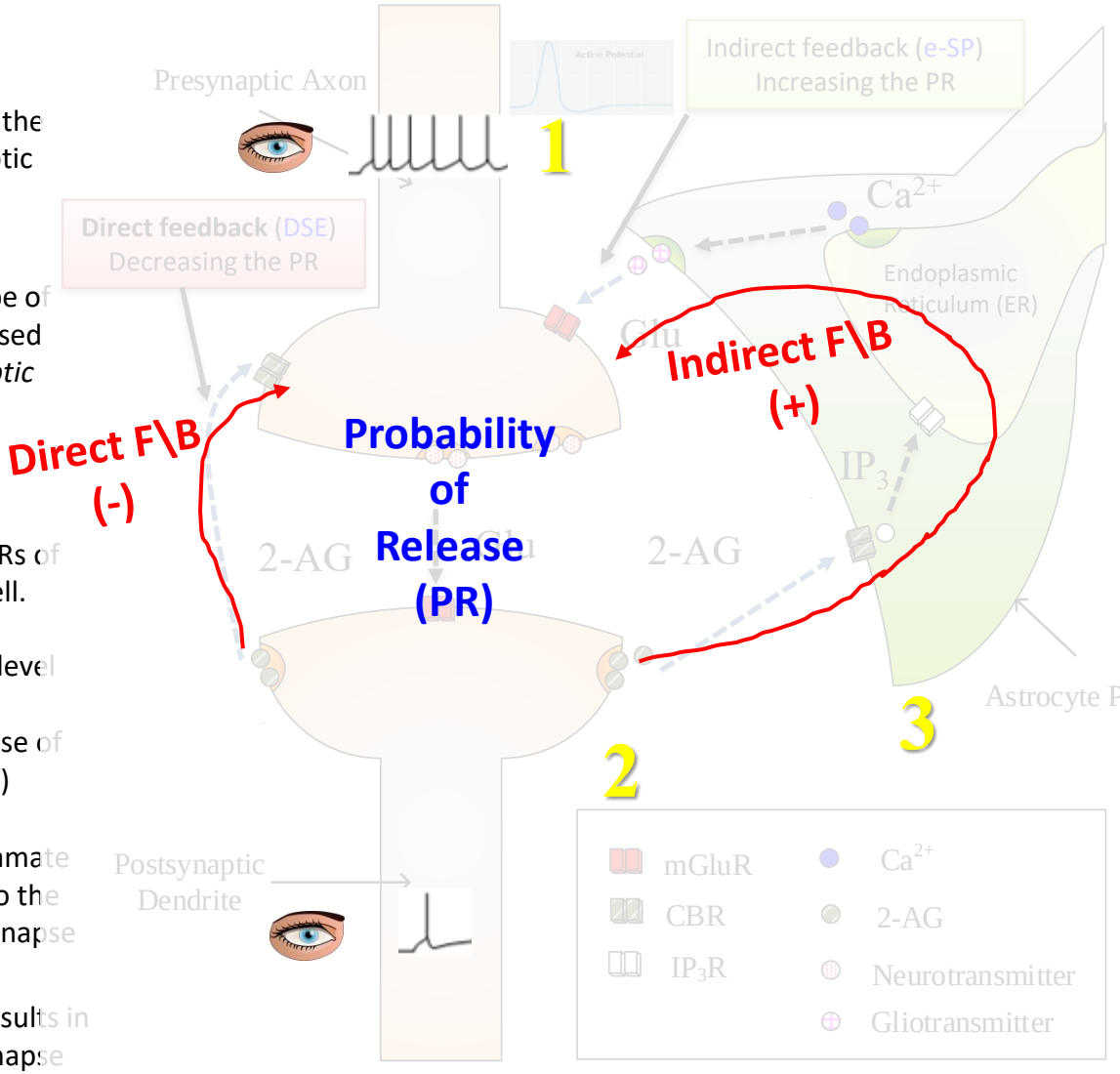
2-AG binds to CB1Rs of the astrocyte cell.

Increasing the IP_3 level

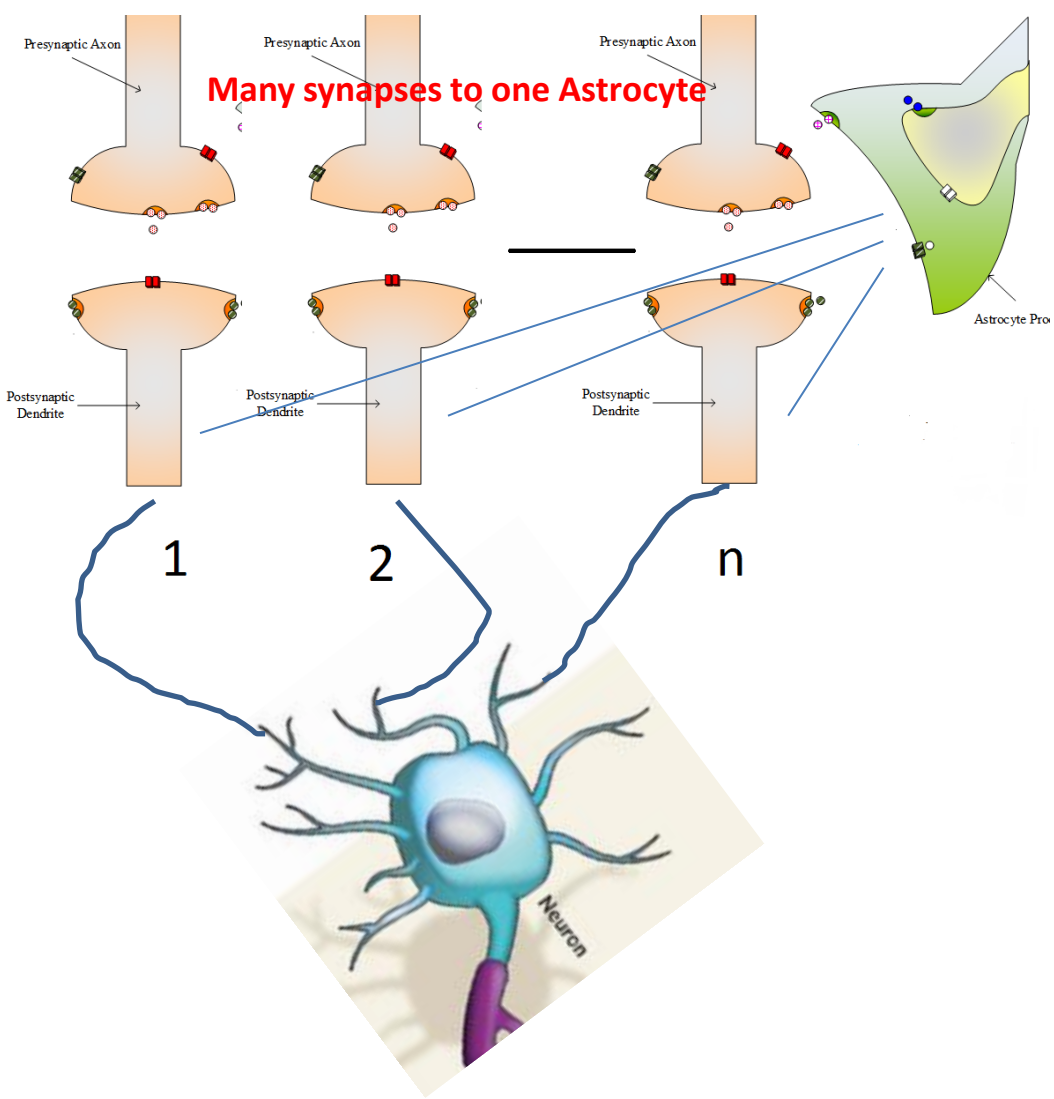
Triggers the release of calcium (Ca^{2+})

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).



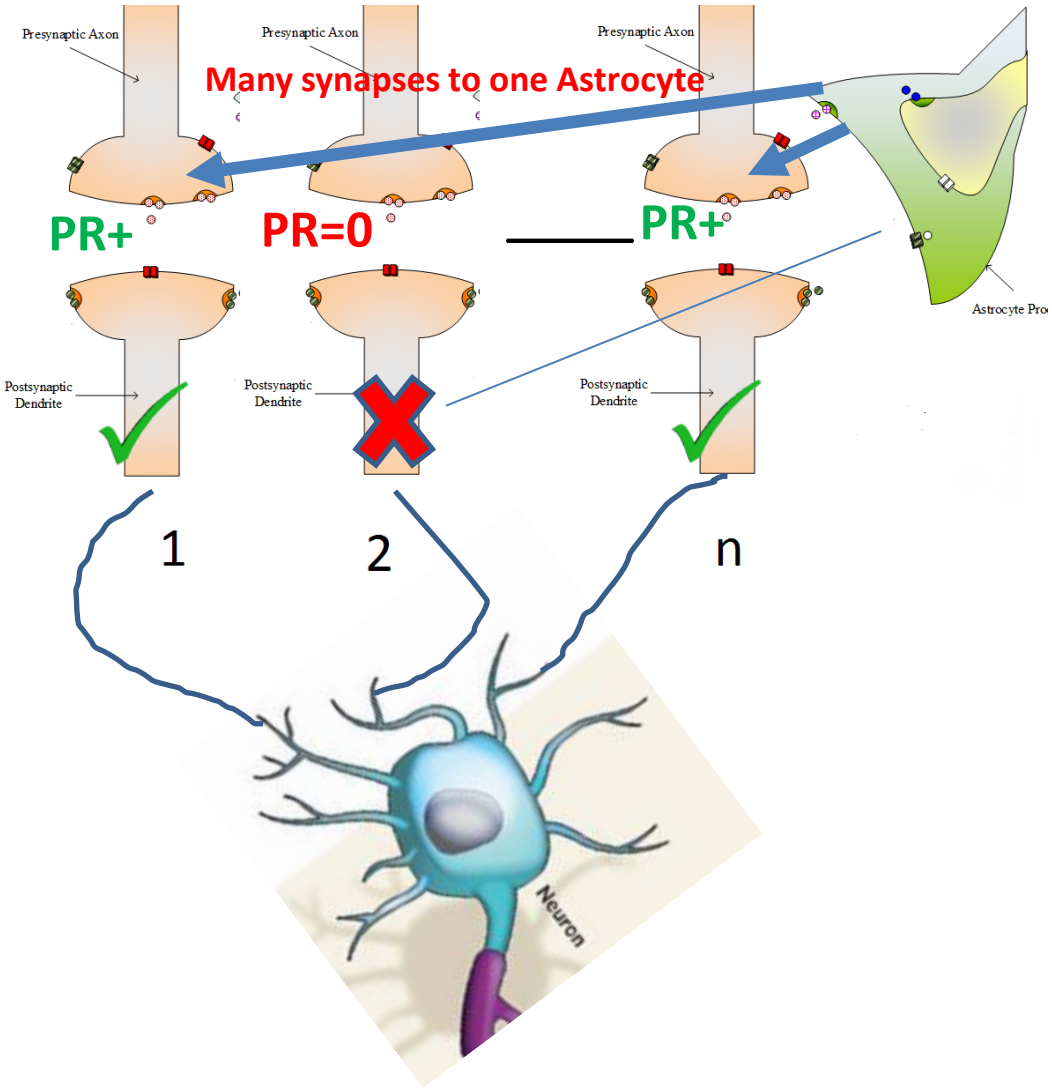
The Bigger Picture - "*Astro-Neuron Network*"



(Calcium signal broadcast) IP_3

J. Wade, L. McDaid, J. Harkin, *et. al*, "***Self-Repair in a Bidirectionally Coupled Astrocyte-Neuron (AN) System based on Retrograde Signalling***," Frontiers in Computational Neuroscience, 6(76), pp. 1–12, 2012.

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

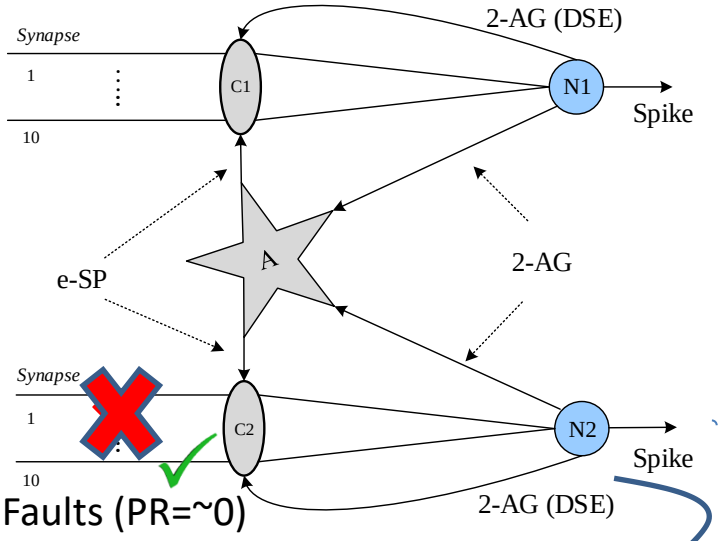


(Calcium signal broadcast) IP_3

J. Wade, L. McDaid, J. Harkin, *et. al*, "***Self-Repair in a Bidirectionally Coupled Astrocyte-Neuron (AN) System based on Retrograde Signalling***," Frontiers in Computational Neuroscience, 6(76), pp. 1–12, 2012.

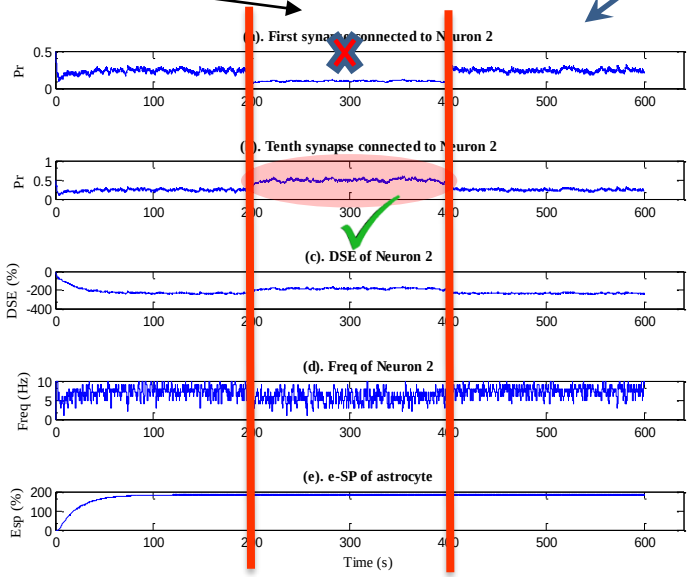
Self-repair of a 'Small' Astrocyte-Neuron Network

Temporary Faults injected



PR of Synapse #1

PR of Synapse #10

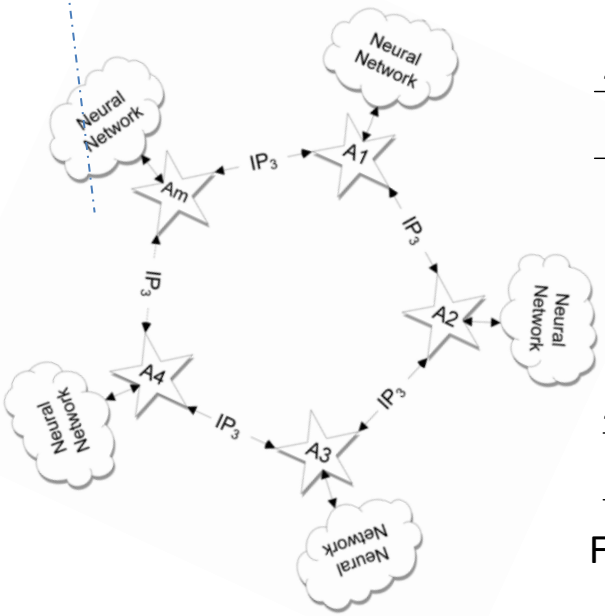


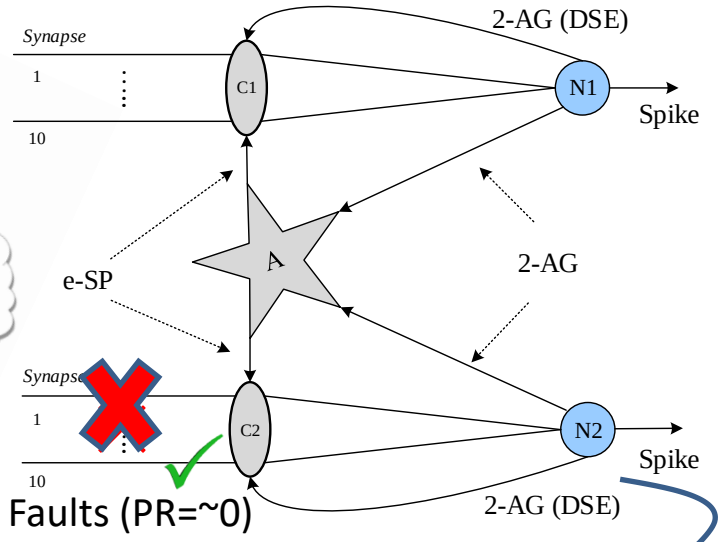
Neuron #2 under **80% fault rate** with temporary faults. (*80% - severely damaged*)

Ulster University

Naeem M, McDaid L, Harkin J, Wade J, Marsland J, "***On The Role of Astroglial Syncytia in Self-Repairing Spiking Neural Networks***", IEEE Transactions Neural Networks & Learning Systems, 2015.

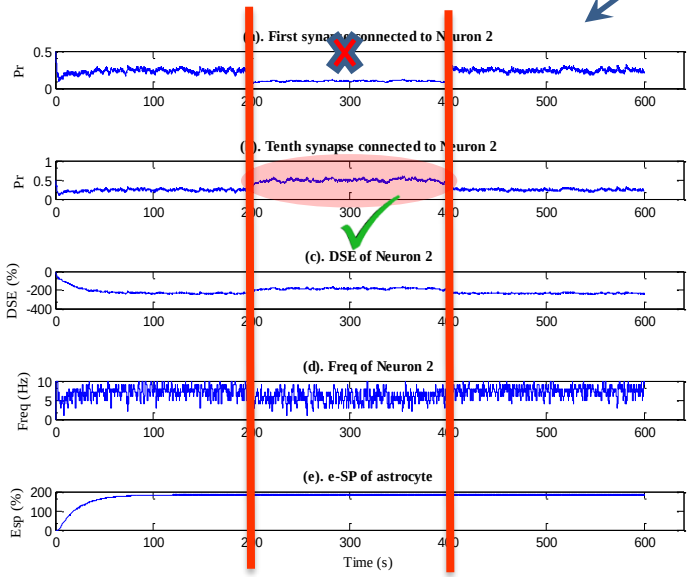
Self-repair of a ‘Small’ Astrocyte-Neuron Network





PR of Synapse #1

PR of Synapse #10



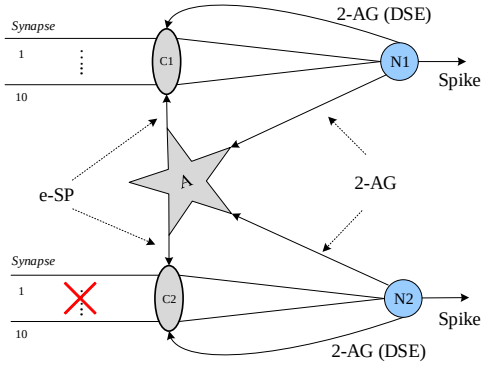
Neuron #2 under **80% fault rate** with temporary faults. (*80% - severely damaged*)

Ulster University

Naeem M, McDaid L, Harkin J, Wade J, Marsland J, “***On The Role of Astroglial Syncytia in Self-Repairing Spiking Neural Networks***”, IEEE Transactions Neural Networks & Learning Systems, 2015.

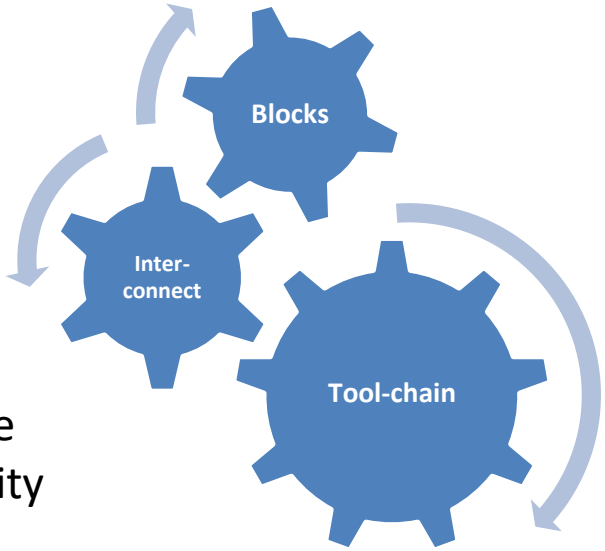
The Bigger Picture - "*Astro-Neuron Network*"

- Astrocytes are the key player in how brain performs **distributed** repair.
- What do Astrocyte-neuron networks afford?
- Increased reliability via:
  - ✓ **Fine-grained repair** at synapse-level
  - ✓ **Distributed** repair mechanism (astrocyte cell)
- New type of neural network paradigm?



Ulster University

Goal and Challenges



- **Goal -** Effectively map astrocyte-neuron networks to hardware (i.e. reliability via self-repair).
  - **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.

Reliability Challenge

Brain-inspired Hardware

Brain-inspired Information Processing

Reliability via Self-repair (the concept)

Building Self-repairing Hardware

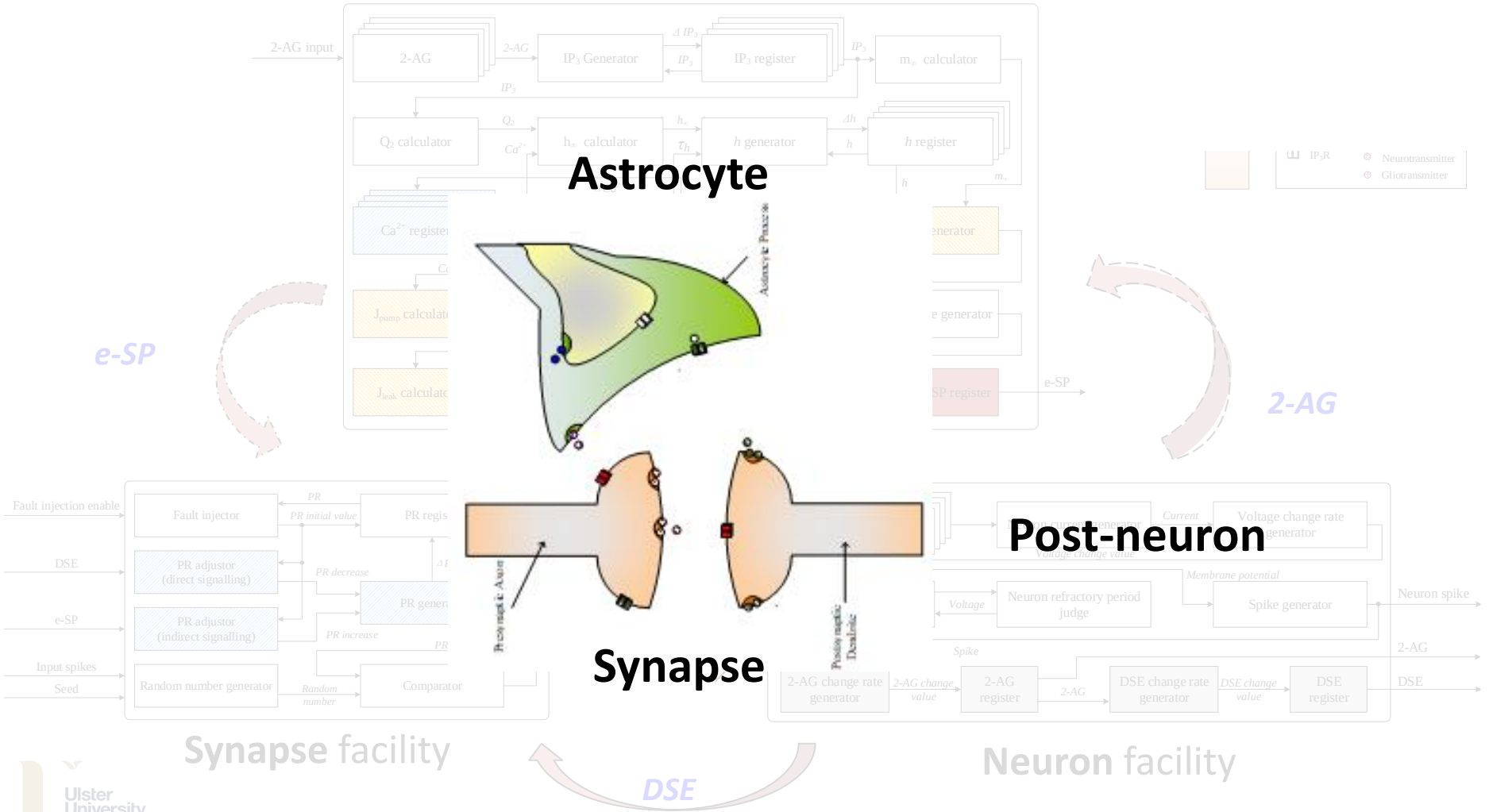
On-chip Communication Challenge and NoC Solution

The Opportunities

Ulster University

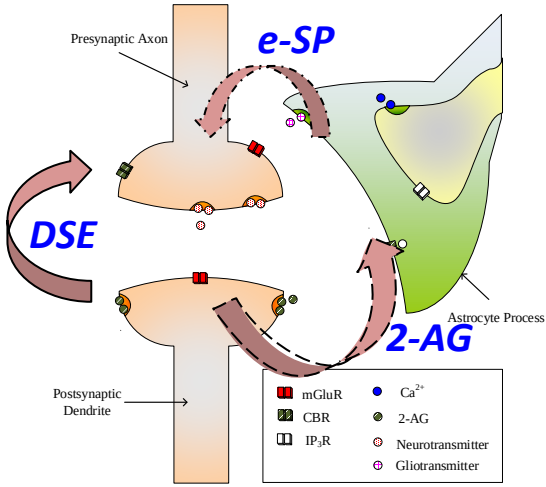
Moving to Hardware

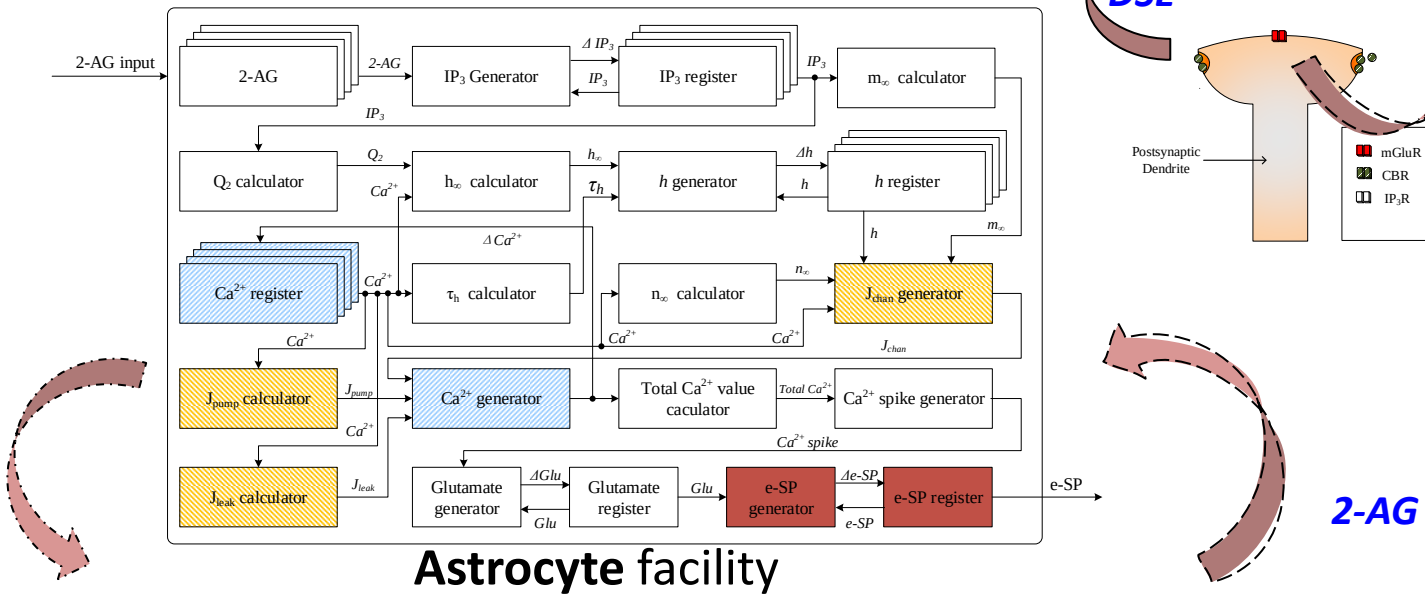
- Explored the mapping to hardware as it enables repairs to achieved when the underlying hardware is unreliable.

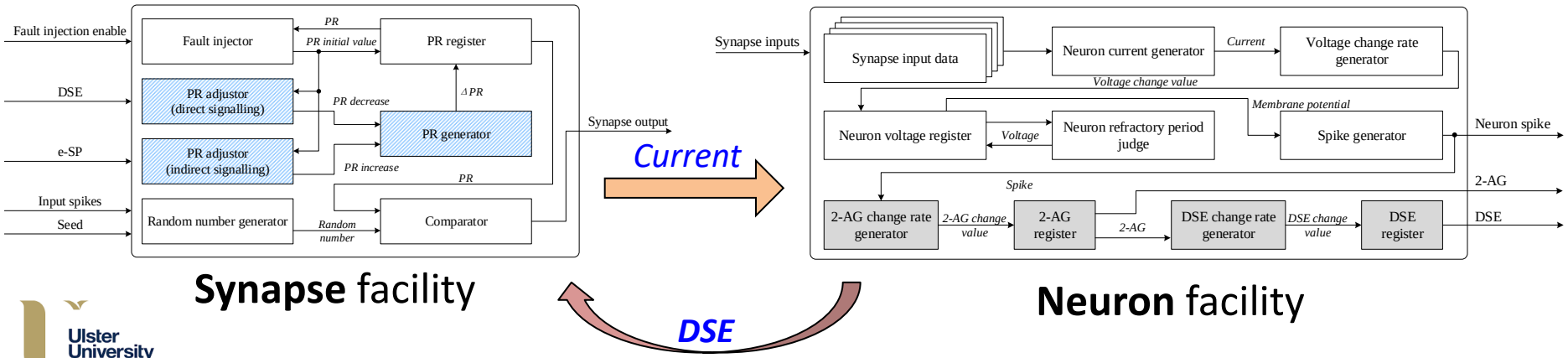


Moving to Hardware

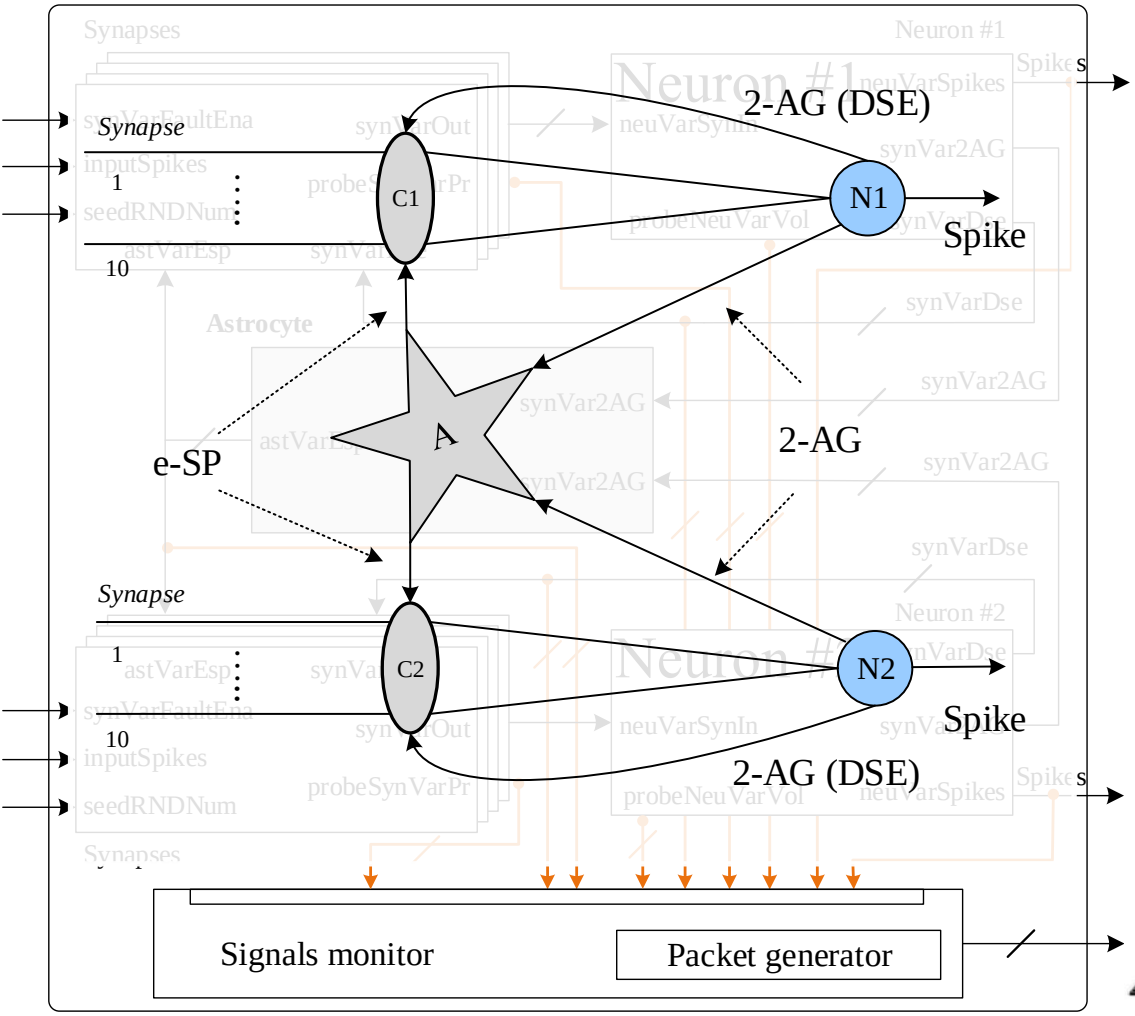
- Explored the mapping to hardware as it enables repairs to achieved when the underlying hardware is unreliable.



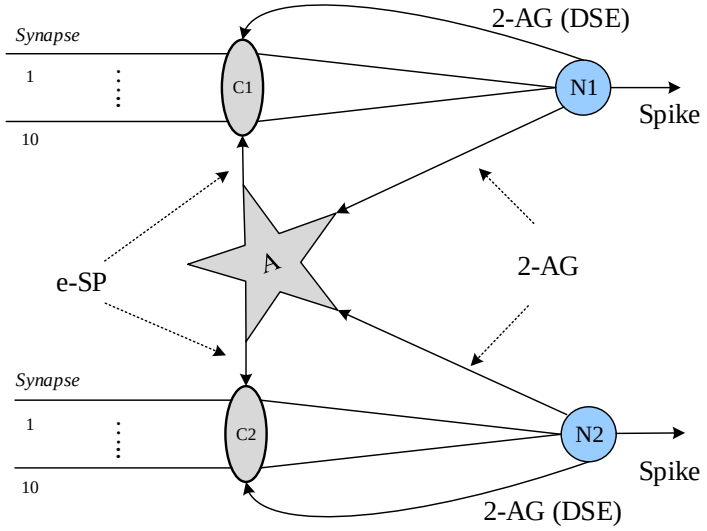




FPGA Hardware Implementation



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



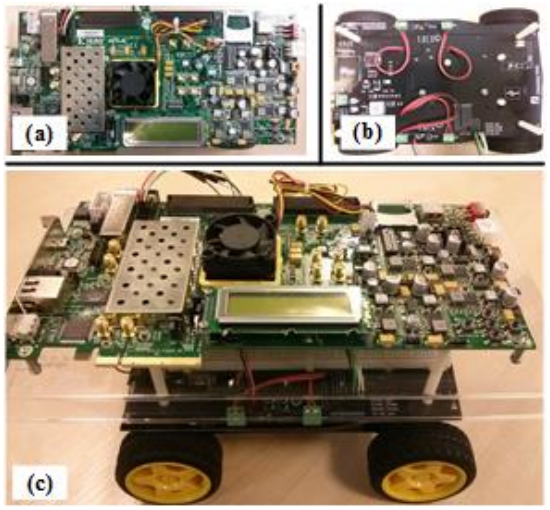
Xilinx VC707 development board

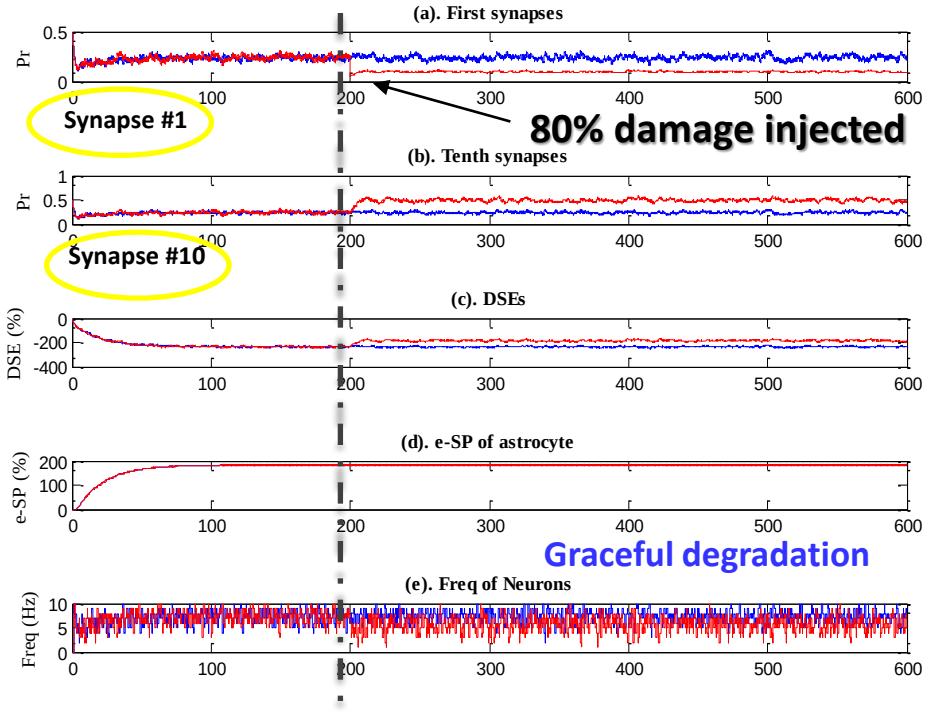
- Xilinx Virtex-7 XC7VX485T
- Vivado High Level Synthesis tool to generate the IP blocks

Ulster University

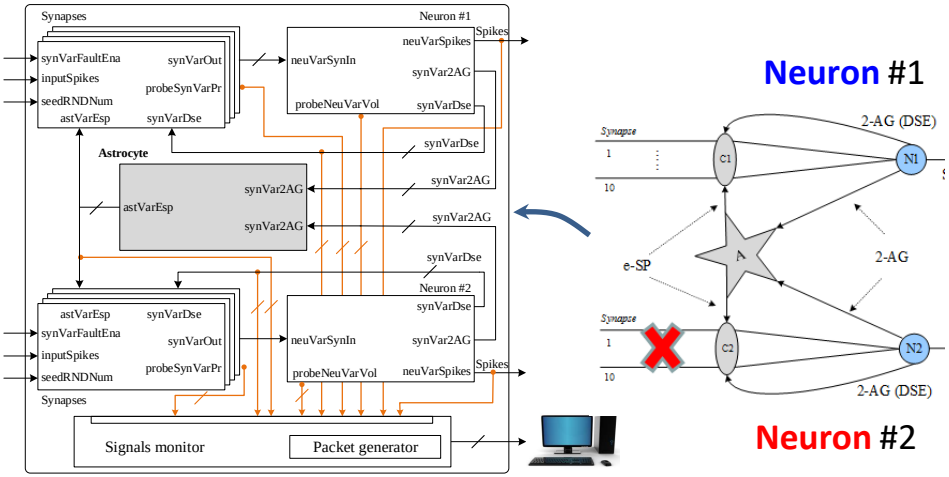
‘Small’ Astrocyte-Neuron Network to FPGAs

- 1 astrocyte, 2 neurons (20 synapse)

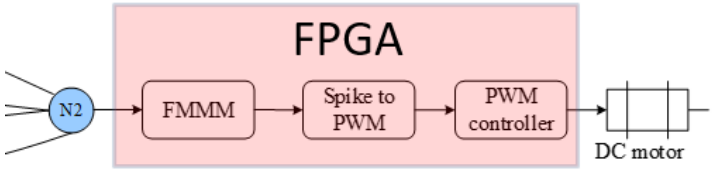




Xilinx Virtex-7 XC7VX485T

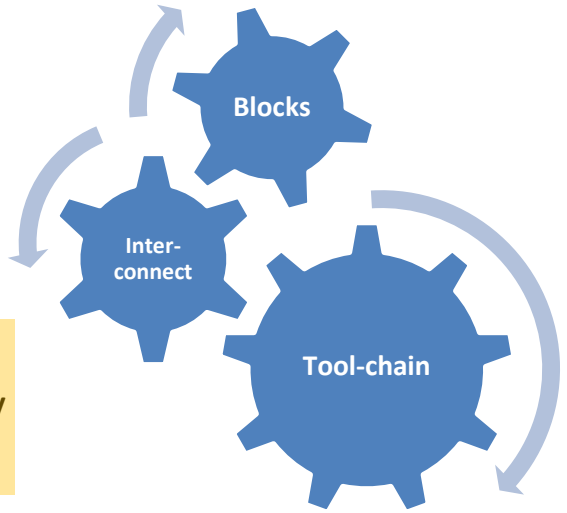


Fault density	Platform	Average output frequency	
		Neuron 1	Neuron 2
0%	Simulation	7.19	7.20
	Hardware	**7.28**	**7.27**
40%	Simulation	7.38	6.81
	Hardware	**7.37**	**6.88**
80%	Simulation	7.38	5.68
	Hardware	**7.37**	**5.75**



Liu J, Harkin J, Maguire LP, McDaid L, Wade J; "***Self-Repairing Hardware with Astrocyte-Neuron Networks***", 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.

Key Research Problem

How to implement **large-scale accelerators** on a multiple FPGA platform and **send/receive data** from the platform easily ?

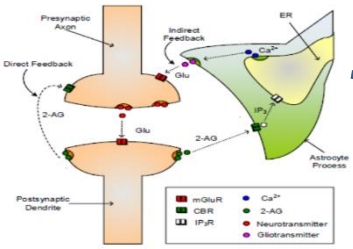
Requires the system to achieve:

- More area efficient astrocyte cells.
- Monitoring and exchanging data with the multiple FPGA platform.
- Interconnect (NoC) to channel information (inter/intra FPGA).

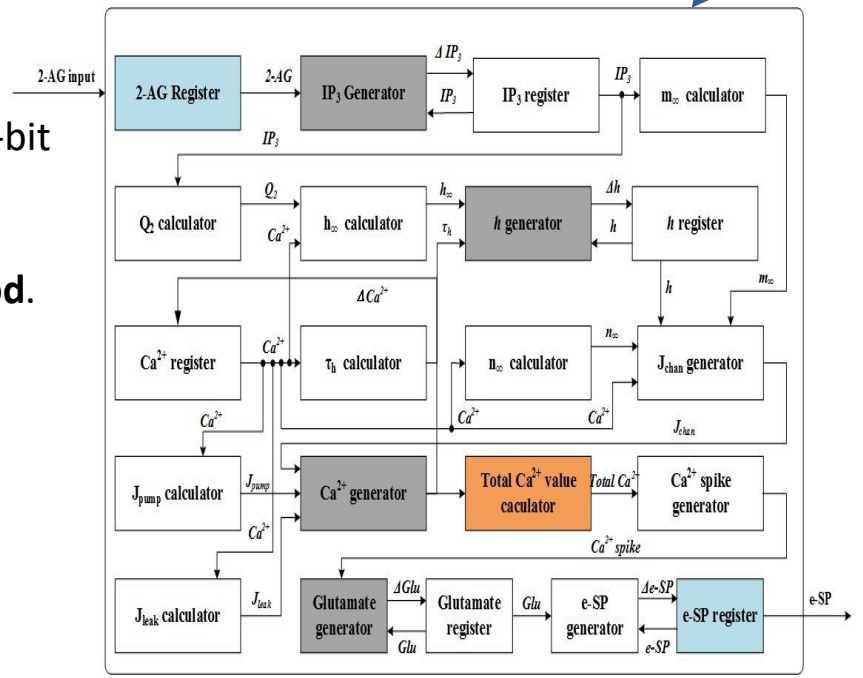
Ulster University

Refining Astrocyte Hardware

- The Astrocyte model implemented based on original block functions.



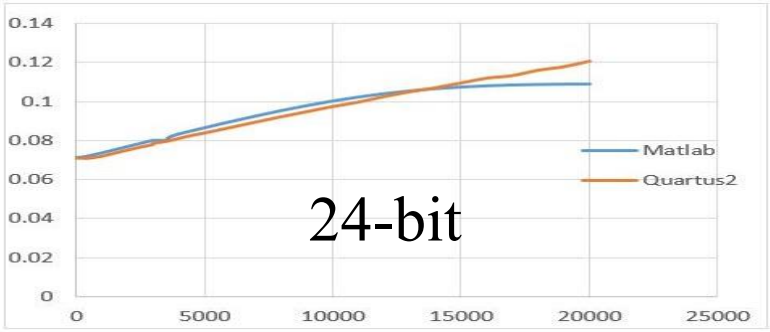
- **Astrocyte hardware** is **faithful** to the original model.
- **Each block** is sub-function in the **original model**.
- **Equations** belonging to each sub-function were **realized** using VHDL: **32-bit fixed point** (8-bit real, 24-bit fractional).
- Differential equations represented using **Euler method**.
- An **Euler constant of 0.001** seconds was used.

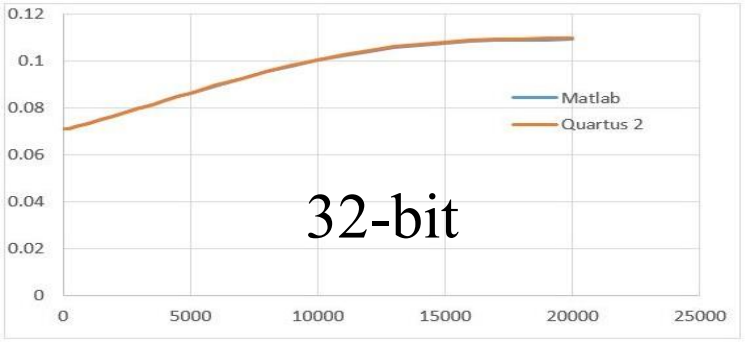


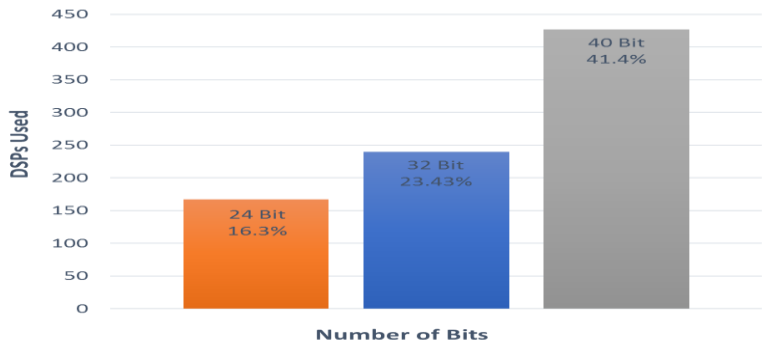
S. Karim, J Harkin, LJ McDaid et al., "**Assessing Self-Repair on FPGAs with Biologically Realistic Astrocyte-Neuron Networks**," Proc. IEEE Computing Society Annual Symposium on VLSI (ISVLSI), July, pp. 421–426, 2017

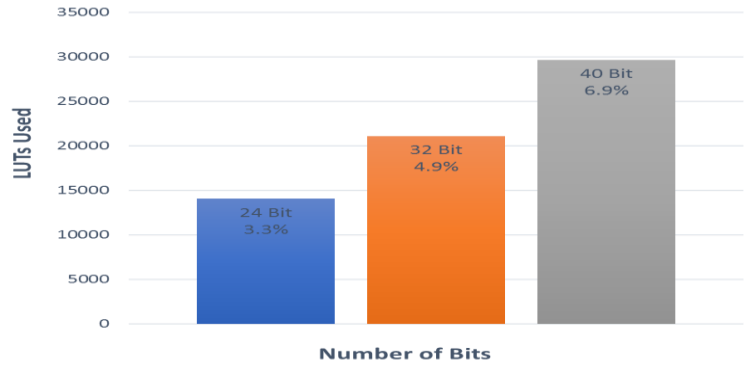
Astrocyte Precision Analysis

- Accuracy compared against double float Matlab implementation.
- 24 bit precision results in more than 10% loss in accuracy.
- 40-bit precision significantly increases FPGA resource usage without improving accuracy to any degree.
- 32 bit precision was a reasonable trade-off between resource usage and accuracy..





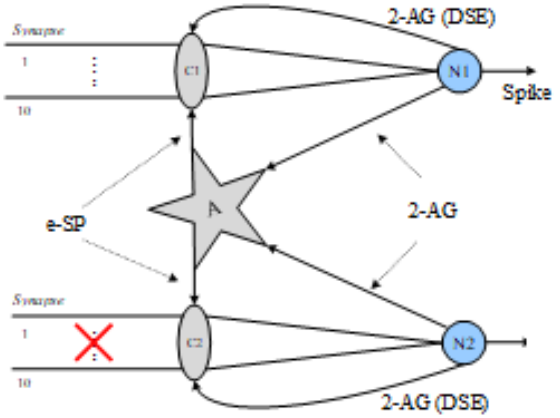




S. Karim, J Harkin, LJ McDaid et al., "**Assessing Self-Repair on FPGAs with Biologically Realistic Astrocyte-Neuron Networks**," Proc. IEEE Computing Society Annual Symposium on VLSI (ISVLSI), July, pp. 421–426, 2017

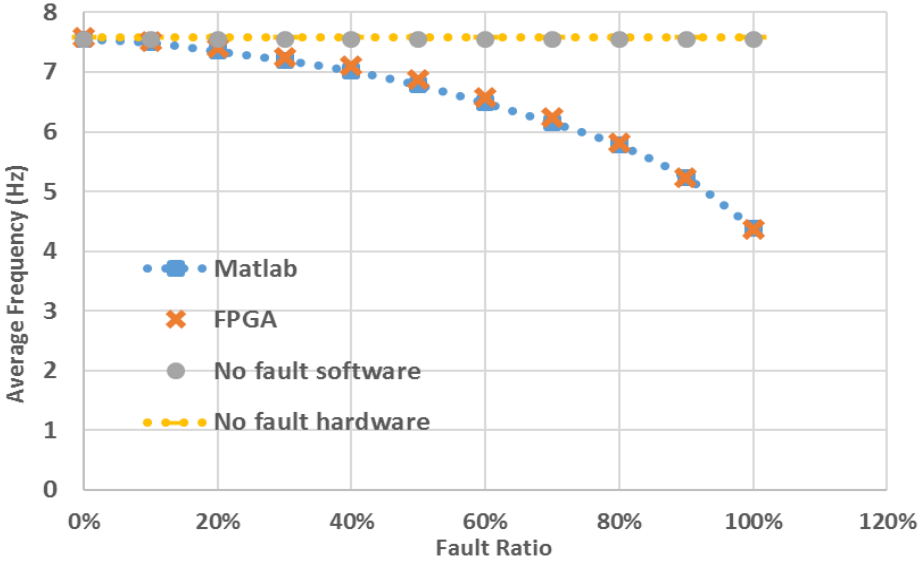
Experiments on Graceful Degradation

- **Faults** were **injected** at **specific time** during the simulation.
- **Faults severe** (reducing PR from 0.5 to zero) or **partial** (PR 0.5→0.1).
- N1 and N2 have the similar avg. O/P frequency up to **severe faults** injected.



Partial fault

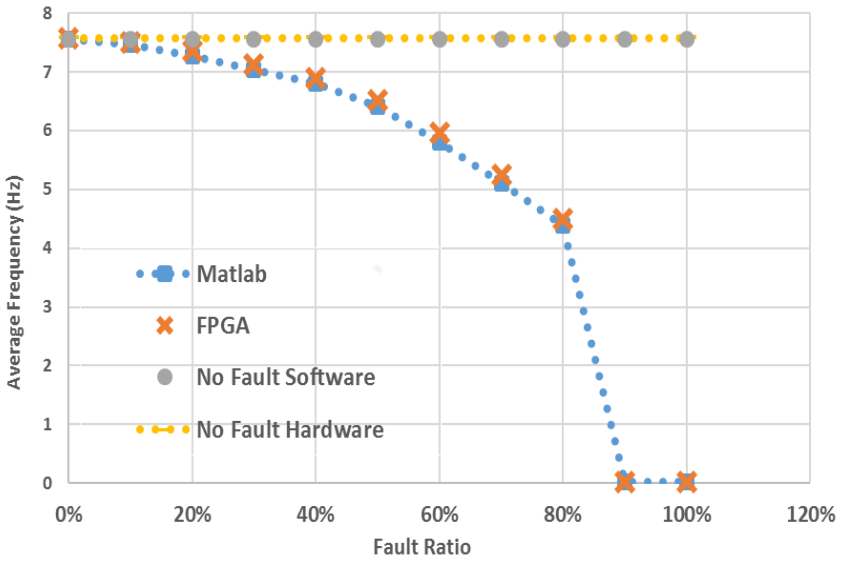
- **Spiking rate reduces gradually** to rates of **partial faults**.



(reducing PR from 0.5 to 0.1)

Severe fault

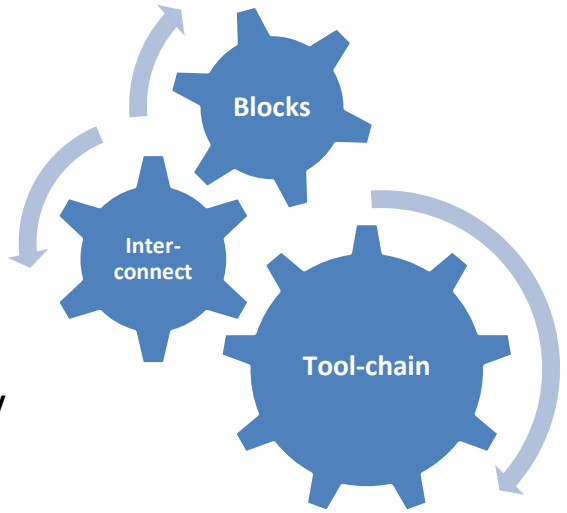
- **Spiking rate reduces more steeply** to rates of **severe faults**.



(reducing PR from 0.5 to zero)

Ulster University

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.

Reliability Challenge

Brain-inspired Hardware

Brain-inspired Information Processing

Reliability via Self-repair (the concept)

Building Self-repairing Hardware

On-chip Communication Challenge and NoC Solution

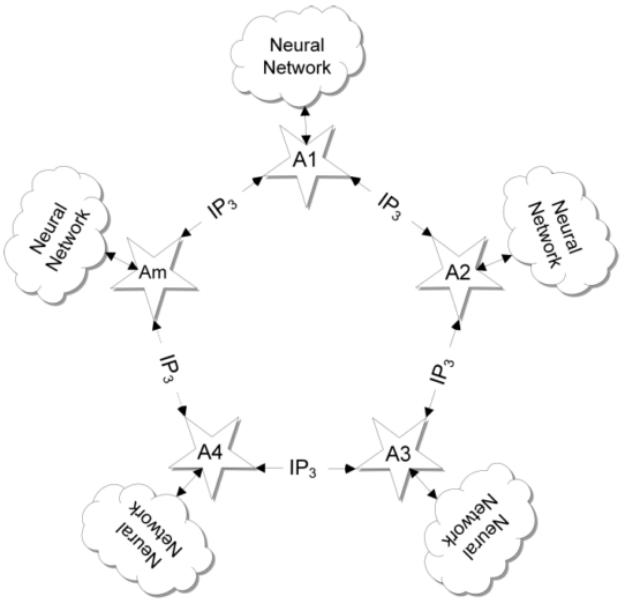
The Opportunities

Ulster University

Interconnect Challenge

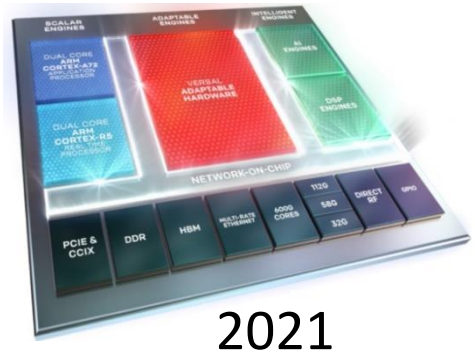
- Large **volumes of neurons\astrocytes** with **different communication** requirements:
  - **high speed** **temporal spike event** for the neuron network
  - **low speed** **numerical** inositol trisphosphate (IP_3) information exchange for astrocyte network

- Astrocyte-neuron network viewed as **two-tiered network** (neuron and astrocyte networks).



Background: Evolution Timeline

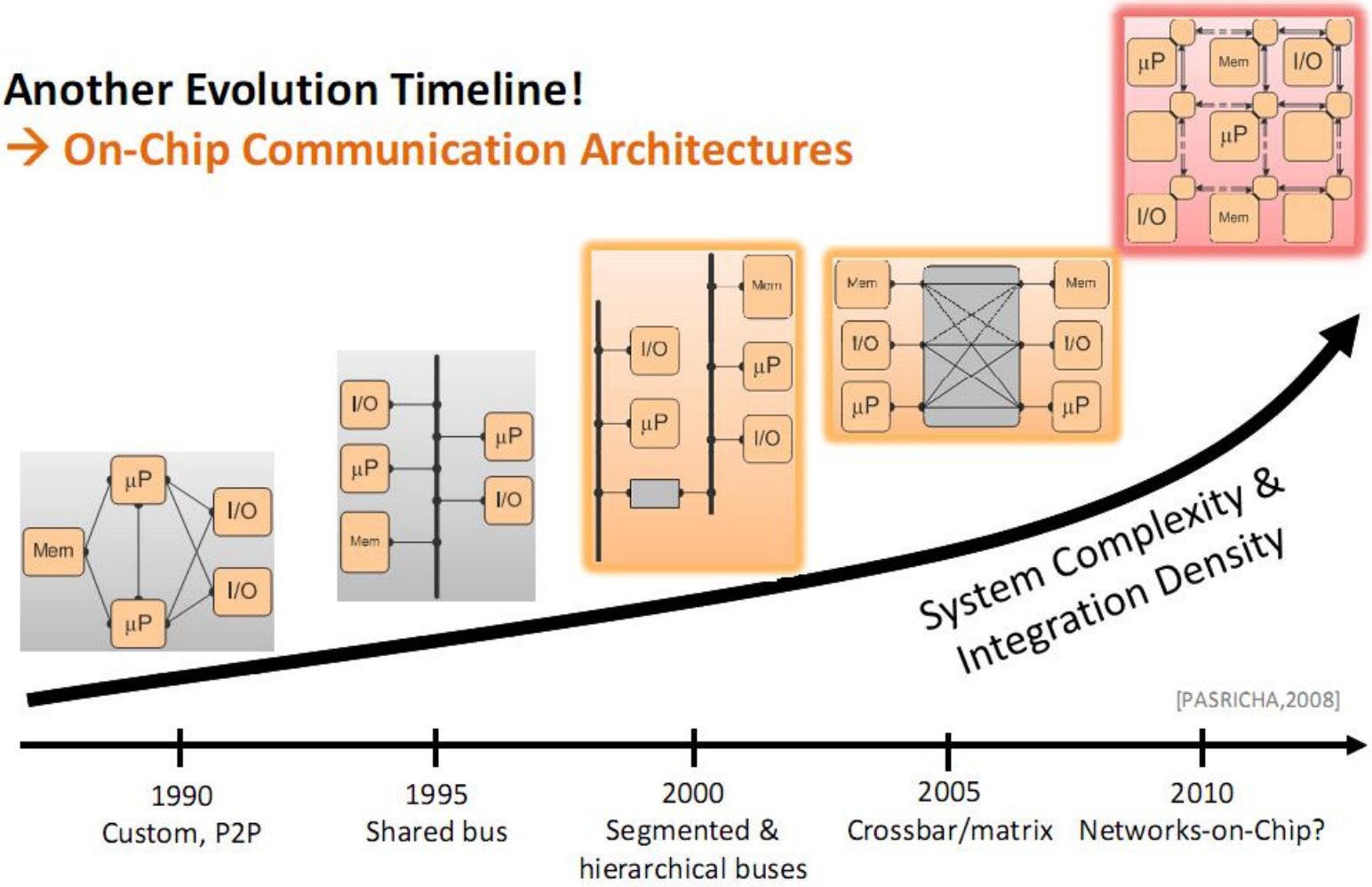
Xilinx Versal FPGA



2021

Another Evolution Timeline!

→ On-Chip Communication Architectures



PROCESSORS

French firm and TSMC put 256 processors on 28nm chip

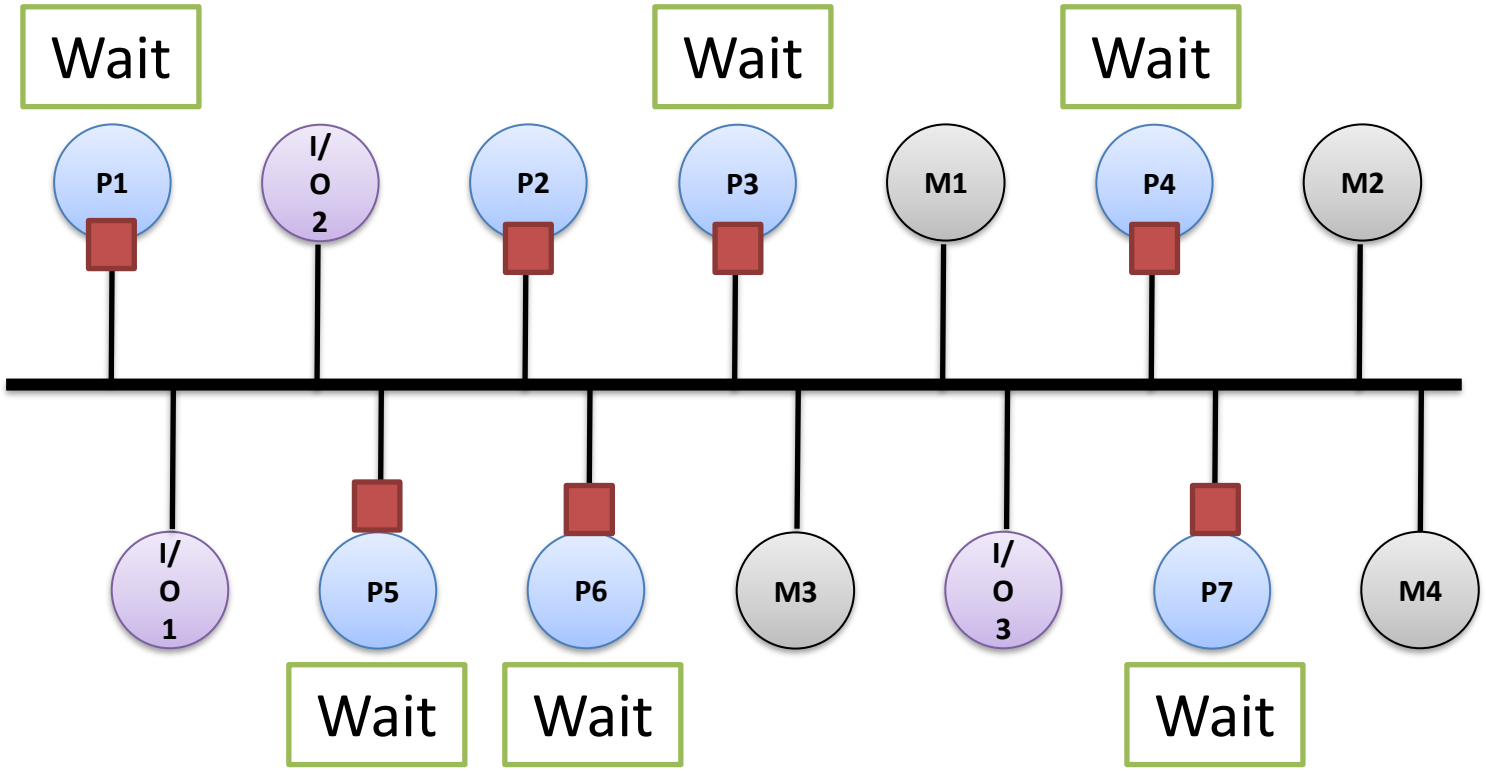
French chip firm Kalray is sampling a 28nm multicore processor. Its Multi-Purpose Processor Array (MPPA)-256 processor is fabricated by TSMC. First products to be ramped in volume will be processors for signal processing in an imaging application. Product qualification is scheduled for completion in November 2012.

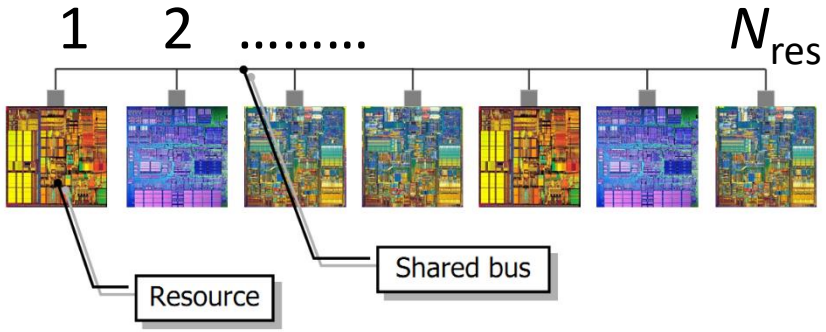
The Paris Orsay-based fabless semiconductor and software company is backed by French investment funds, local funds, private investors, and OSEO, a French public-sector institution which finances innovative projects brought by SME's.

The first MPPA-256 processor integrates 256 processors onto a single silicon chip, organised as 16 clusters of 16 processors. Multiple

On-chip Interconnection

Bus-based was Traditional



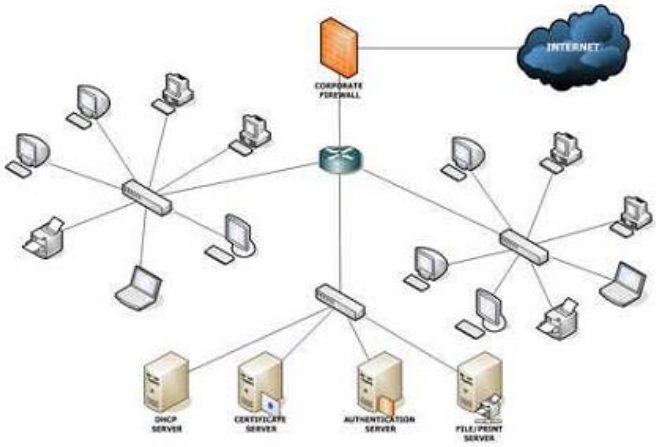


Aggregated data rate:

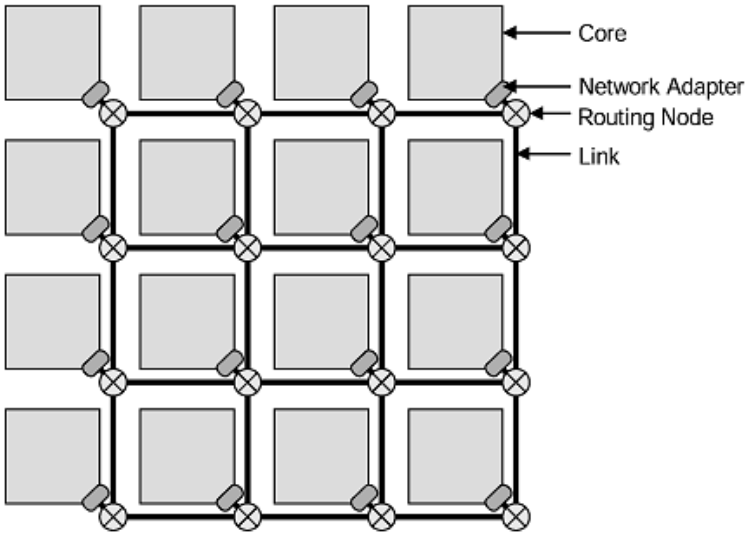
$$DR_{agg} \propto \frac{1}{N_{res}}$$

Introduction to NoC - Basic Components

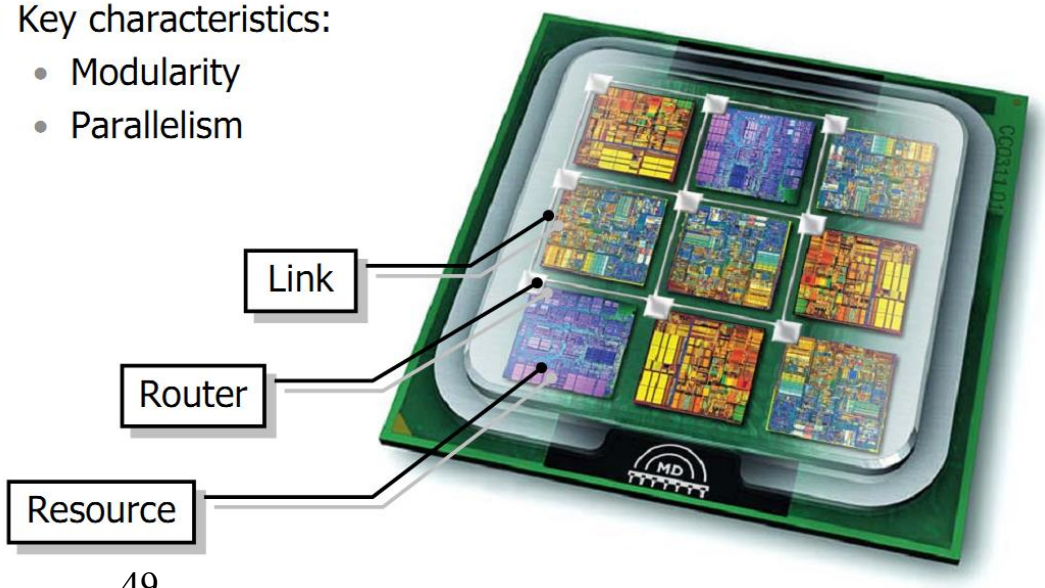
- **What is Networks-on-Chip (NoC) ???**
  - New interconnection paradigm that is used to connect electronic systems **inside a chip**



- **Why do we need it ???**
  - Number of transistors (Moore's Law)
  - Modularity (IP cores + re-used)
  - Reduce time to market

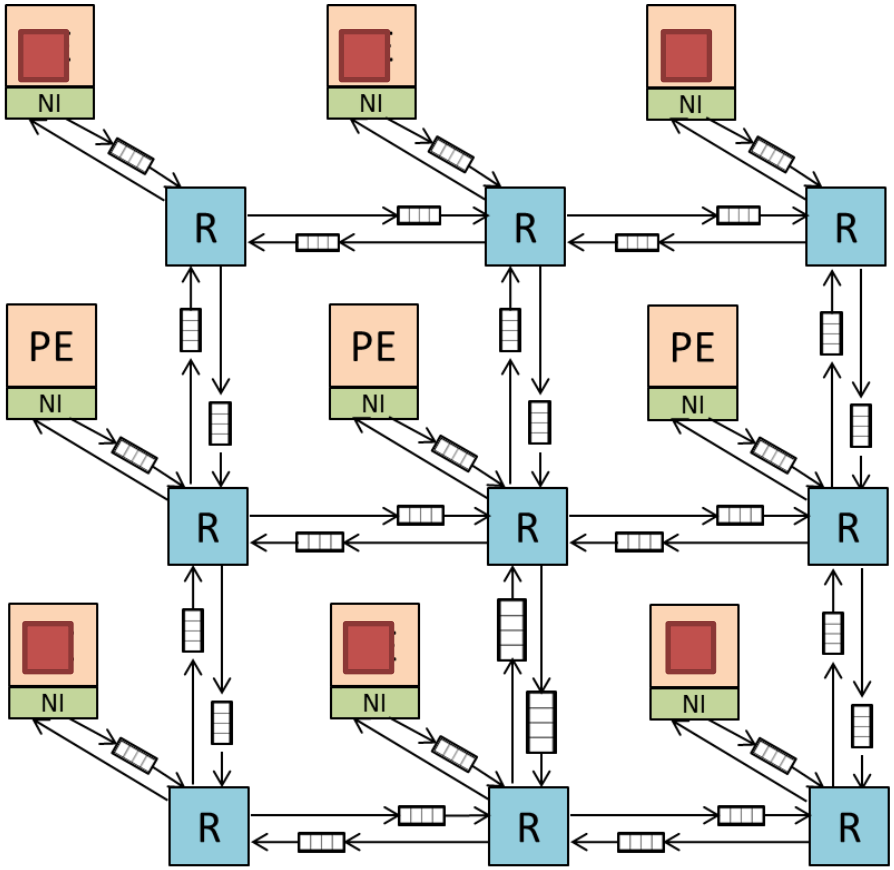


- Key characteristics:
  - Modularity
  - Parallelism



Networks-on-Chip (NoC)

Mesh Based

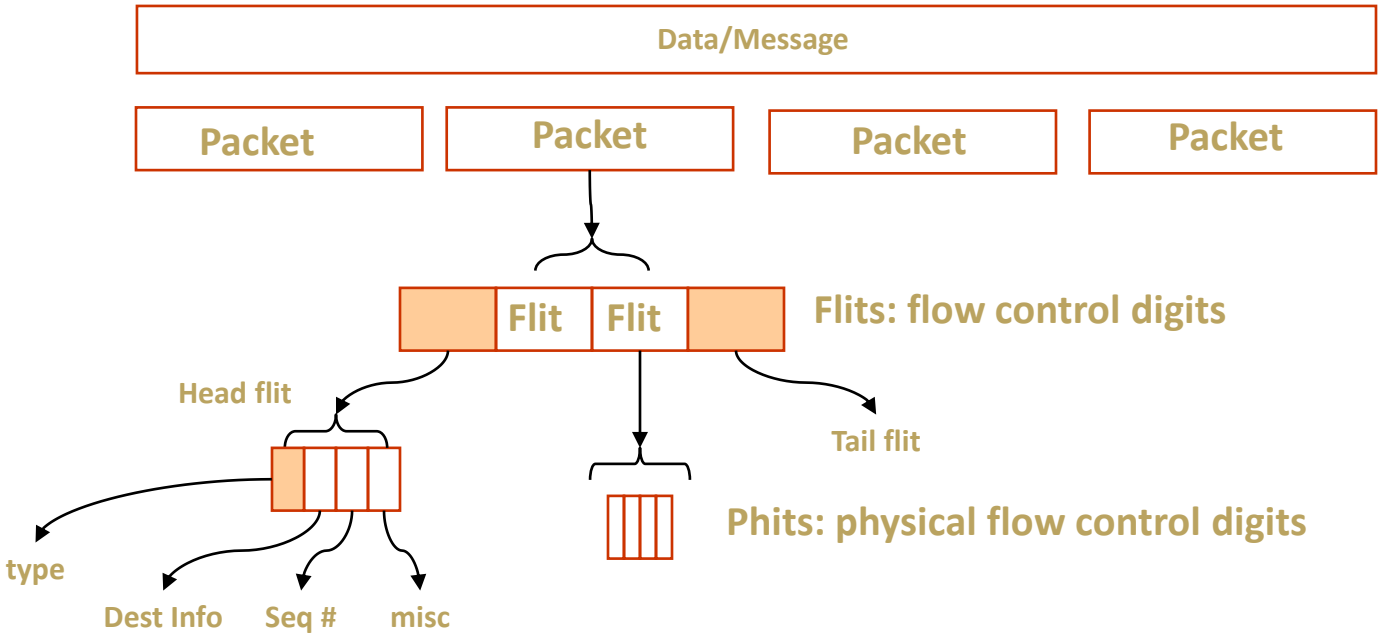


R Router

PE NI Processing Element Network Interface

Communication Data

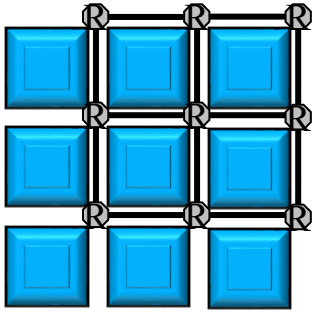
Messaging Units



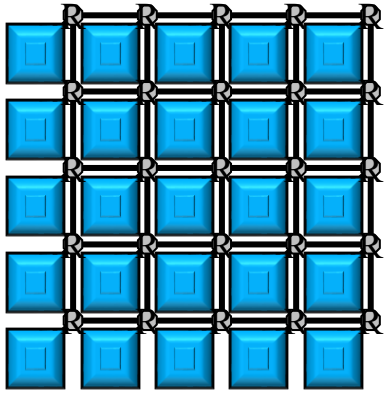
- Data is transmitted based on a hierarchical data structuring mechanism
  - Messages → packets → flits → phits
  - flits and phits are fixed size, packets and data may be variable sized
  - phit is a unit of data that is transferred on a link in a single cycle
  - typically, **phit** size = **flit** size

NoC – Its Abstraction and Challenges

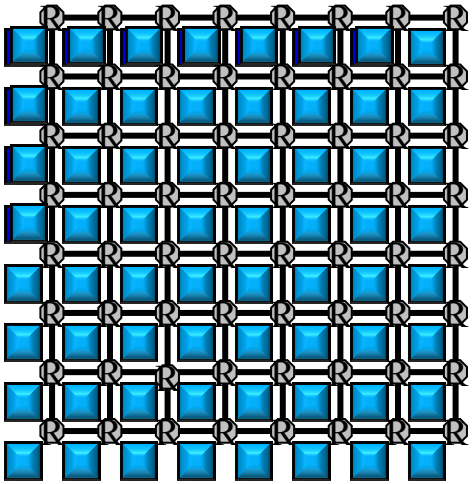
- **NoC Advantages**
  - Scalability
  - Modularity
  - High level abstraction



3x3 array

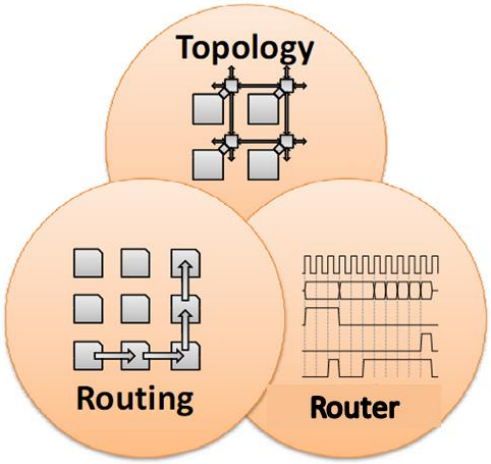


5x5 array



8x8 array

- **NoC Challenges**
  - Design Efficient (low area/power) Routers
  - Effective Hierarchical Topologies
  - High-throughput Routing Algorithms



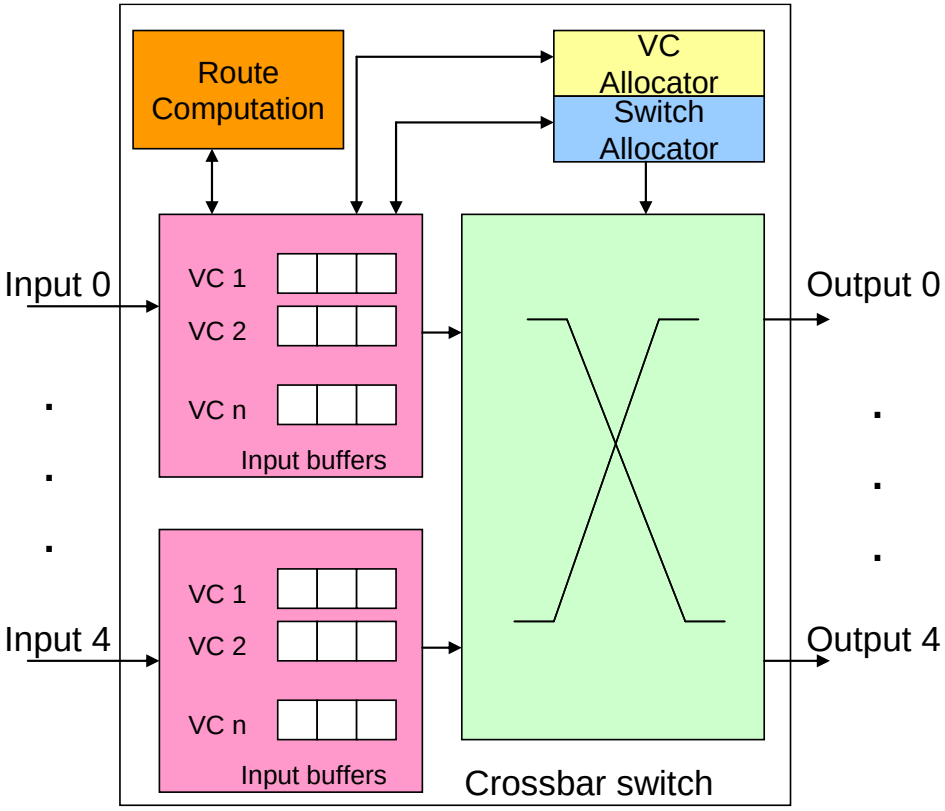
NoC Concepts

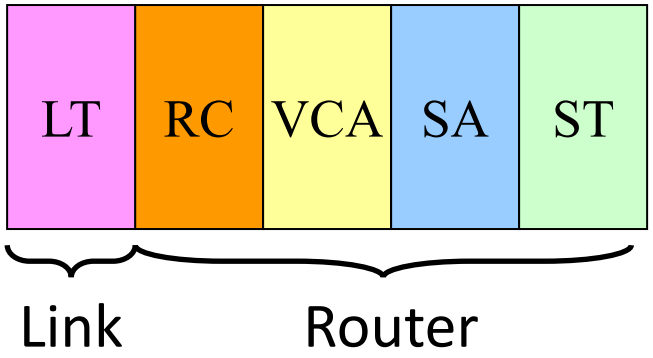
- NoC Architecture (a.k.a. NoC router)
  - *Combinational and sequential components*
- Topology
  - *How the nodes are connected together*
- Routing Algorithms
  - *Path selection between a source and a destination node in a particular topology*
- Switching Mechanisms
  - *Allocation of network resources (bandwidth, buffer capacity, …) to information flows*
- Flow control Mechanisms
  - *How the downstream node communicates forwarding availability to the upstream node*

NoC Concepts

- NoC Architecture (a.k.a. NoC router)
  - *Combinational and sequential components*
- Topology
  - *How the nodes are connected together*
- Routing Algorithms
  - *Path selection between a source and a destination node in a particular topology*
- Switching Mechanisms
  - *Allocation of network resources (bandwidth, buffer capacity, …) to information flows*
- Flow control Mechanisms
  - *How the downstream node communicates forwarding availability to the upstream node*

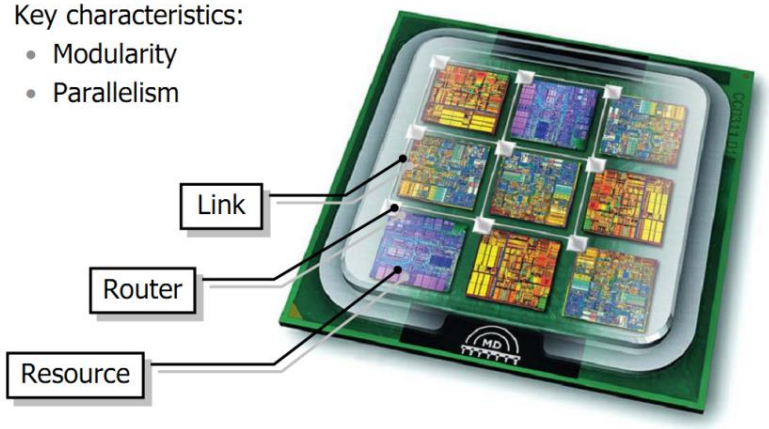
Generic NoC Router Architecture





By using prediction/speculation, pipeline can be made more compact

- Key characteristics:
  - Modularity
  - Parallelism



LT: Link Traversal

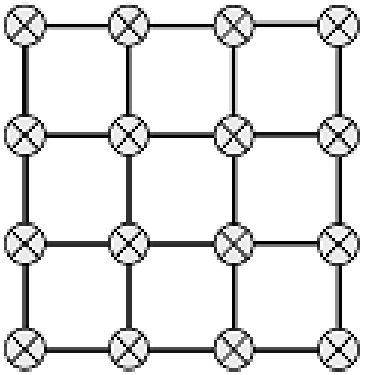
RC: Route Computation

VCA: VC Allocation

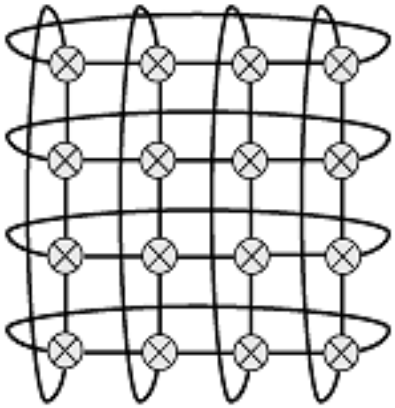
SA: Switch Allocation

ST: Switch Traversal.

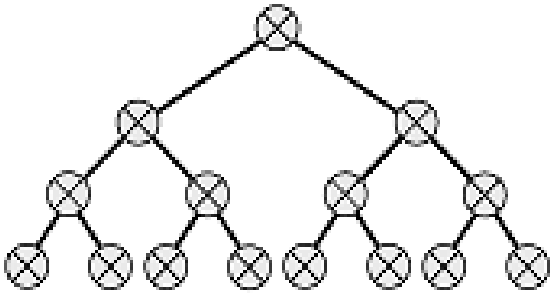
NoC Topologies



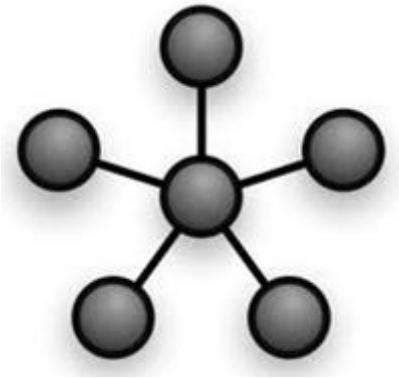
✓ Mesh



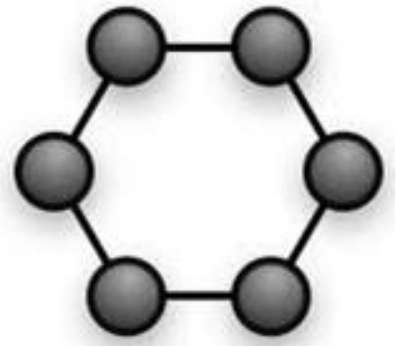
✓ Torus



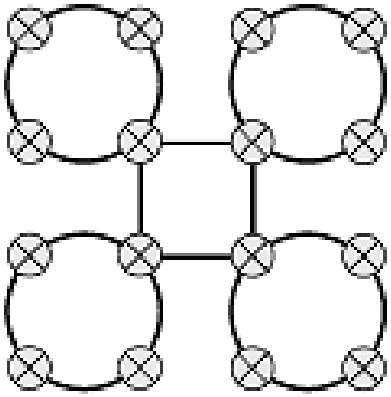
✓ Binary-Tree



✓ Star



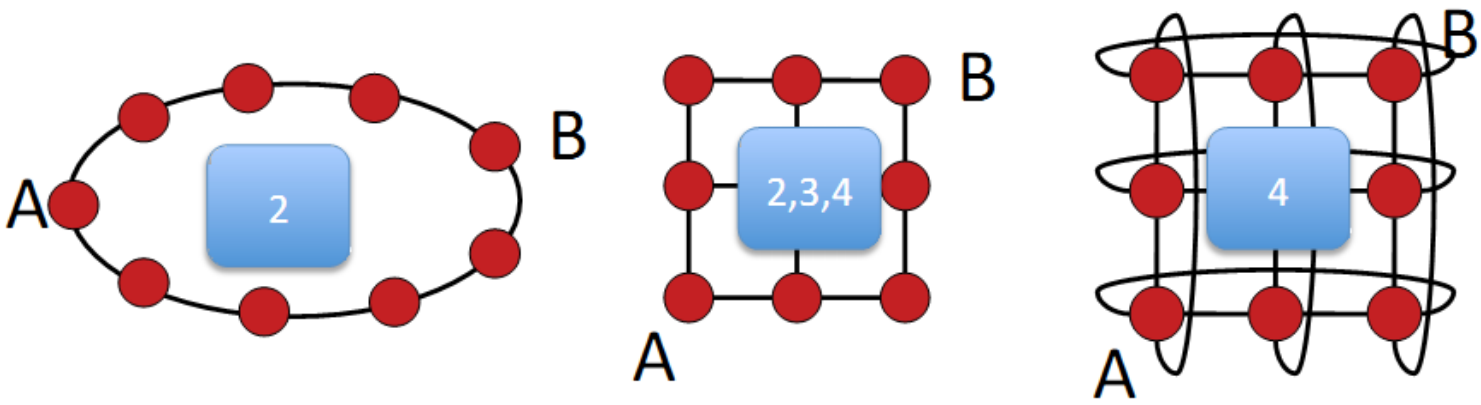
✓ Ring



✓ Hybrid

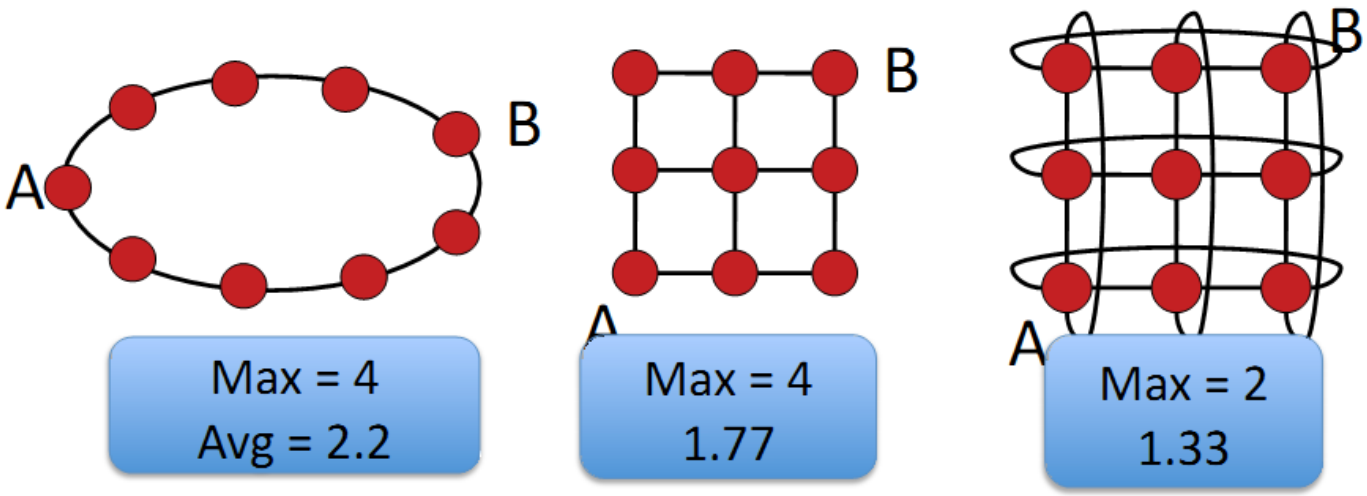
Abstract Metrics: Degree

- Switch Degree: number of links at a node
  - Proxy for estimating **cost**
    - Higher degree requires more links and port counts at each router

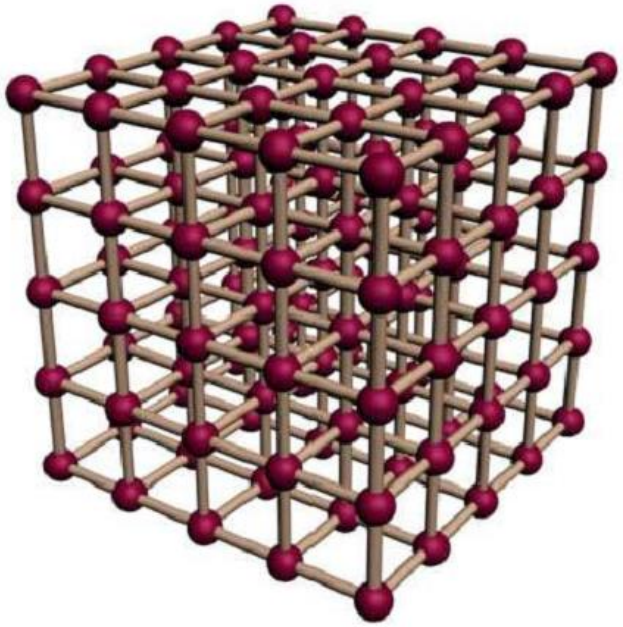


Abstract Metrics: Hop Count

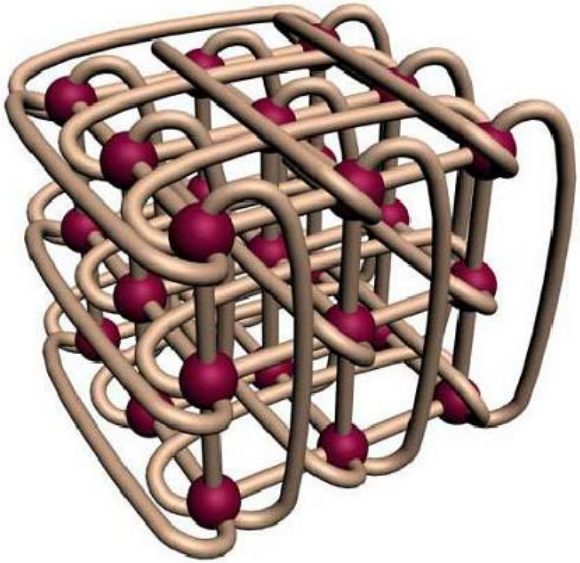
- Hop Count: number of hops a message takes from source to destination
  - Simple, useful proxy for network **latency**
    - Every node, link incurs some propagation delay even when no contention



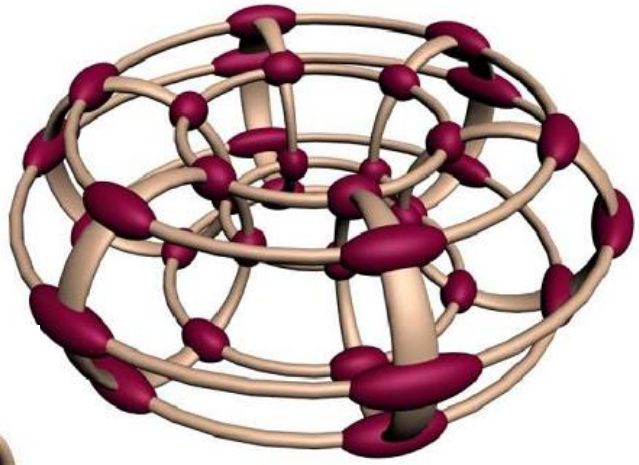
Unconventional NoC Topologies



3-dimensional 5-sided array



3-ary 3-cube

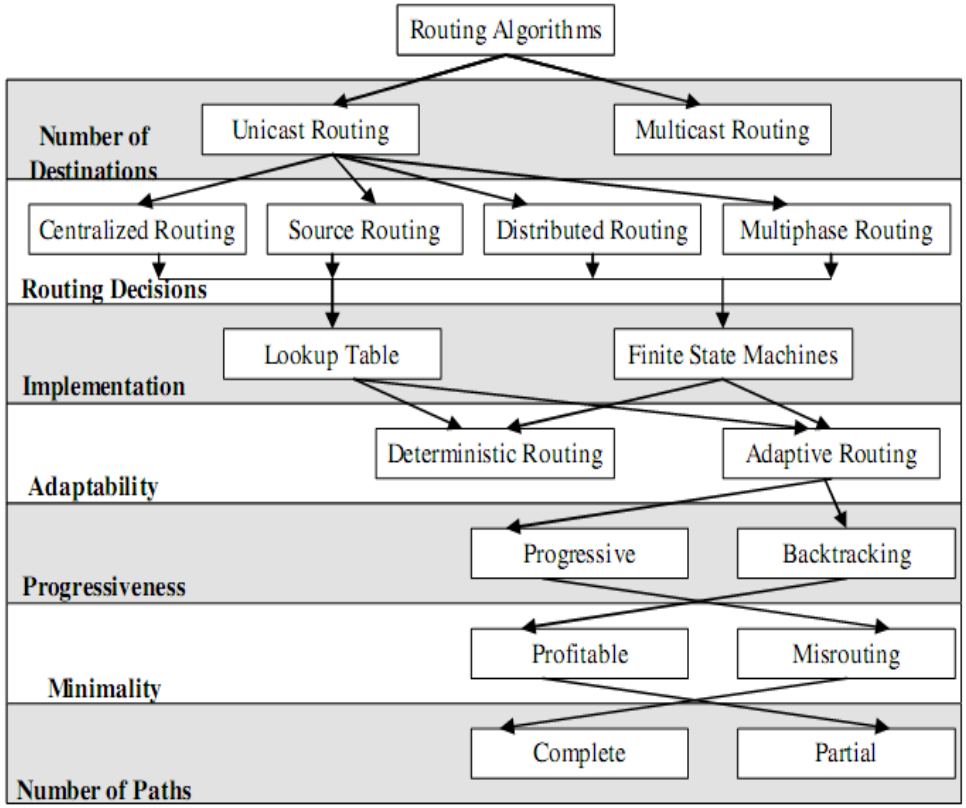


6x6 torus

NoC Concepts

- NoC Architecture (a.k.a. NoC router)
  - *Combinational and sequential components*
- Topology
  - *How the nodes are connected together*
- Routing Algorithms
  - *Path selection between a source and a destination node in a particular topology*
- Switching Mechanisms
  - *Allocation of network resources (bandwidth, buffer capacity, …) to information flows*
- Flow control Mechanisms
  - *How the downstream node communicates forwarding availability to the upstream node*

Taxonomy for Routing Algorithms



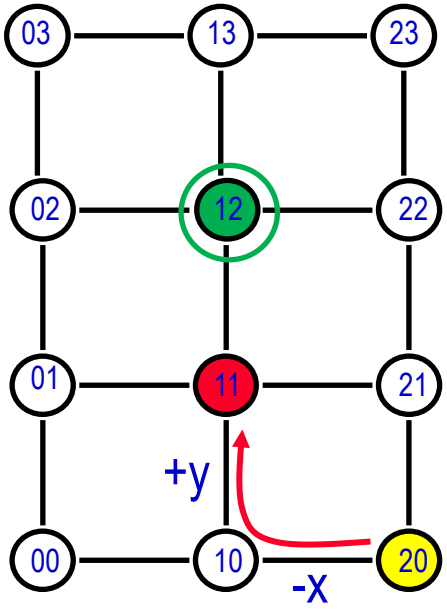
- A routing algorithm aims to **balance traffic loads** to maximise throughput of data packets

- Routing algorithm **determines the path selection** by a packet to reach its destination

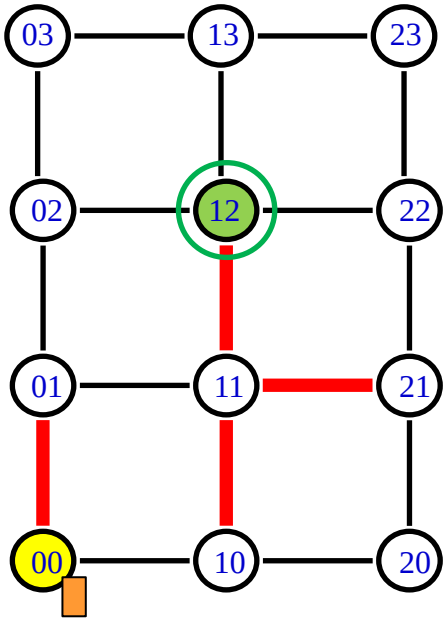
Ulster University

Static vs. Dynamic Routing

- **Static routing**: **fixed paths** are used to transfer data between source & destination
  - does not take into account current state of the network
  - easy to implement, since very little additional router logic is required



- **Dynamic routing**: routing **decisions are made according to the current state** of the network
  - considers factors such as availability and load on links
  - able to better distribute traffic in a network
  - more resources needed to monitor network state and dynamically change routing paths



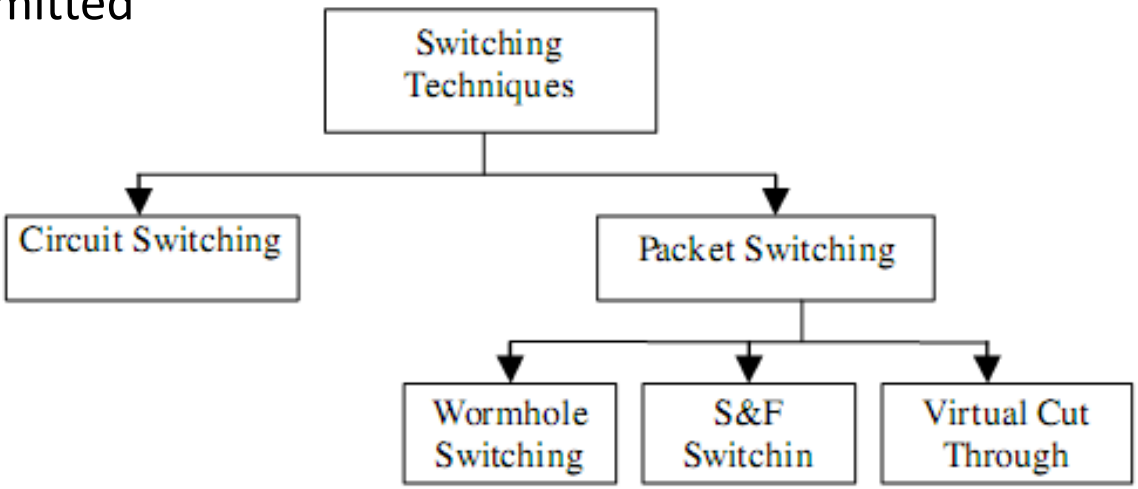
Ulster University

NoC Concepts

- NoC Architecture (a.k.a. NoC router)
  - *Combinational and sequential components*
- Topology
  - *How the nodes are connected together*
- Routing Algorithms
  - *Path selection between a source and a destination node in a particular topology*
- Switching Mechanisms
  - *Allocation of network resources (bandwidth, buffer capacity, …) to information flows*
- Flow control Mechanisms
  - *How the downstream node communicates forwarding availability to the upstream node*

Switching Techniques

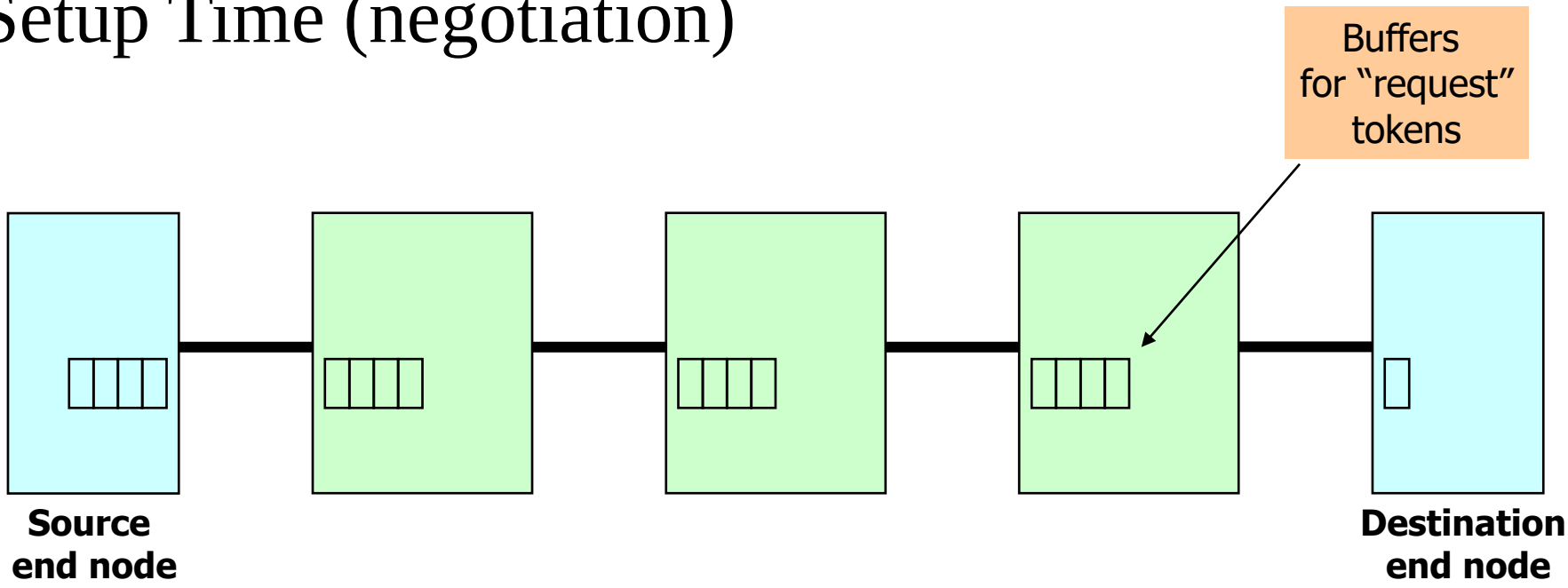
- A switching technique determines the way a packet is transmitted across the network…



- The Switching techniques can be classified based on…
  - **Circuit switched networks** **reserve a physical path** before transmitting the data packets.
  - **Packet switched networks** transmit the packets without reserving the entire path.

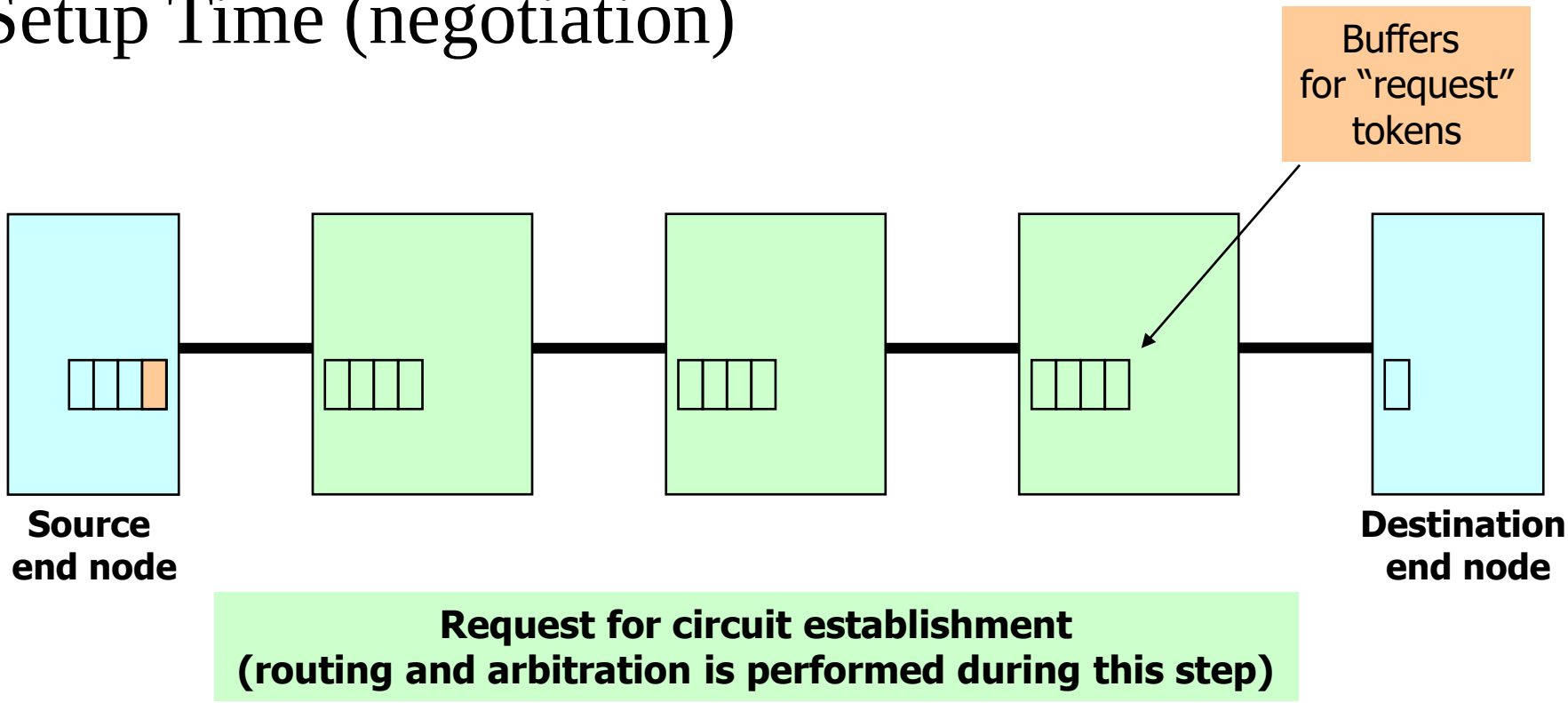
Circuit Switching

- Setup Time (negotiation)



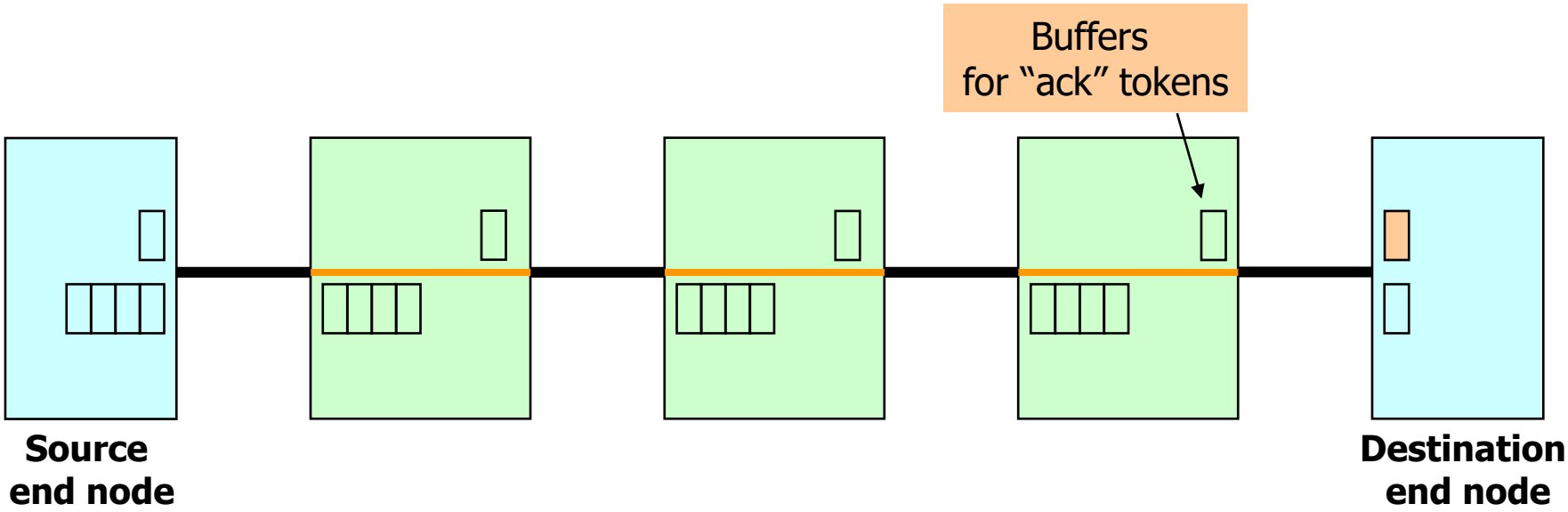
Circuit Switching

- Setup Time (negotiation)



Circuit Switching

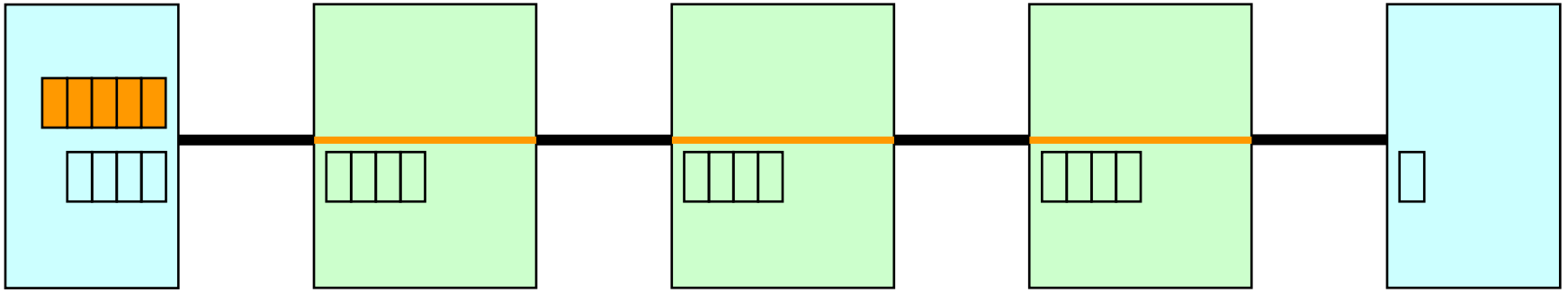
- Setup Time (negotiation)



Acknowledgment and circuit establishment
(as token travels back to the source, connections are established)

Circuit Switching

- Data Transmission



Request for circuit establishment

Acknowledgment and circuit establishment

Packet transport
(neither routing nor arbitration is required)

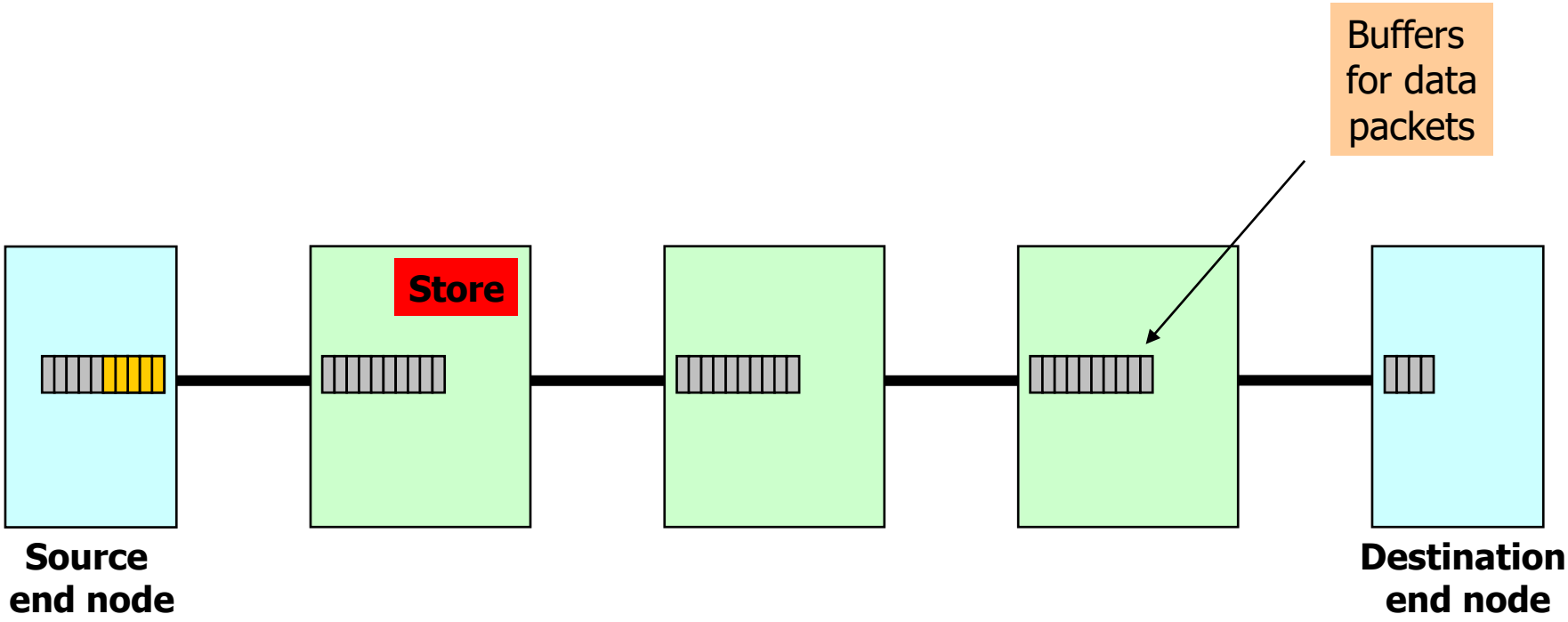
Packet Switching

- Packets are transmitted from source and make their way independently to receiver
  - possibly along different routes and with different delays

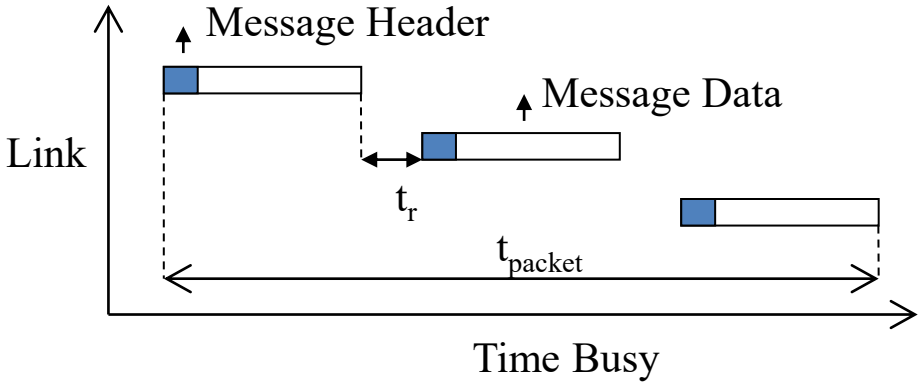
- Zero start up time, followed by a variable delay due to contention routers along packet path
  - QoS guarantees are harder to make

- Three main packet switching scheme variants:
  1) **Store and Forward (SAF) switching**
  2) **Virtual Cut Through (VCT) switching**
  3) **Wormhole (WH) switching**

Packet Switching (Store & Forward)

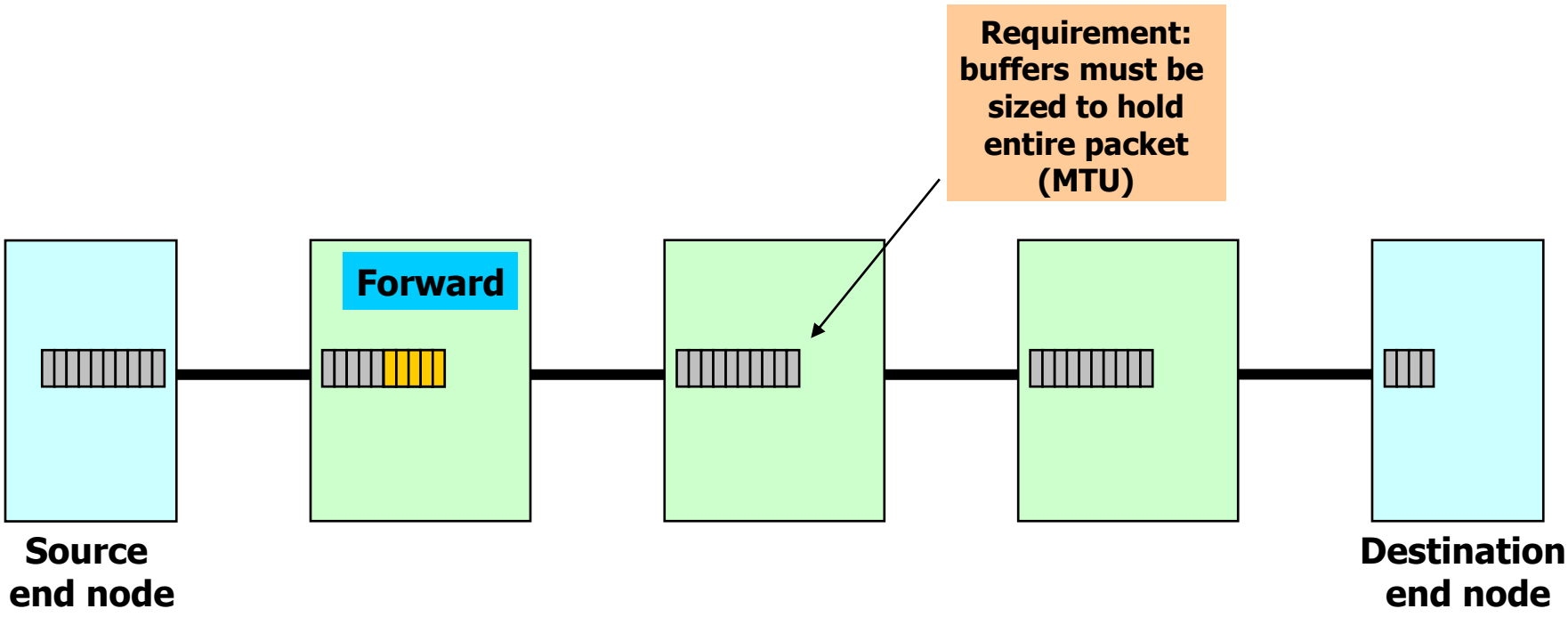


Packets are completely stored before any portion is forwarded

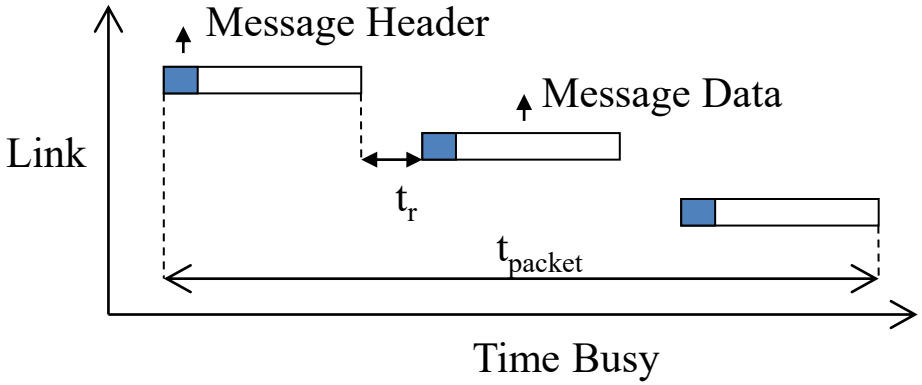


Ulster University

Packet Switching (Store & Forward)

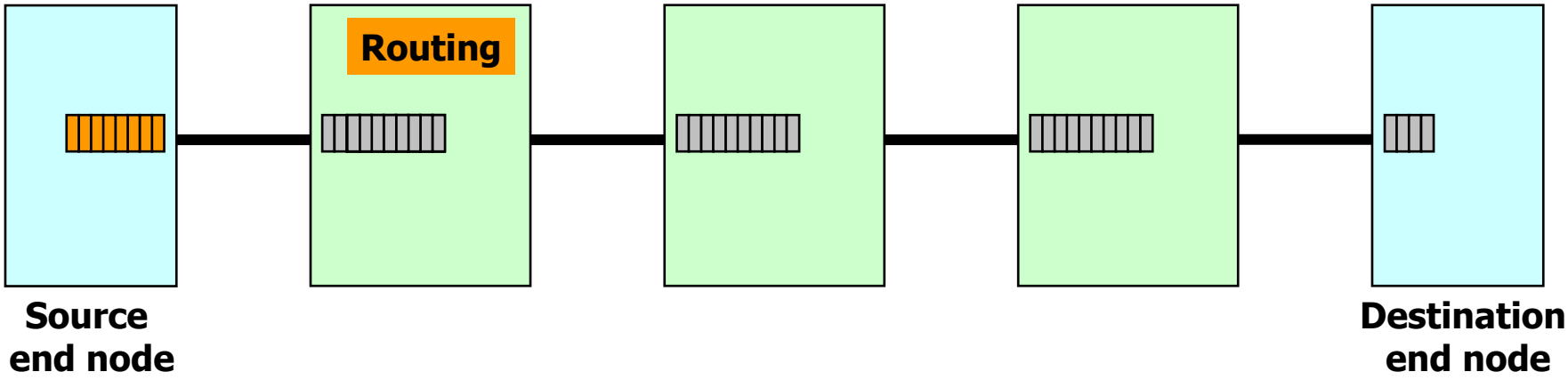


Packets are completely stored before any portion is forwarded



Ulster University

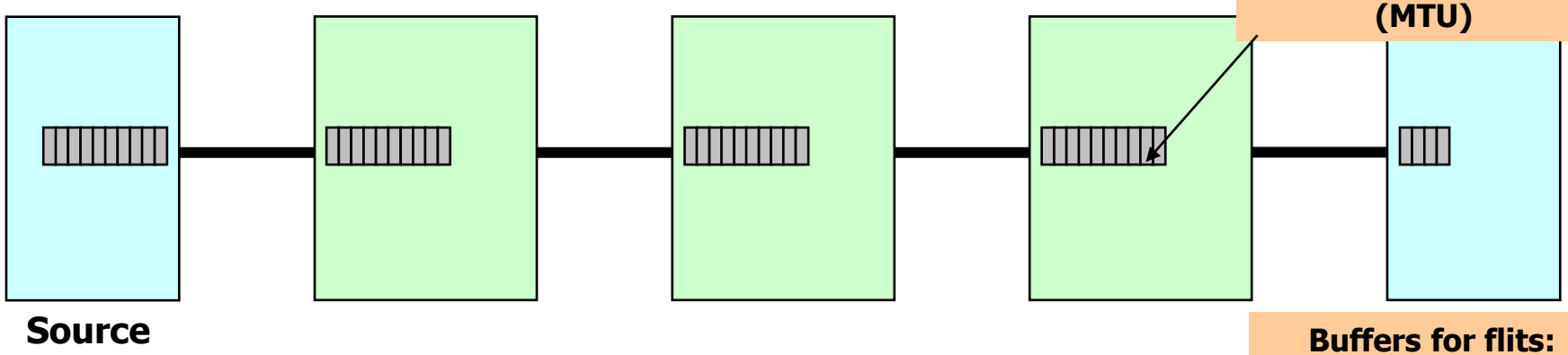
Packet Switching (Virtual Cut-through)



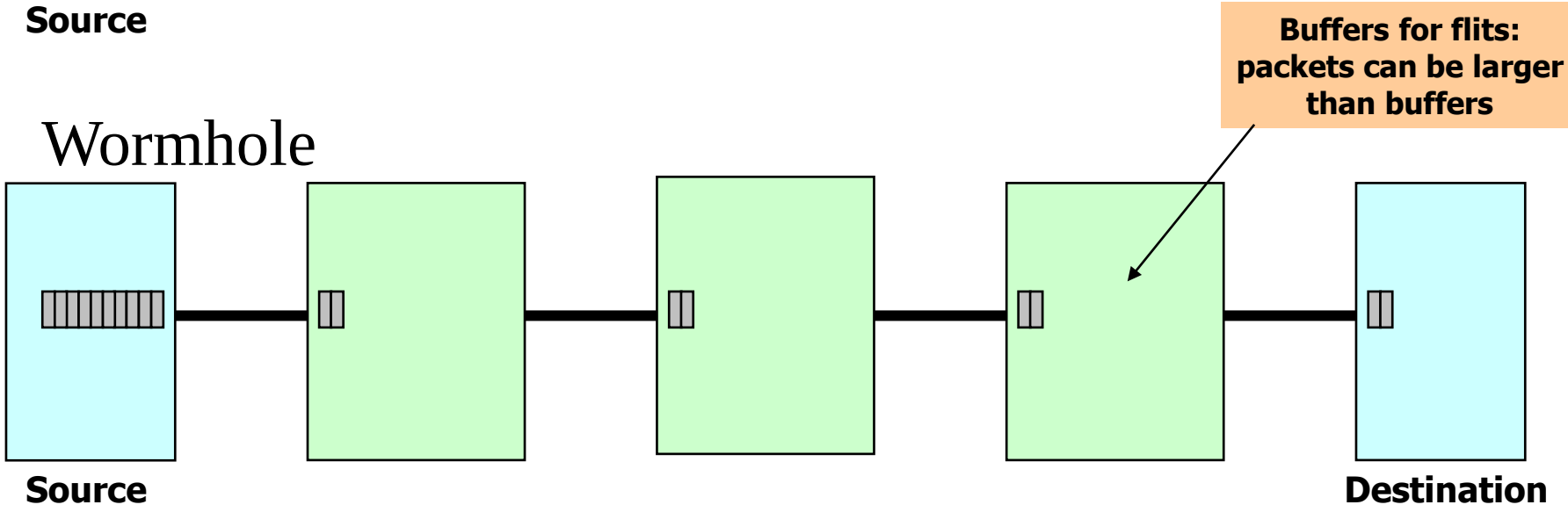
Portions of a packet may be forwarded ("cut-through") to the next switch before the entire packet is stored at the current switch

Virtual Cut Through vs. Wormhole

- Virtual Cut Through

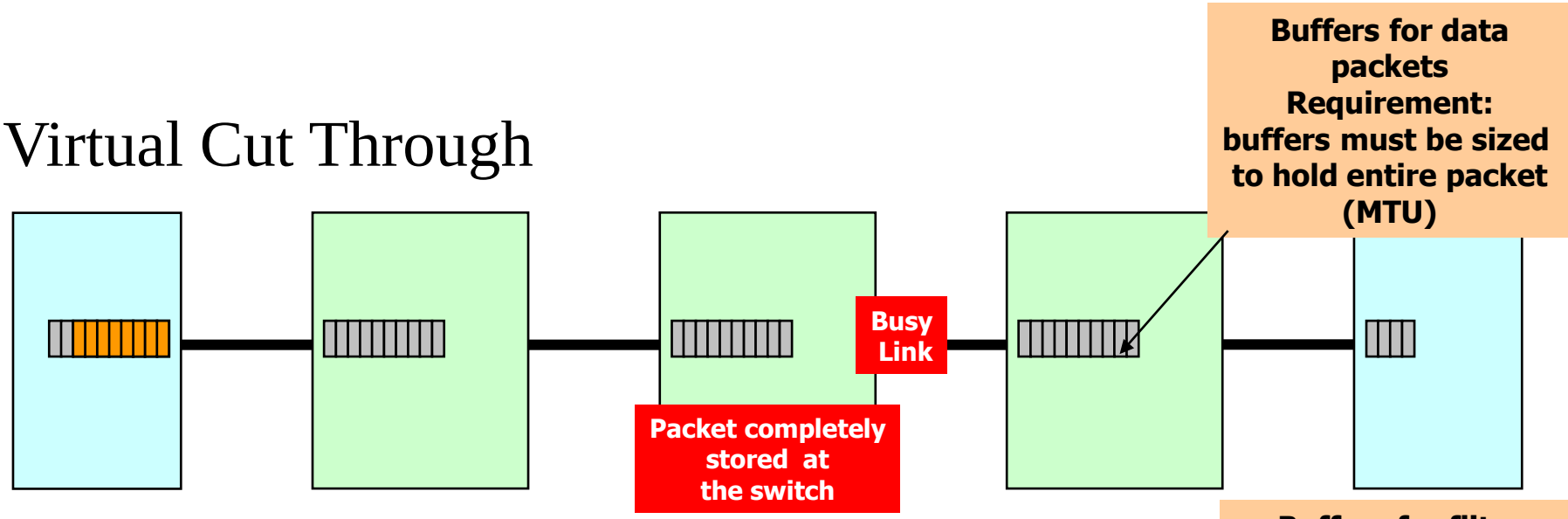


- Wormhole

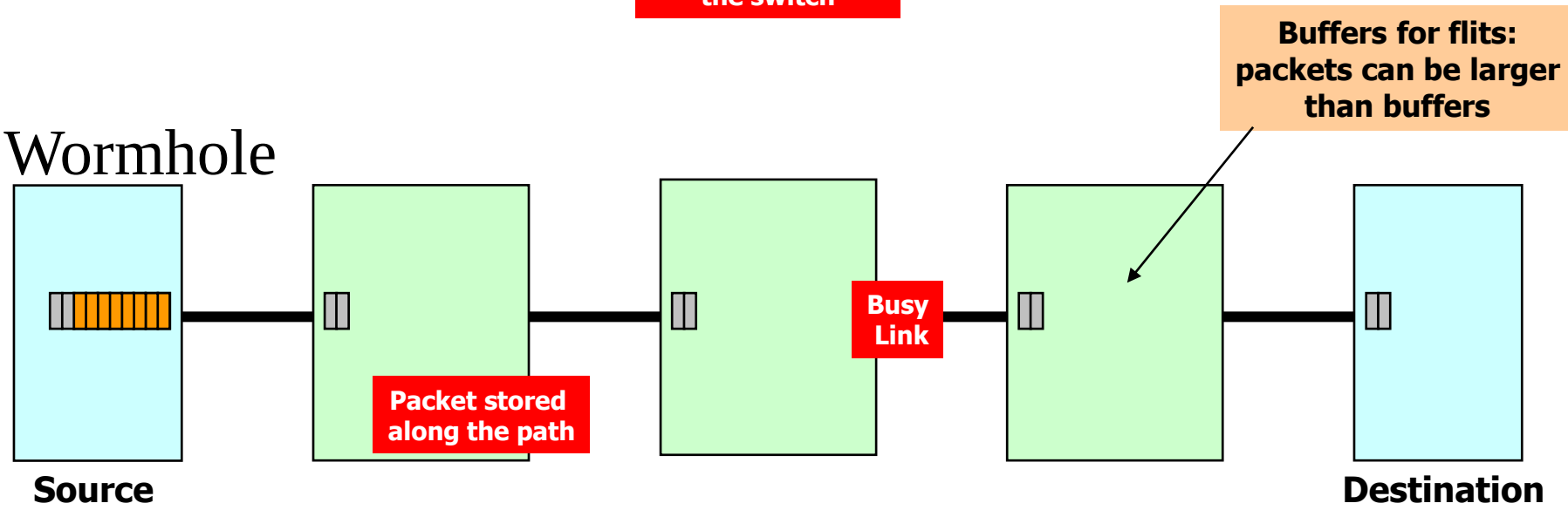


Virtual Cut Through vs. Wormhole

- Virtual Cut Through

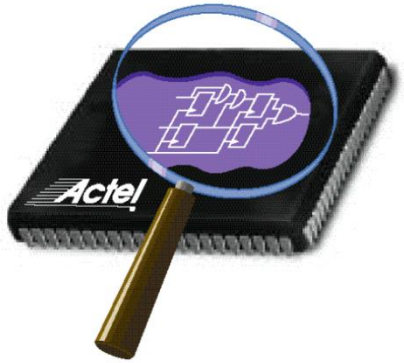


- Wormhole



Evaluation Metrics

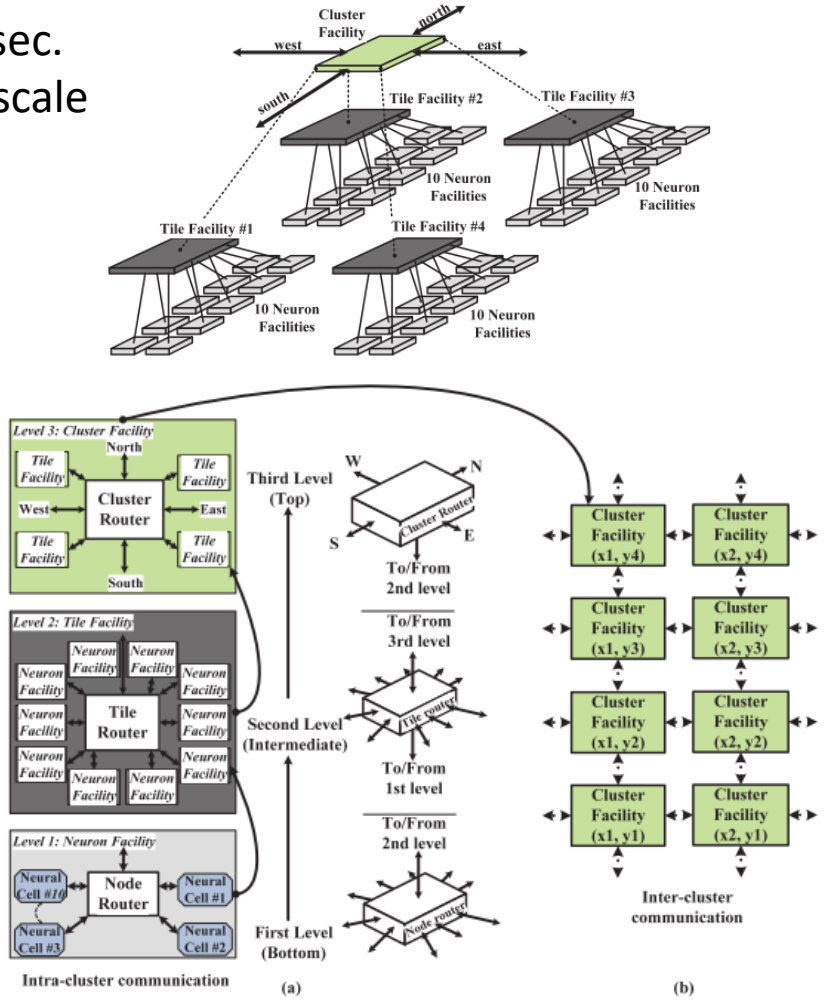
- ✓ **Throughput** ➔ frequency operation, number of pipeline stages, etc…
- ✓ **Latency** ➔ routing algorithm, switching technique, packet format, etc…
- ✓ **Area Utilisation** ➔ buffers, look-up tables, routing engine, CMOS technology etc…
- ✓ **Power Consumption** ➔ throughput spec + latency spec + area spec !!!



NoC for Brain Information Communication?

Hierarchical NoC (**H-NoC**) for spiking neural network: Neuron, Tile & Cluster levels.

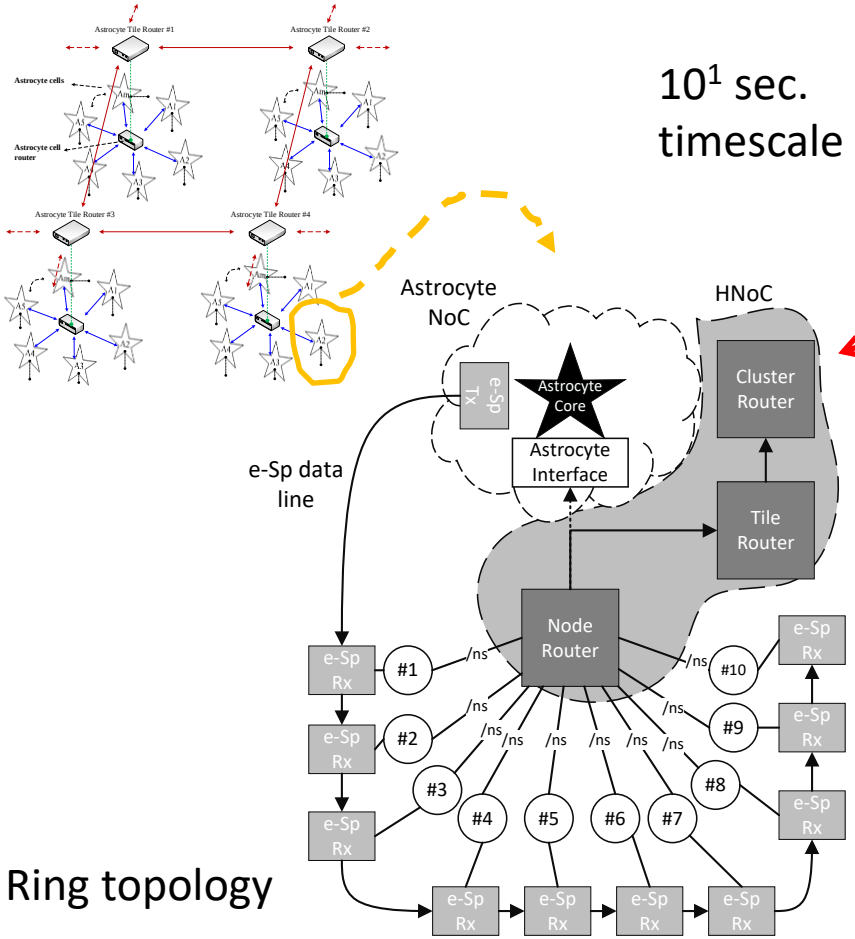
10^{-3} sec. timescale



Carrillo, S., Harkin, J., McDaid, et. al: ***"Scalable Hierarchical NoC Architecture for Spiking Neural Network Hardware Implementations"***. IEEE Trans. on Parallel and Distributed Systems. (2013)

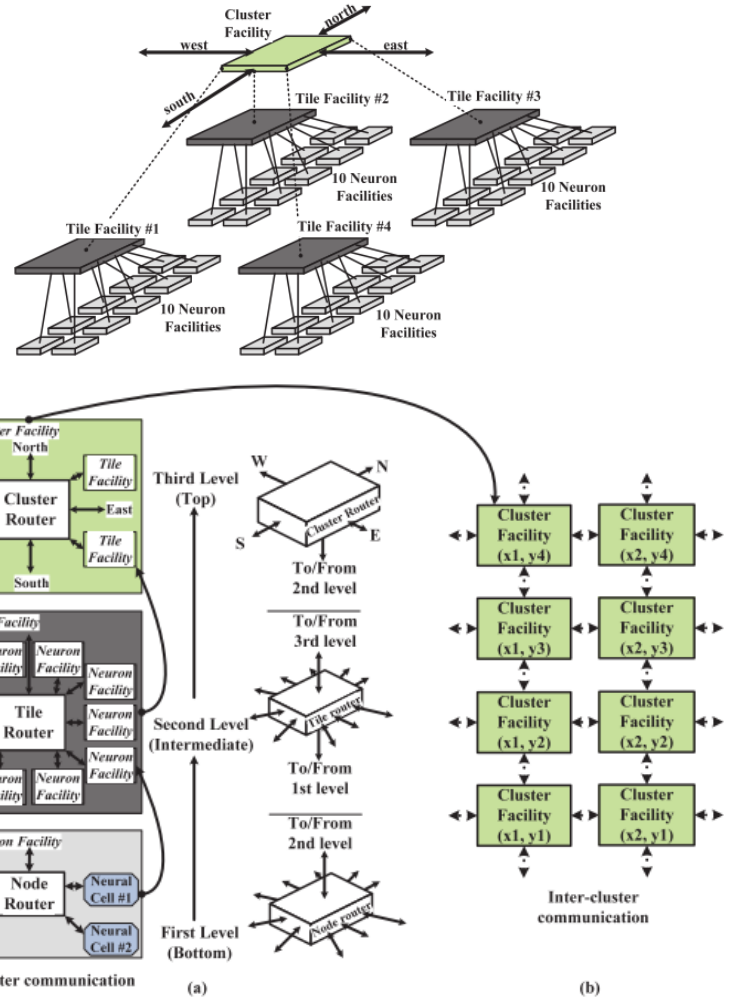
NoC for Brain Information Communication?

Hierarchical Astrocyte Network Architecture (**HANA**)



Martin G, Harkin J, McDaid L et al. "***On-chip communication for neuro-glia networks***" IET Computers & Digital Techniques, 12(4), pp. 130-138, 7 2018

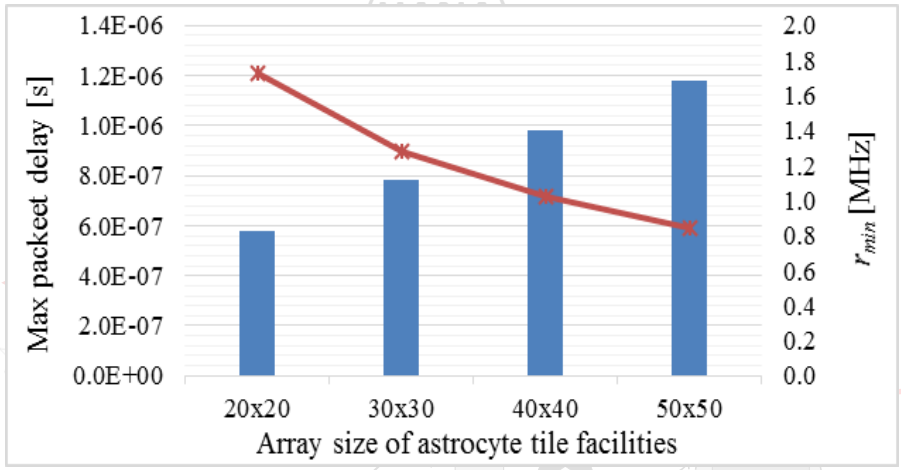
Hierarchical NoC (**H-NoC**) for spiking neural network: Neuron, Tile & Cluster levels.



Carrillo, S., Harkin, J., McDaid, et. al: ***"Scalable Hierarchical NoC Architecture for Spiking Neural Network Hardware Implementations"***. IEEE Trans. on Parallel and Distributed Systems. (2013)

NoC Works!

Hierarchical Astrocyte Network Architecture

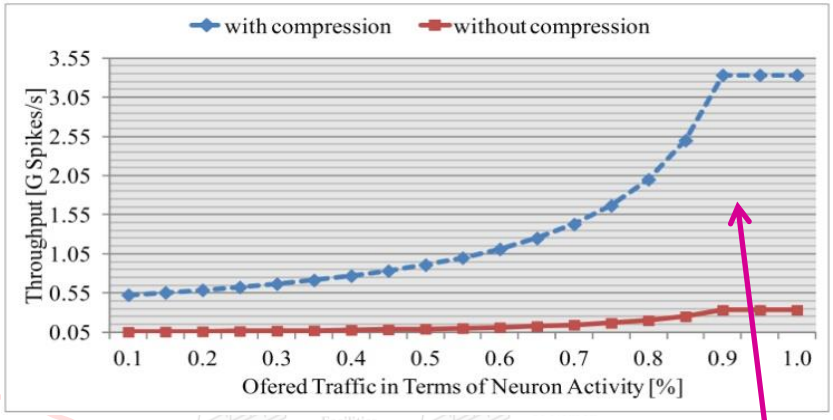


The approach		Topology	Area overhead (mm²)	
			Router	Device technology
[8] H-NoC		2D Mesh	0.056	90nm CMOS
[9] Many-core		2D Mesh	0.182	SAED 90nm
CG router [10]		2D Mesh	0.237	SAED 90nm
FG router [10]		2D Mesh	0.267	SAED 90nm
HANA	Astrocyte cell router	Star	0.024	SAED 90nm
HANA	Astrocyte tile router	2D Mesh	0.156	SAED 90nm

0.18mm²

Packet size has major influence on overall area overhead.

Hierarchical NoC (**H-NoC**) for spiking neural



Increased throughput under load testing

Project Reference	Quality of Service (QoS)	Congestion Mechanism	Throughput [Events/s]	Improvement
EMBRACE	Best Effort	Yes	3.33x10⁹	--
FACETS	Best/Guaranteed Effort	No	1.50x10⁹	2.2x
SpiNNaker	Best Effort	No	0.200x10⁹	16.5x

HNoC router performance

[8]. Carrillo, S., Harkin, J., McDaid, *et. al:* Advancing Interconnect Density for Spiking Neural Network Hardware Implementations using Traffic-aware Adaptive Network-on-Chip Routers. Neural Networks. 33, 42–57 (2012).

[9]. Liu, J., Harkin, J., Li, Y., Maguire, L.: Online Traffic-Aware Fault Detection for NoC. Journal of Parallel and Distributed Computing. 74, 1984–1993 (2014).

[10]. Liu, J., Harkin, J., Li, Y., Maguire, L.P.: Fault Tolerant Networks-on-Chip Routing with Coarse and Fine-Grained Look-ahead. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems. 35, 260–273 (2016)

for Astrocyte-Neuron Network Hardware", 25th Inter. Conf. on Artificial Neural Nets (2016)

Carrillo, S., Harkin, J., McDaid, et. al: *"Scalable Hierarchical NoC Architecture for Spiking Neural Network Hardware Implementations"*. IEEE Trans. on Parallel and Distributed Systems. (2013)

Reliability Challenge

Brain-inspired Hardware

Brain-inspired Information Processing

Reliability via Self-repair (the concept)

Building Self-repairing Hardware

On-chip Communication Challenge and NoC Solution

The Opportunities

What is Next?

- Major potential gains in the provision of **adaptive systems** that provide **high-levels of reliability**.
- **New generation of engineers** required to think differently!
- Exploring beyond information processing brings **new models of reliable computation**.

- **Next big wave** in how we design modern embedded systems.

EFUTURES

The UK Landscape in Artificial Intelligence and Brain-Inspired Computing Hardware: the potential for establishing a new Centre of Excellence

August 2021

GIACOMO INDIVERI,
UNIVERSITY OF ZURICH AND ETH ZURICH

WALID NAJJAR,
UNIVERSITY OF CALIFORNIA, RIVERSIDE

Generative Pre-trained Transformer (GPT)

large amounts of data [1]. The methods used, however, to develop the latest and most powerful networks, such as GPT-3 [2] require thousands of petaflop-days to train (over 10^{23} floating-point operations). It has been estimated that the multiple training sessions used to develop GPT-3 required "9,998 total days" worth of GPU time (more than 27 GPU-years). Taking all these runs into account, the researchers estimated that building this model generated over 35 tonnes of carbon dioxide emissions: more than the average American adult will produce in two years."[3]

Report supported by the EPSRC-funded network *eFutures 2.0*

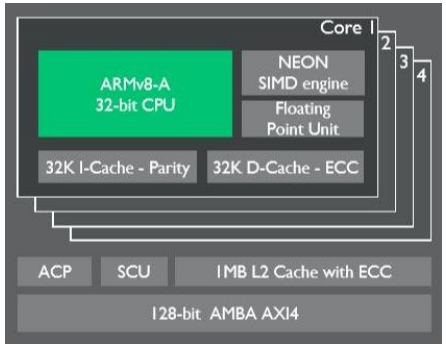
Opportunities

- Build new computers that allow Neuroscientists to **understand brain diseases** (e.g. Parkinson's), pharmaceutical companies **design new drugs**.

- **Harsh environments**: Robotics in nuclear industry, Wireless sensor networks in environmental disasters; forest fire, structural health monitors (earthquake damage).

- **Remote locations**: Border security, space, deep-sea

- **Error resilience** in computing: Plastic electronics, robotics, modern microprocessors



Ulster University

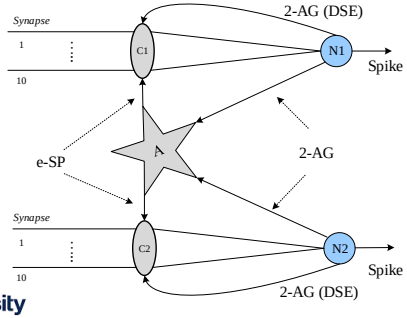
Opportunities

- Build new computers that allow Neuroscientists to **understand brain diseases** (e.g. Parkinson's), pharmaceutical companies **design new drugs**.

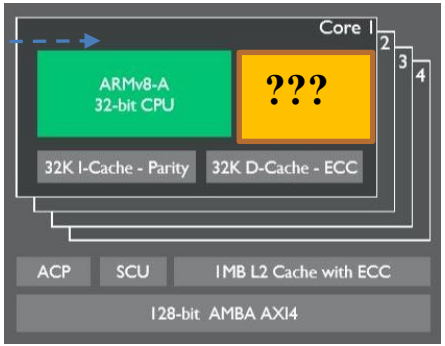
- **Harsh environments**: Robotics in nuclear industry, Wireless sensor networks in environmental disasters; forest fire, structural health monitors (earthquake damage).

- **Remote locations**: Border security, space, deep-sea

- **Error resilience** in computing: *Plastic electronics*, robotics, modern microprocessors.



Augment existing architectures – not replace!



Ulster University

Take Home Message

- **Embedded AI** - many challenges remaining
- **Astrocyte-Neuron systems** - Disruptive approach
- Major potential gains in the provision of **adaptive systems** that provide **high-levels of reliability**

Computational Neuroscience and Neuromorphic Engineering Team

Modelling: Prof. L McDaid, Dr J Wade Dr B Flanagan (M Toman, A Reza, D Harkin)

Hardware: Dr J Liu, M McElholm, Dr G Martin A Javid (D Simpson, N Hamilton, K Madden)

Ulster University

HFSP Human Frontiers Science Programme

EPSRC **SPANNER** project (EP/N00714X/1)

Thank You

Modelling: Prof. L McDaid, Dr J Wade Dr B Flanagan (M Toman, A Reza, D Harkin)

Hardware: Dr J Liu, M McElholm, Dr G Martin A Javid (D Simpson, N Hamilton, K Madden)

Ulster University

Human Frontiers Science Programme

EPSRC **SPANNER** project (EP/N00714X/1)