The City College of New York

CSC 36000:
Modern Distributed Computing *with AI Agents*

By Saptarashmi Bandyopadhyay
Email: sbandyopadhyay@ccny.cuny.edu
Assistant Professor of Computer Science
City College of New York and Graduate Center at City University of New York

Today's Lecture

Bottlenecks in Traditional Computing

Neuromorphic Computing

How is Neuromorphic Computing Distributed?

Spiking Neural Networks

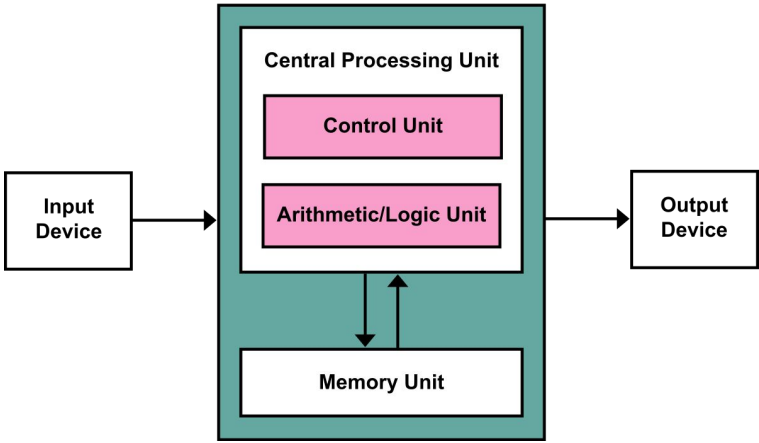
Bottlenecks in Traditional Computing

The Von Neumann Architecture (Traditional Computers)

Separation of Memory (DRAM) and Compute (Core).

Data must be shuttled back and forth constantly.

Energy Cost: Moving data consumes 100x more energy than the computation itself.



The Problem for Distributed Agents

The data bus introduces a lot of latency and energy usage, especially for AI Agents at a Distributed Scale.

This is not just inconvenient for the user, but it could be safety-critical, especially in use-cases like Autonomous Vehicles.

Furthermore, Power consumption limits scalability on Edge devices (Drones, Rovers).

A Biologically Inspired Solution

Are there any systems in real-life that have the computing power we need without the latency or energy cost?

The ***brain*** collocates memory and compute (synapses and neurons).

Operates on *~20 Watts* vs. *Megawatts* for Supercomputers.

Neuromorphic Computing

Neuromorphic Computing

Instead of a central clock (Synchronous), use Event-Driven (Asynchronous) processing.

Computation only happens when data changes (Sparsity).

This mimics how the brain processes individual events; we often don’t have a precise internal clock but we know the order of events and adapt when information changes.

Real-World Neuromorphic Processors

Chips like Intel Loihi or IBM TrueNorth contain thousands of cores.

Each core simulates hundreds of neurons.

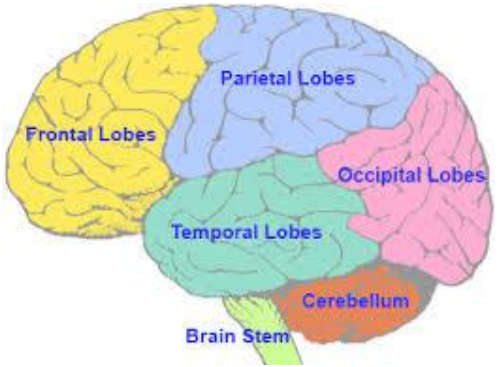
No central shared memory; memory is distributed per core.

How is Neuromorphic Computing Distributed?

A single neuromorphic chip is a distributed system of cores communicating via a packet-switched network.

Your brain is constantly communicating information between regions with vastly different purposes and capabilities!

This requires a sophisticated distributed system to share information and make decisions.

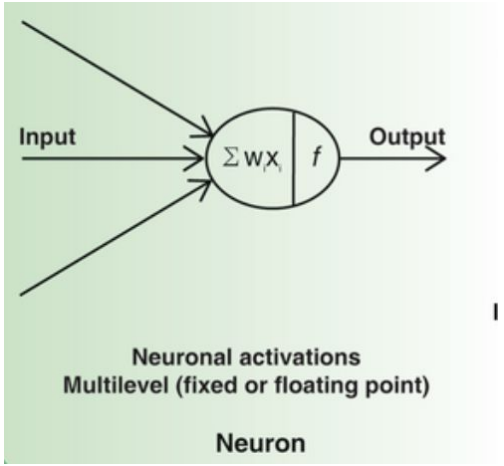


Spiking Neural Networks

Artificial Neural Networks (Traditional Deep Learning)

Continuous values (ReLU, Sigmoid).

Synchronous layers (Layer 1 -> Layer 2 -> Layer 3).

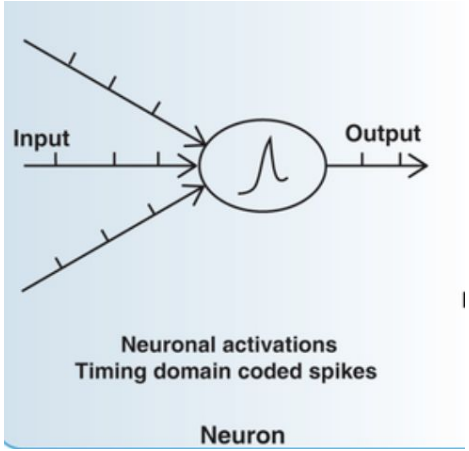


Spiking Neural Networks (Neuromorphic)

Binary "Spikes" (0 or 1) over time.

Information is encoded in the timing and frequency of spikes.

Asynchronous: Neurons fire whenever they reach a threshold, independent of a global clock.



Questions?