Aalto University

Neuromorphic Computer Architectures

Aleksi Korsman

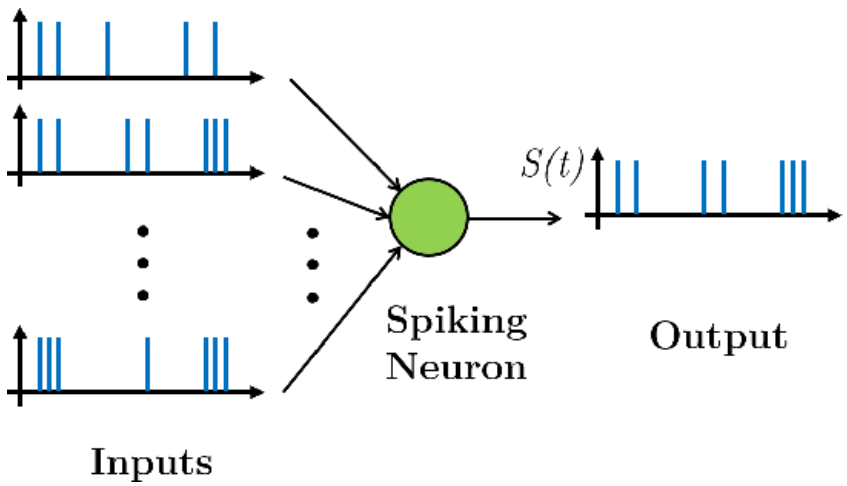
10.5.2023

Outline

1. Spiking Neural Networks
2. Neuromorphic chip building blocks
3. Example architectures

Spiking Neural Networks

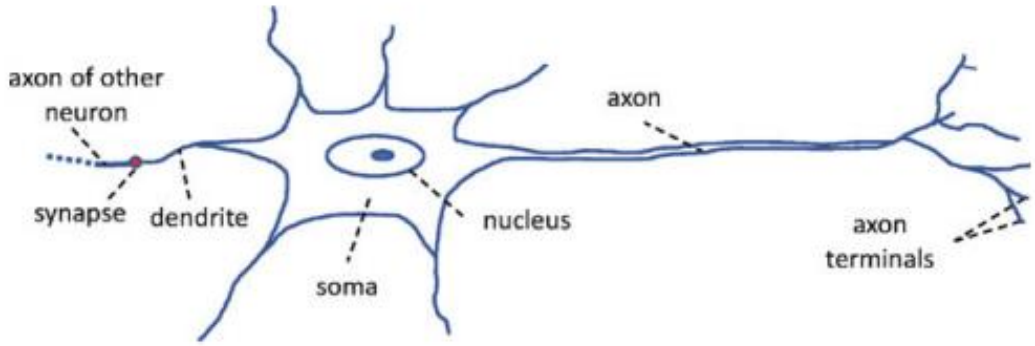
- A model of computation with neurons as the basic processing element
- At some instant of time, one or more neurons might send out impulses, or spikes, to connected neurons
- Therefore, SNNs incorporate time as an explicit dependency
- Connections between neurons are called synapses, and the travelling time can be non-zero
- Neurons integrate the incoming spikes to its membrane potential, and fire a spike once a threshold is reached



Anwani, Navin and Bipin Rajendran. "Training Multilayer Spiking Neural Networks using NormAD based Spatio-Temporal Error Backpropagation." ArXiv abs/1811.10678 (2018): n. pag.

Spiking Neural Networks

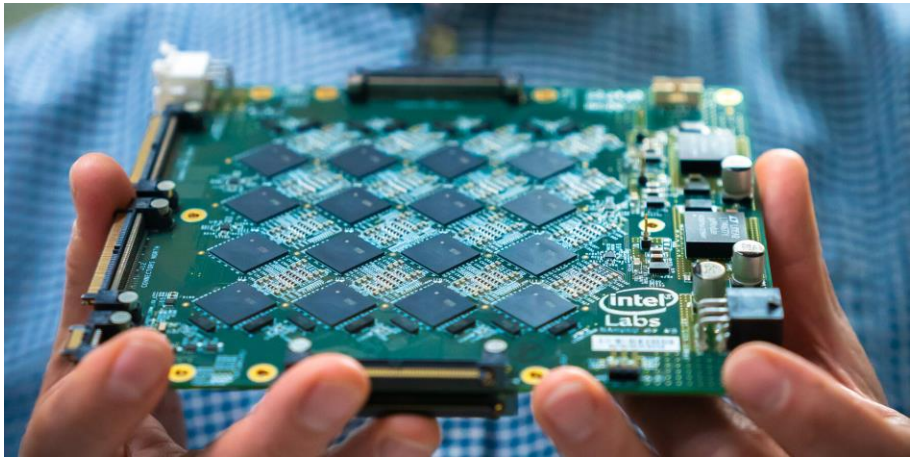
- Try to mimic the behavior of biological neural networks
- Neuromorphic computing uses SNNs
- Important terms include:
  - Axon
  - Dendrite
  - Synapse
  - Soma & Nucleus



Abdallah, A. B., & Dang, K. N. (n.d.). *Neuromorphic computing principles and organization*. SpringerLink.

Neuromorphic chip building blocks

- Neuron
- Synapse (Network)
- Routing



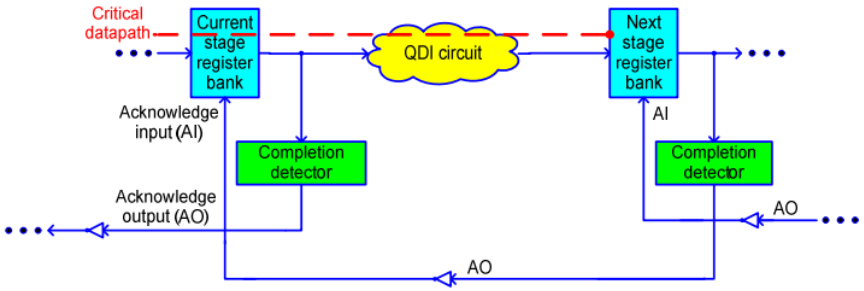
Tim Herman / Intel Corporation

Design Methodology – Event driven

- Neuromorphic chips are often event driven
- Neuron firing rate is relatively low, max 10s of Hz for real-time applications
- Most of the circuit is idle during any given time
- If all circuit elements were clocked, power consumption would be very high, and clock gating is difficult to manage separately for 1 million+ elements
- Event driven - bits are flipped only when computation is required
- This is mostly achieved with asynchronous circuit design

Design methodology - Quasi Delay-Insensitive Async Circuits

- Prevalent in large neuromorphic systems
- QDI only makes timing assumptions on the propagation delay of signals that fan-out to multiple gates
- Makes no timing assumption on gate delays
- Input register bank consists of 2-input C-elements
- C element outputs 1 if both its inputs are 1, 0 if both its inputs are 0, otherwise maintains output
- Computation on a QDI circuit is initiated with a handshake
- M-of-n encoding scheme: 1-of-n is most common, which is essentially one-hot encoding
- Encoding includes a spacer value where all bits are 0. This must be inserted between all computations.



Balasubramanian, P.; Mastorakis, N.E. Quasi-Delay-Insensitive Implementation of Approximate Addition. Symmetry 2020, 12, 1919. https://doi.org/10.3390/sym12111919

Neuron model

- A model is required for a circuit implementation
- Leaky Integrate-and-fire (LIF) model is most common
- LIF consists of an integrator with leak and a threshold

$$V_j(t) = V_j(t-1) + \sum_{i=0}^{n} A_i(t) w_{i,j} - \lambda_j$$

  - V_j is membrane potential for neuron j
  - A_i is 1 if there was an incoming spike from neuron i, 0 otherwise
  - $w_{i,j}$ is synaptic weight between neuron i and j
  - λ_j is leak
- If V_j is larger than threshold α_j, V_j is reset to R_j , and neuron j generates a spike

Other neuron models

- LIF is hardware-friendly, but not biologically exact
- More biologically accurate models exist
- Hodgkin-Huxley model

$$\frac{dv}{dt} = \left(\frac{1}{C}\right)I - g_k(V - V_k) - g_{Na}(V - V_{Na}) - g_L(V - V_L)$$

- Izhikevich Model

$$\frac{dv}{dt} = 0.04v^2 + 5v + 140 - u + I$$

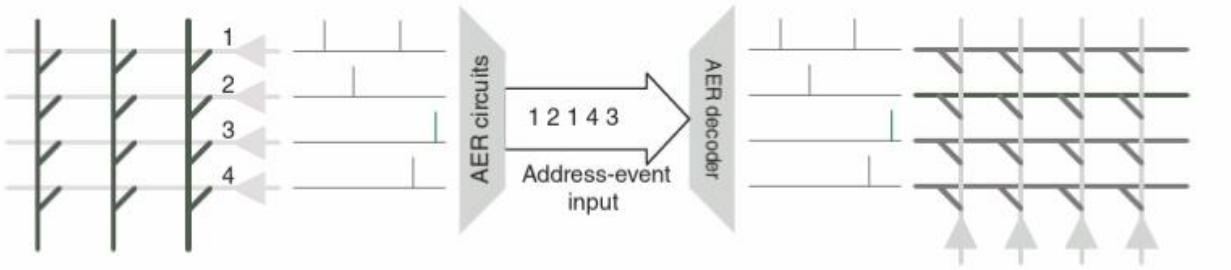
$$\frac{du}{dt} = a(bv - u) \qquad \begin{cases} v \leftarrow c \\ u \leftarrow u + d \end{cases} \quad if\, v \geq 30\,\mathrm{mV}$$

Synapses

- A neuron's axon (output) connects to another neuron's dendrite (input) via a synapse
- One neuron might connect to thousands of others
- However, neuron spiking rates are very low (tens of Hz) compared to the gate-delays of modern-process transistors
- Therefore, many architectures use time-multiplexing for interneuron communication
- Spike events are transmitted in packets that request access for shared communication resources
- Packets can be generated asynchronously or synchronously by using a global tick

Address-Event Representation

- Address-Event Representation is commonly used, where the address of the spiking neuron is broadcast over the network
- Address mapping is needed to know which source neuron n is connected to a specific destination axon a
- Mapping tables implemented with on- and off-chip memories are used



Event-Based Neuromorphic Systems, edited by Shih-Chii Liu, et al., John Wiley & Sons, Incorporated, 2015. ProQuest Ebook Central, https://ebookcentral.proquest.com/lib/aalto-ebooks/detail.action?docID=1895762.

Routing

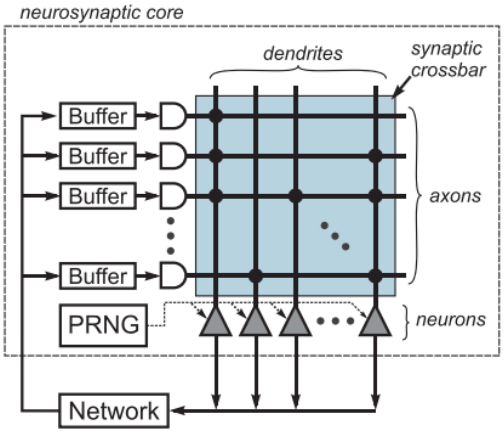
- There can be multiple communication buses that connect neighboring neuron clusters
- To maintain connectivity between all neurons, the spike packets need to be routed
- Simplest: Point-to-Point – all neuron clusters have connectivity to all other neuron clusters
- Rings and 1D arrays: distance- and identifier based addressing
- Meshes: 2D arrays. Distance-based addressing with dx and dy

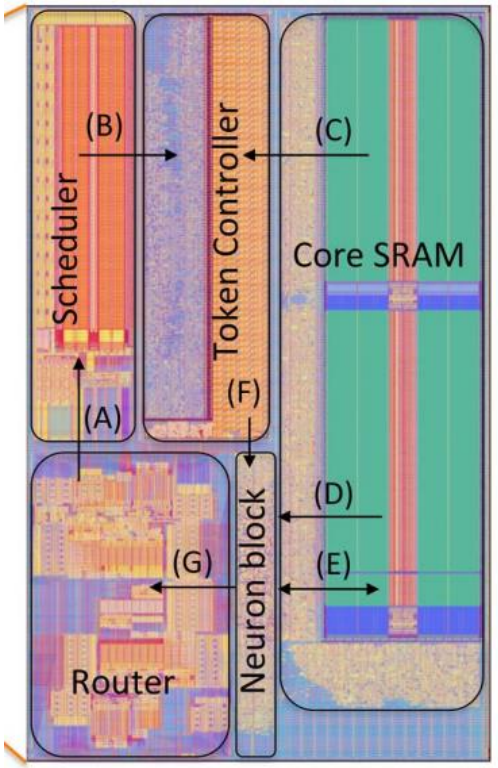
TrueNorth

- 4096 neurosynaptic cores
- Each core contains 256 neurons -> 1 million neurons
- Neurons use synchronous logic
- Otherwise, all logic is asynchronous
- SRAM memory stores parameters such as neuron connectivity, membrane potential, and leak
- 1 ms global tick
- Approximate power consumtion 65 mW

Neurosynaptic Core

- Calculates 256 neurons by time-division multiplexing
- Consists of:
  - Scheduler
  - Token controller
  - Core SRAM
  - Neuron
  - Router

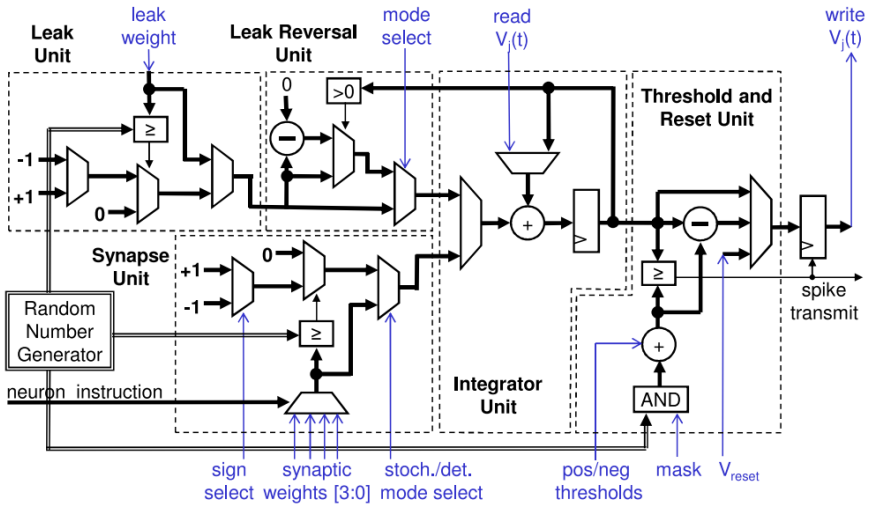




F. Akopyan et al., "TrueNorth: Design and Tool Flow of a 65 mW 1 Million Neuron Programmable Neurosynaptic Chip," in IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, vol. 34, no. 10, pp. 1537-1557, Oct. 2015, doi: 10.1109/TCAD.2015.2474396.

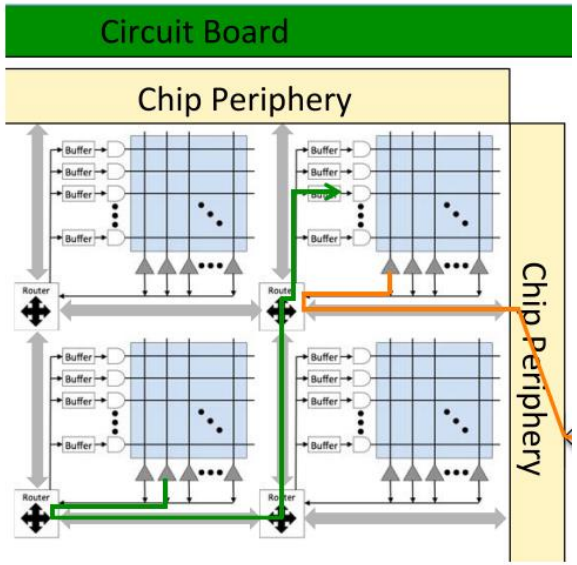
TrueNorth Neuron Block

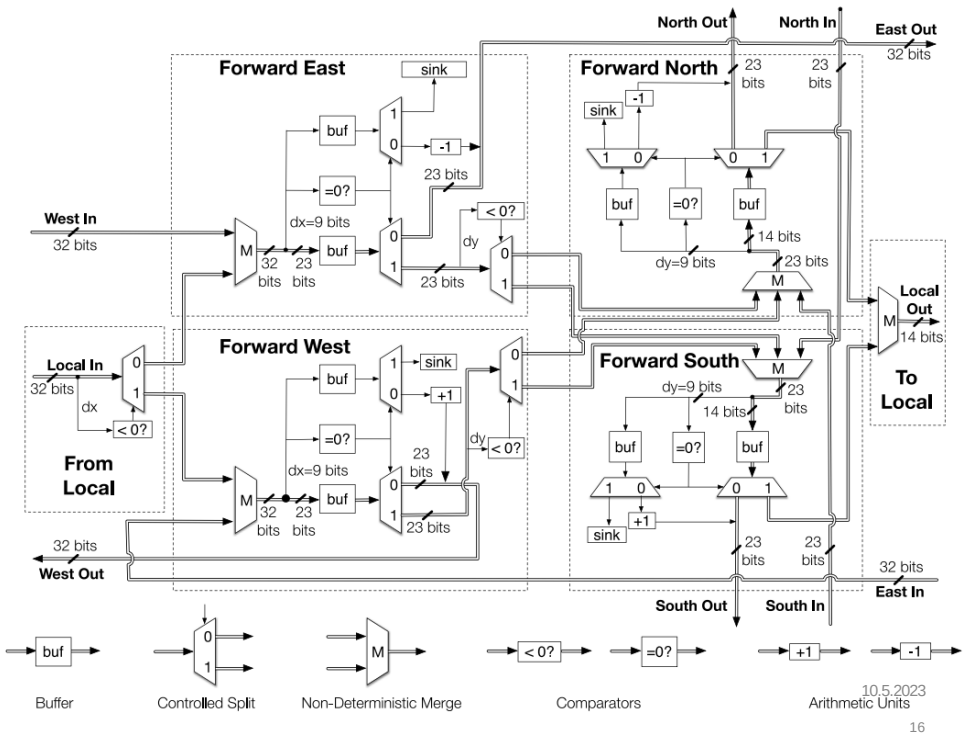
- Augmented integrate-and-fire
- Synchronous logic
- Event-driven – clock signal is generated by token controller only when a spike arrives
- Blue signals indicate input parameters fetched from SRAM



TrueNorth Router

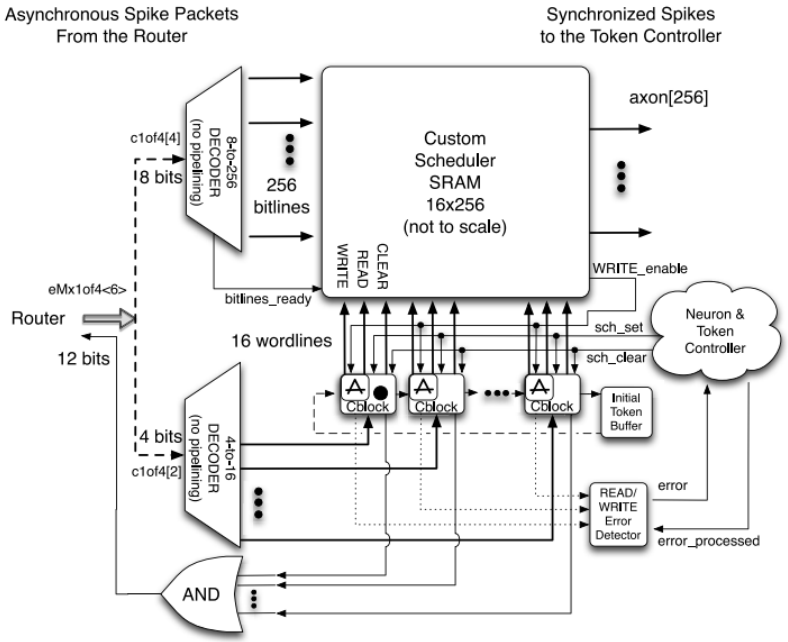
- Communicates with its own core and four neighboring routers in a 2D Mesh network
- Spikes are transmitted via spike packets that consist of:
  - Relative (dx, dy) address of the destination core
  - Destination axon index
  - Destination tick (when spike should be integrated)
- Depending on (dx,dy), the router will either
  - forward the packet to a neighboring core, or
  - send the packet to scheduler for decoding
- Packet is forwarded so that dx is first driven to zero, then dy
- Resources are allocated on a first-come first-serve basis
- Implemented using asynchronous QDI logic





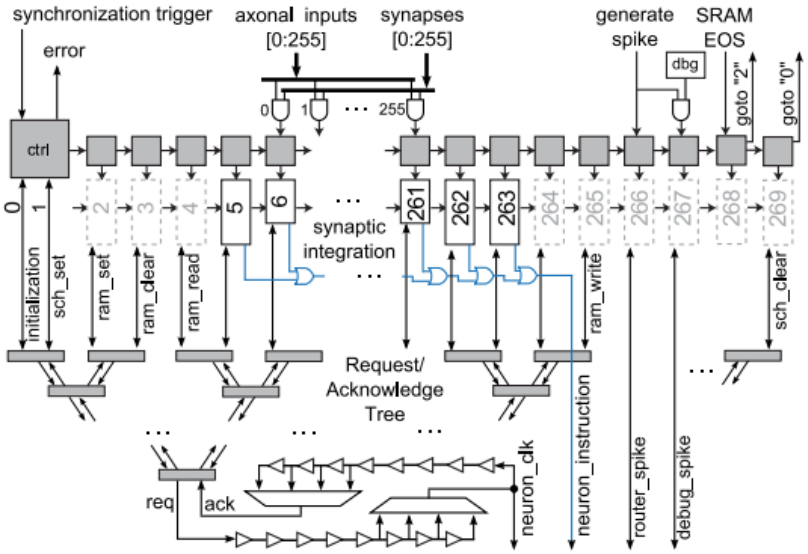
TrueNorth Scheduler

- Receives spike packet (without (dx,dy)) from router
- Delivers the incoming spike to the right axon at the proper tick
- Contains a 16 x 256-bit SRAM
  - 256 bitline corresponds to 256 axons
  - 16 wordlines correspond to 16 delivery ticks
- Writes to SRAM when it receives a spike packet
  - Decodes when (tick) and where (axon) a spike should be delivered
  - Writes 1 to the corresponding bit in the bitline
- In the beginning of a tick, the scheduler reads the corresponding wordline from SRAM and delivers it to token controller
- After that, the corresponding wordline is cleared
- Asynchronous logic



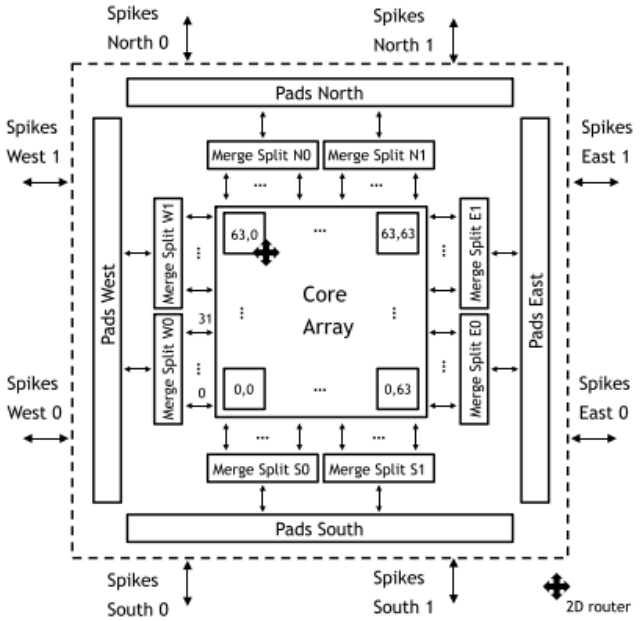
TrueNorth Token Controller

- At every tick, it requests 256 axon states, corresponding to the current tick, from the scheduler
- Next, it sequentially analyzes 256 neurons
- For each neuron, it reads a row of the core SRAM that stores synaptic connections, and combines this data with axon states coming from scheduler
  - If both axon state and synaptic connection is 1, the neuron block is activated
  - Leak is applied
  - Threshold is checked, and membrane potential reset if applicable
  - Membrane potential is written back to core SRAM
- Then, the neuron is evaluated, and a possible pending spike is sent to the router



TrueNorth Chip Periphery

- 64 bi-directional ports on each of the four edges of the 2D mesh boundary
- Corresponds to 8320 and 6272 wires on east/west and north/south edges, respectively
- Merge split blocks take 32 bi-directional ports and (de)multiplex that to one port
- Output of merge split block is connected to pad ring
- Spikes leaving the core array are tagged with their row and column
- Messages use bundled-data two-phase protocol
  - Data is sent on both rising- and falling edges of req signal
- Communication and logic is asynchronous



SpiNNaker

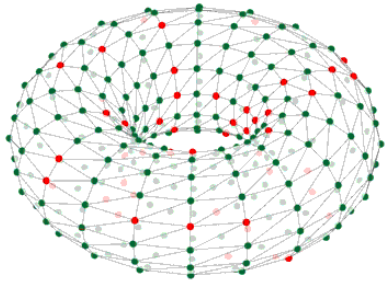
- University of Manchester
- General-purpose large-scale neuromorphic system
- 18 ARM processor nodes
- Specifically designed router for inter-chip communication
- One core is responsible for system management tasks
- 16 cores used for neuromorphic computation
- 1 extra core for spare
- Inter-chip communication is asynchronous
  - 'Globally asynchronous, locally synchronous'
- Approximate power consumption 25-36 W

Neurons and Synapses

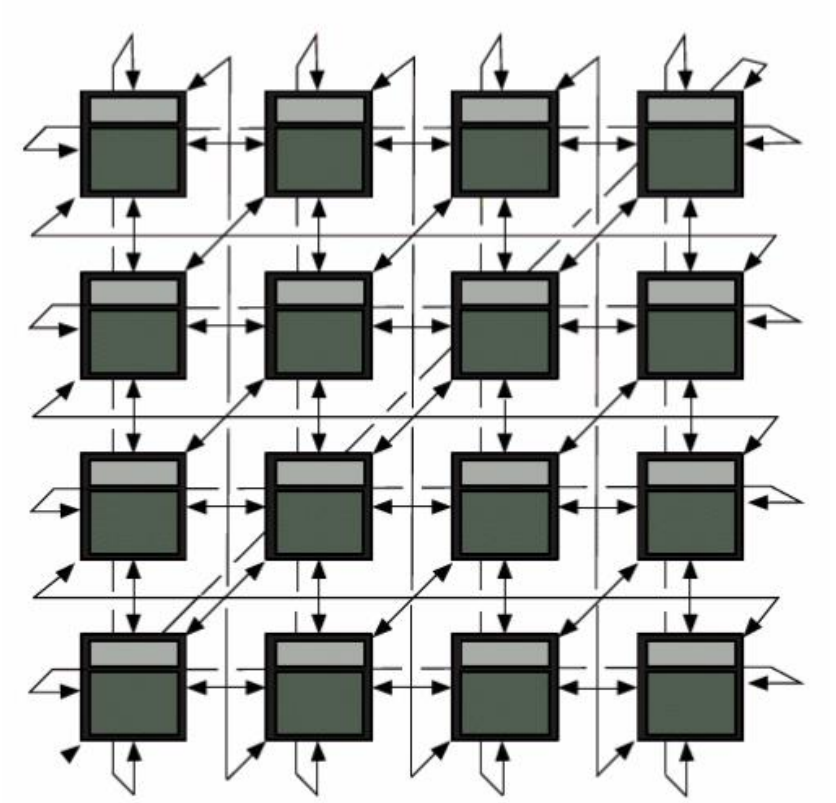
- ARM cores implement neuronal and synaptic computation in software
- This provides maximum flexibility, since neuron model and governing dynamics can be fully configured
- Each ARM core is fast enough to model 1000 neurons
- Synapses are modeled as programmable weights

SpiNNaker Network

- Each chip can communicate with six nearest neighbors
- Horizontal ring, vertical ring, and diagonal ring
- Can be viewed as two-dimensional torus network



https://apt.cs.manchester.ac.uk/projects/SpiNNaker/



Event-Based Neuromorphic Systems, edited by Shih-Chii Liu, et al., John Wiley & Sons, Incorporated, 2015. ProQuest Ebook Central, https://ebookcentral.proquest.com/lib/aalto-ebooks/detail.action?docID=1895762.

Communication

- When a neuron produces an output spike, a source routing key is given and it is delivered to communication fabric
- When the spike arrives to one of six incoming ports, the source key is matched against a 1024-entry content addressable memory (CAM)
- An entry is 32 bits, similarly to the source routing key. Each bit is either 0, 1, or X
- Source key matches the CAM if non-X values match
- When match, a 24-bit vector corresponding to matched entry is retrieved
- The bits here correspond to each on-chip ARM core (18 bits) and six output ports of the communication network (6 bits)
- The spike is propagated to all locations for which the bit is set
- If no match, the packet is forwarded to the next chip in a straight line

End of Presentation

Assignment

1. **What are some fundamental differences in the design principles of TrueNorth and SpiNNaker systems?**