On the Link Between Energy and Information for the Design of Neuromorphic Systems

Narayan Srinivasa

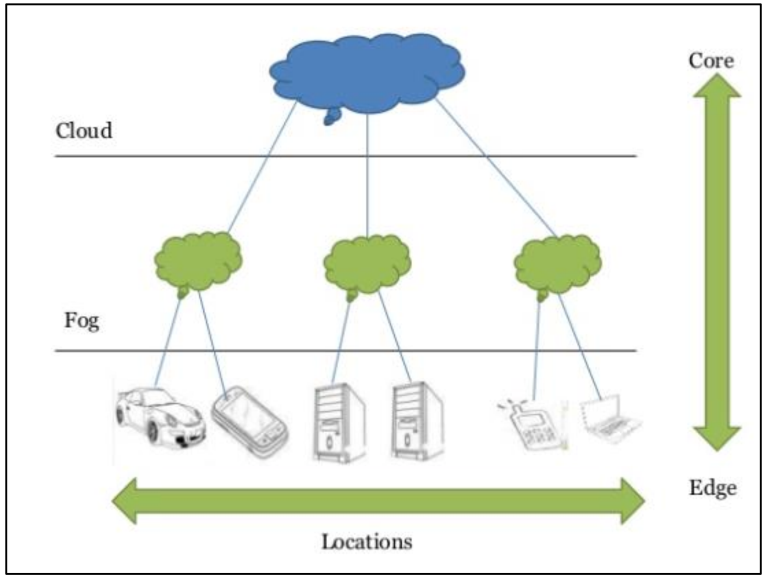
Eta Compute

Simon's Institute "Computational Theories of the Brain"
UC Berkeley, April 18, 2018

Eta Compute
DISRUPTION AT THE CORE

Edge Devices and their relationship with the Cloud

- Edge Device directly interact and communicate with other machines, objects, environment, and infrastructure (**50B** by 2020)
- Edge devices are involved in the fastest sense-infer-act loop
- Processing directly in the fastest loop gives the best agility
- Cloud is more powerful but slow



Eta Compute
DISRUPTION AT THE CORE

Eta Compute is focused on addressing issues at the edge using neuromorphic computing

Each self-driving car generates 4TB/hr

Data Volume

Your Samsung TV is eavesdropping on your private conversations

Privacy

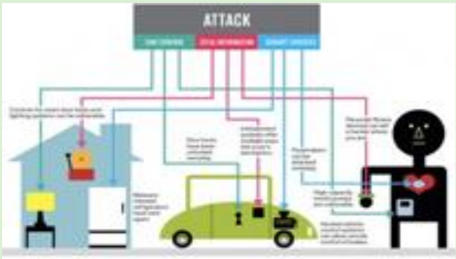
Are Google Home and Amazon Echo listening more than you realize?

Identity

Who needs Santa? 6-year-old orders dollhouse and cookies from Amazon's Alexa

Pollution

Energy



Yes, terrorists could have hacked Dick Cheney's heart

Security

Finns chilling as DDoS knocks out building control system

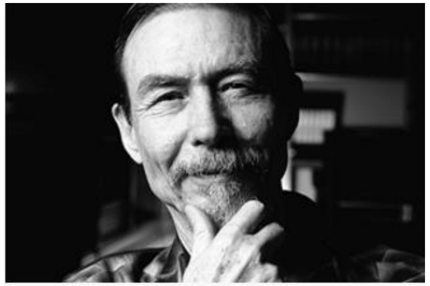
Virtual reality has a motion sickness problem

Horrifying video shows a drone dispensing candy crash into crowd below

Latency

Neuromorphic Computing

Original term "neuromorphic electronics" to describe electronic *analog* circuits that *mimic* neurobiological circuits and architectures in the nervous system



Carver Mead
1985

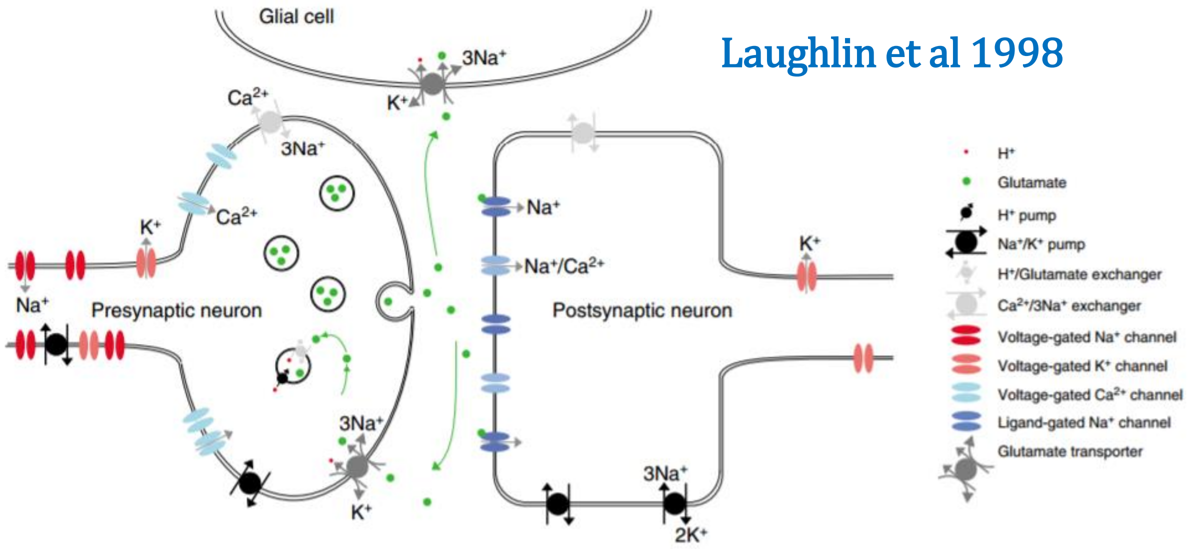
"Neuromorphic engineering/computing" was introduced to expand the scope to include analog, digital, mixed-mode analog/digital VLSI and software systems

Eta Compute
DISRUPTION AT THE CORE

Brain evolved to balance energy consumption with information processing

- Brain is unique – serves as an interface between morphology, physiology and behavior
- Brain evolution is shaped by two key selective pressures
- Selective pressure #1: to generate adaptive behavior via information processing under changing conditions (Benefit)
- Selective pressure #2: To minimize the energy consumed during this process (Cost)

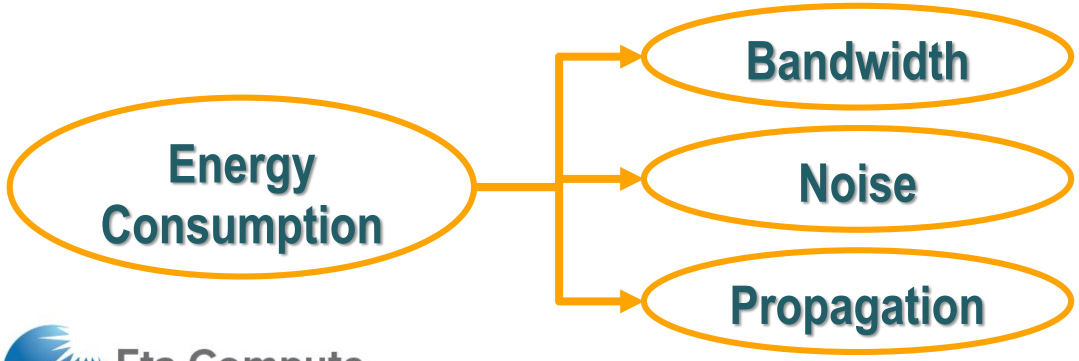
Bulk of energy consumed for spike generation & for its speed, quality and propagation needs



Laughlin et al 1998

Nervous system consumes 20% of energy for just 2% of body mass

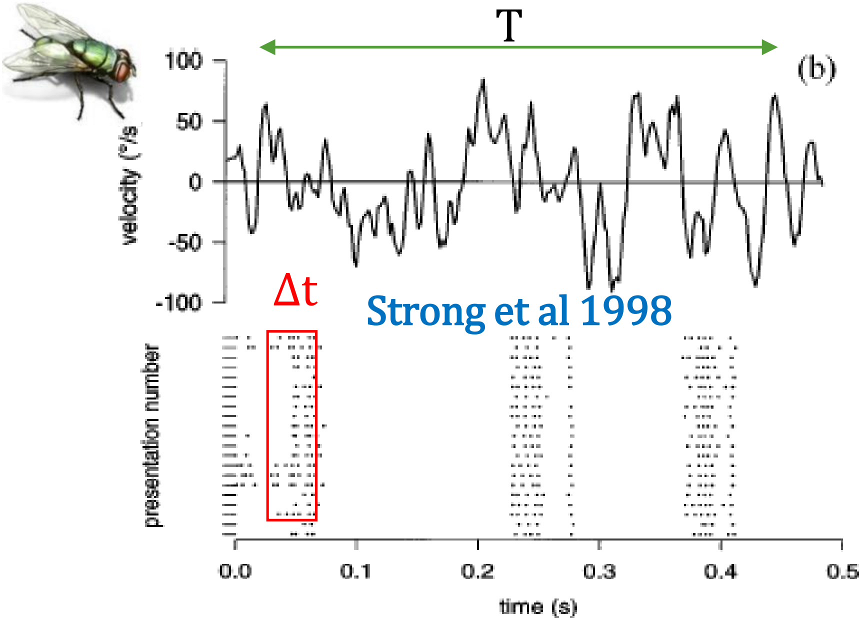
- Action potential generation
- Action potential maintenance
- Synaptic transmission

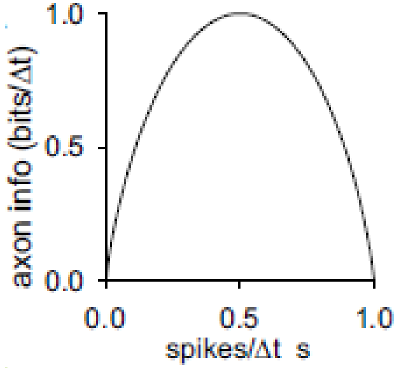


Feature	Purpose
Bandwidth	Signaling speed
Noise	Signal quality
Propagation	Signal transmission

Eta Compute
DISRUPTION AT THE CORE

Asynchronous Spike Timing Maximizes Information Rates



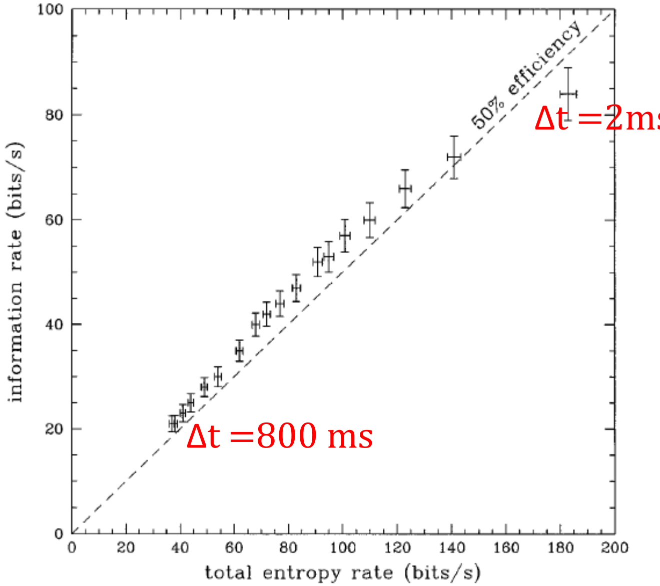


Useful Information:

$$I = H_s - H_N \ bits$$

Efficiency

$$\sigma = I/H_s$$

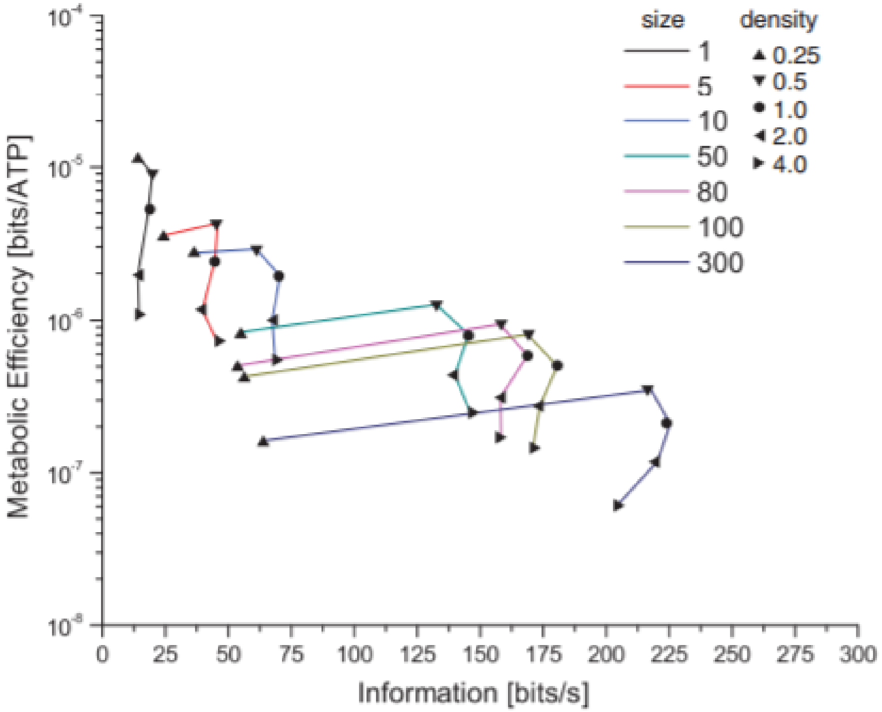


- In response to long stimuli – variability in spike train or its spike entropy H_s was calculated
- For repeated identical stimuli – variability in spike train or its noise entropy H_N was calculated

Eta Compute
DISRUPTION AT THE CORE

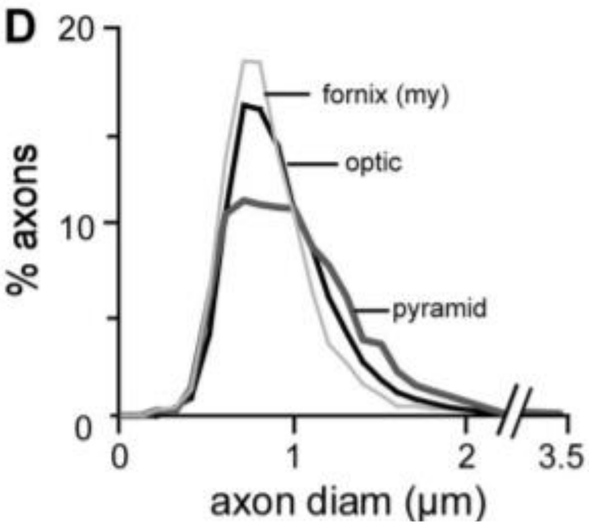
Neuronal morphology biased to deliver information at the lowest allowed energy

Smaller neurons ⇒ lower Information rates
⇒ higher energy efficiency

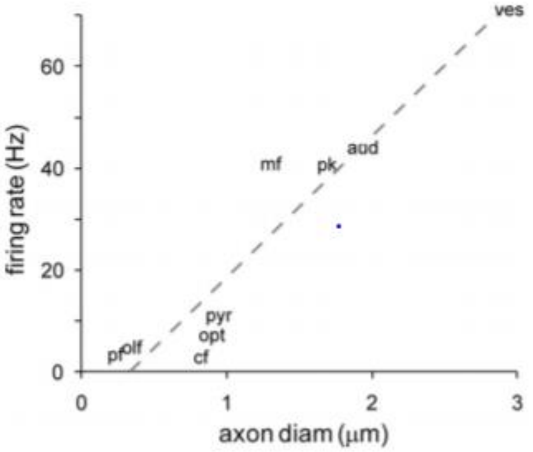


Distribution of axonal tracts skewed towards thin axons

Perge et al, 2012

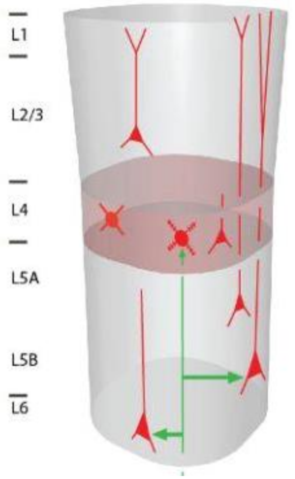


Thin axons have lower firing rates than thicker ones



Eta Compute
DISRUPTION AT THE CORE

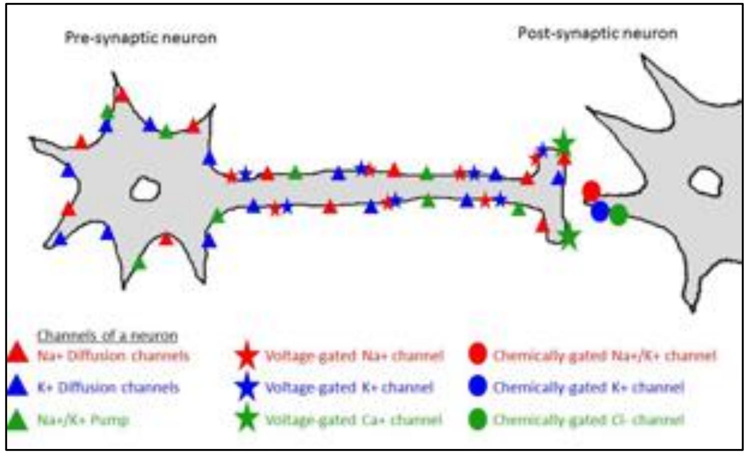
Balance between dense short & sparse long connectivity enables energy efficient behavior



- Distributions of dense short connections are “tuned” to extract information from the environment
- But adaptive behavior requires many of these local computations to be integrated rapidly
- Thicker and longer axons encode information spread over several low information rate thin axon tracts
- These sparse long range connections enables constant synaptic path length ⇒ rapid information exchange

Eta Compute
DISRUPTION AT THE CORE

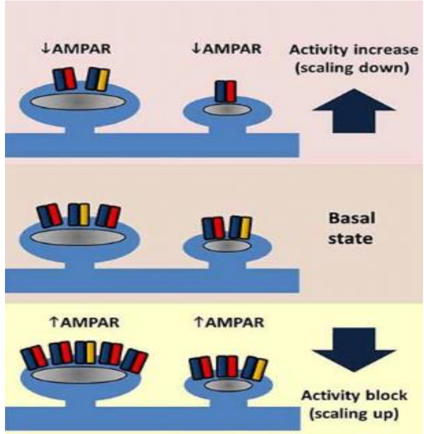
Neurons evolved to balance information rates with energy consumed



Improve channel properties (kinetics & sensitivity) ⇒ high IR but high energy

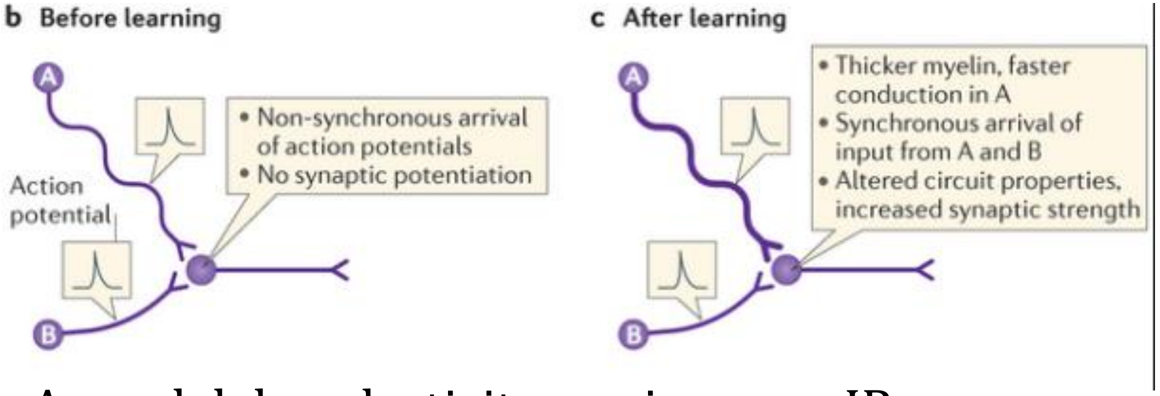
Voltage gated channels control threshold ⇒ control IR and energy consumed

Turrigiano & Nelson, 2004; Modjeski et al, 2016



Homeostatic plasticity regulates IR and hence energy consumed

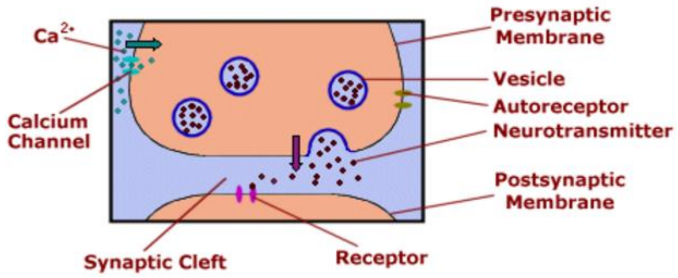
Fields, 2015



Axonal delay plasticity can improve IR per joule by controlling myelination

Eta Compute
DISRUPTION AT THE CORE

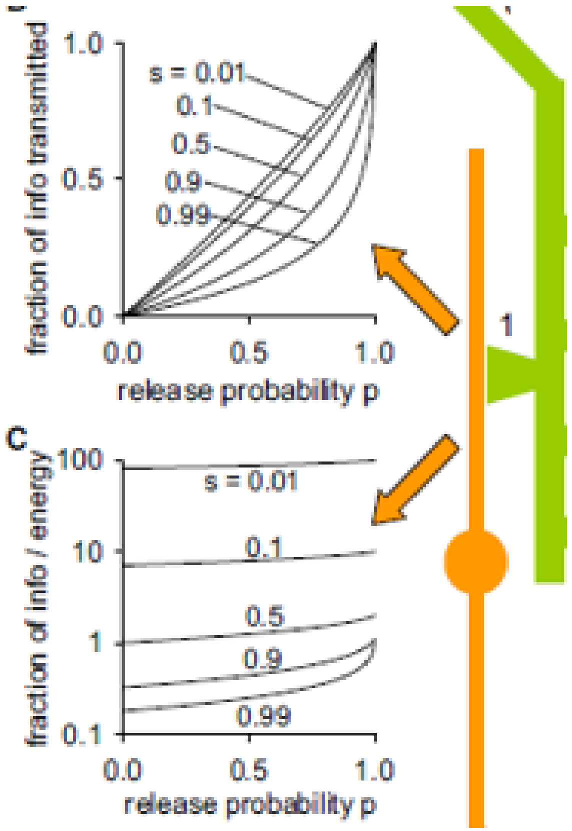
Biophysics at synapse optimized for energy efficient information processing



p – probability of vesicle released onto postsynaptic cell
s – probability of presynaptic spike occurrence

- Information transmitted at a synapse at increasing *p* is linear for lower *s* but less-than-linear for higher *s*

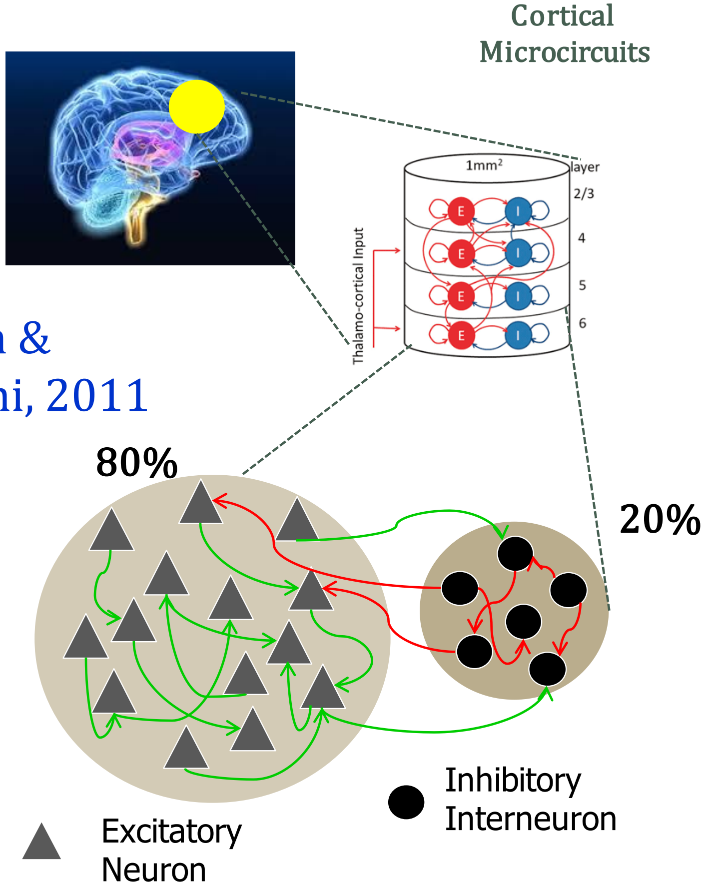
- Since lower *s* implies lower energy consumed, the transmitted information per joule is higher for lower *s* irrespective of *p*



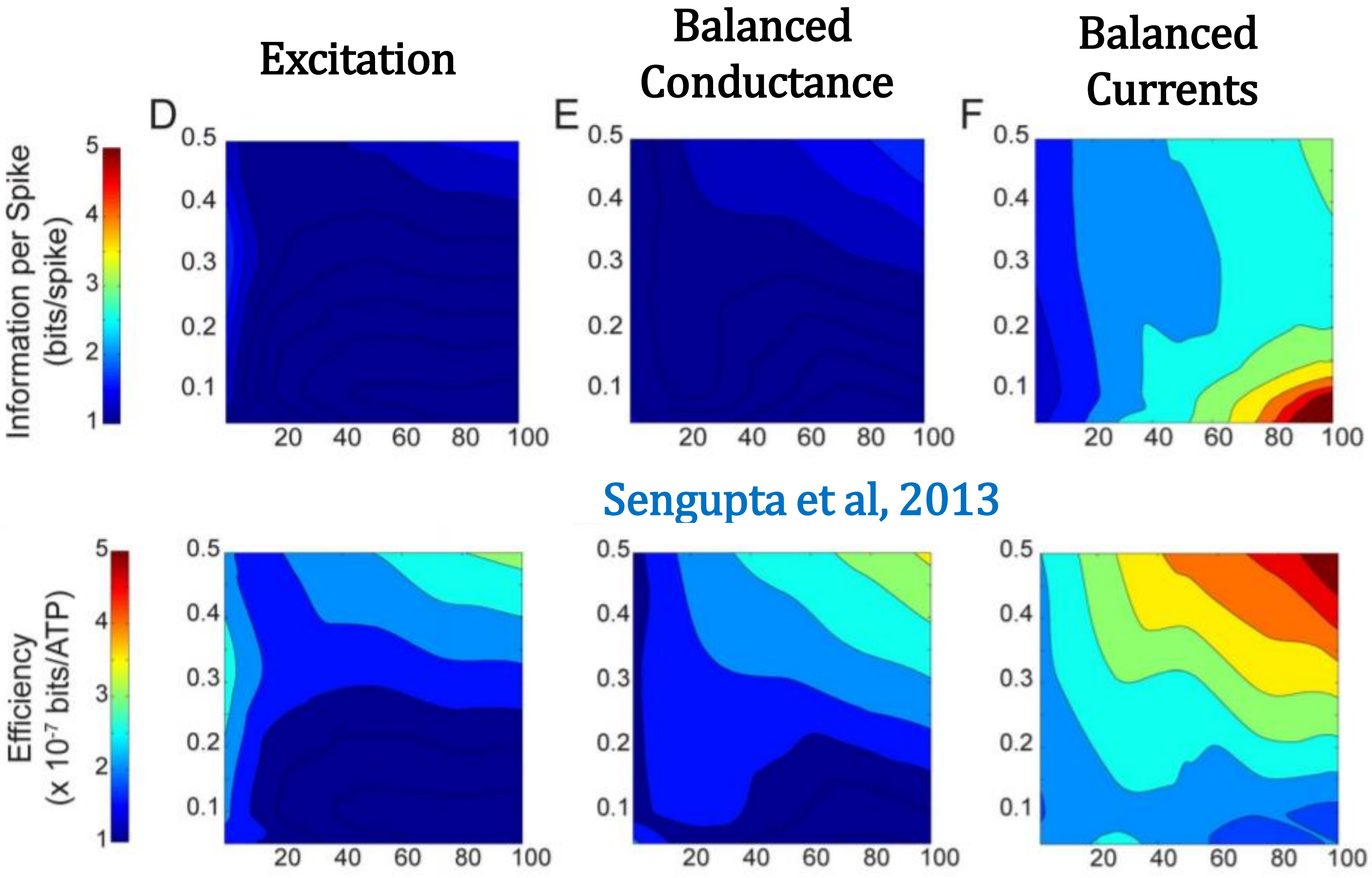
Harris et al, 2012

Eta Compute
DISRUPTION AT THE CORE

Balanced synaptic currents promote efficiency in both information coding & energy



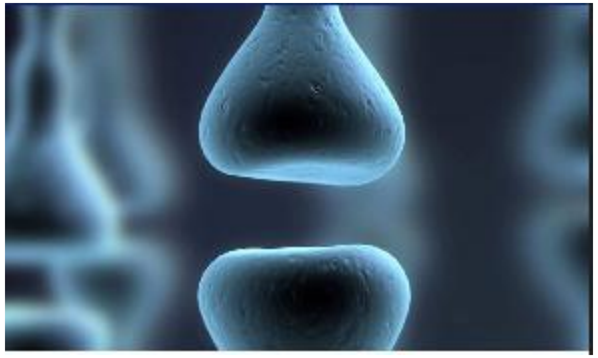
Isaacson & Scanziani, 2011

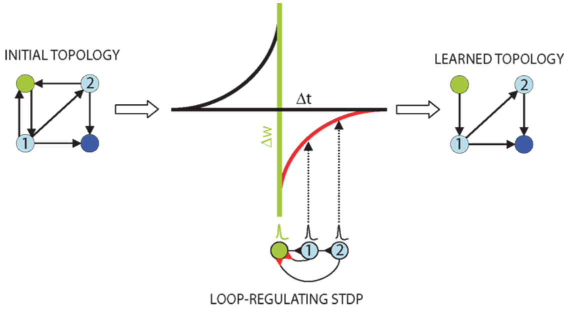


Sengupta et al, 2013

Eta Compute
DISRUPTION AT THE CORE

Synaptic plasticity enables optimal filtering of information for efficient information transfer

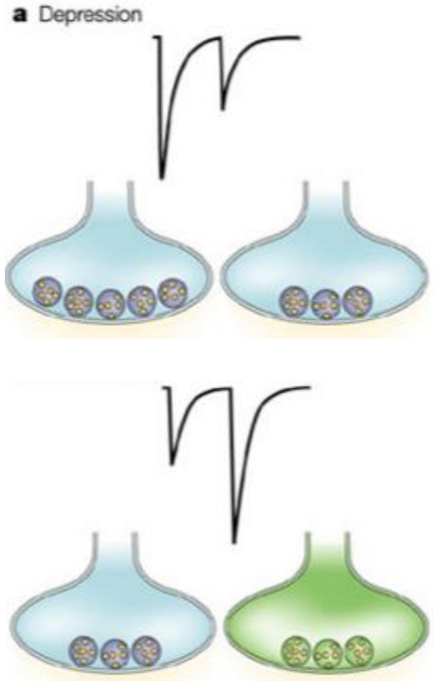




Markram et al 1997, Bi & Poo, 1998

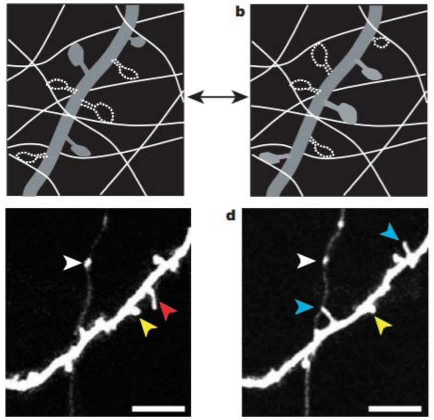
Spike timing dependent plasticity can optimize unsupervised filtering of information

Blitz et al, 2004



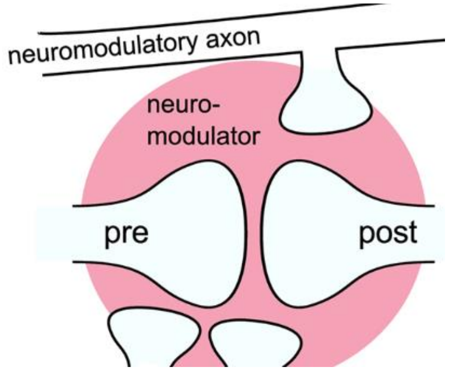
Short-term plasticity can regulate information rates while lowering energy

Chklovksii et al, 2004



Structural changes to optimize information transfer rates via spine growth & pruning

Pawlak et al, 2010

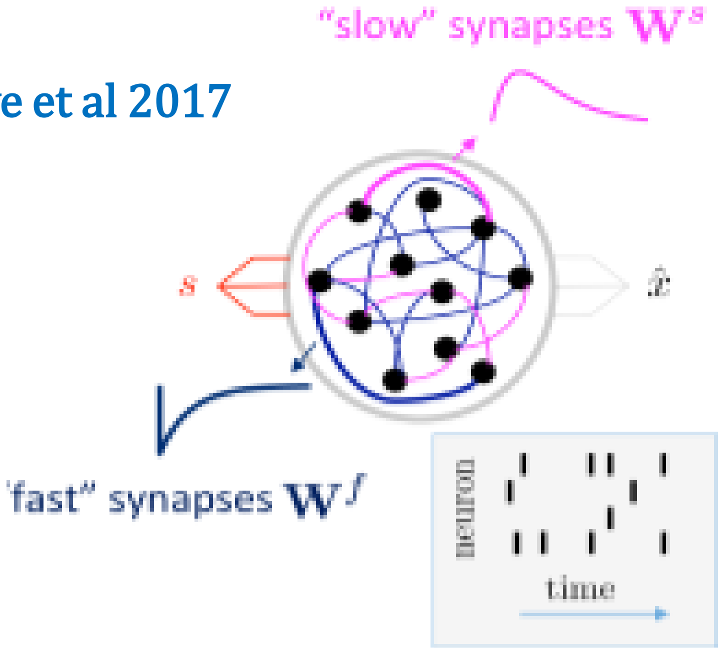


Neuromodulation of synaptic plasticity can also optimize filtering of information at different time scales

Eta Compute
DISRUPTION AT THE CORE

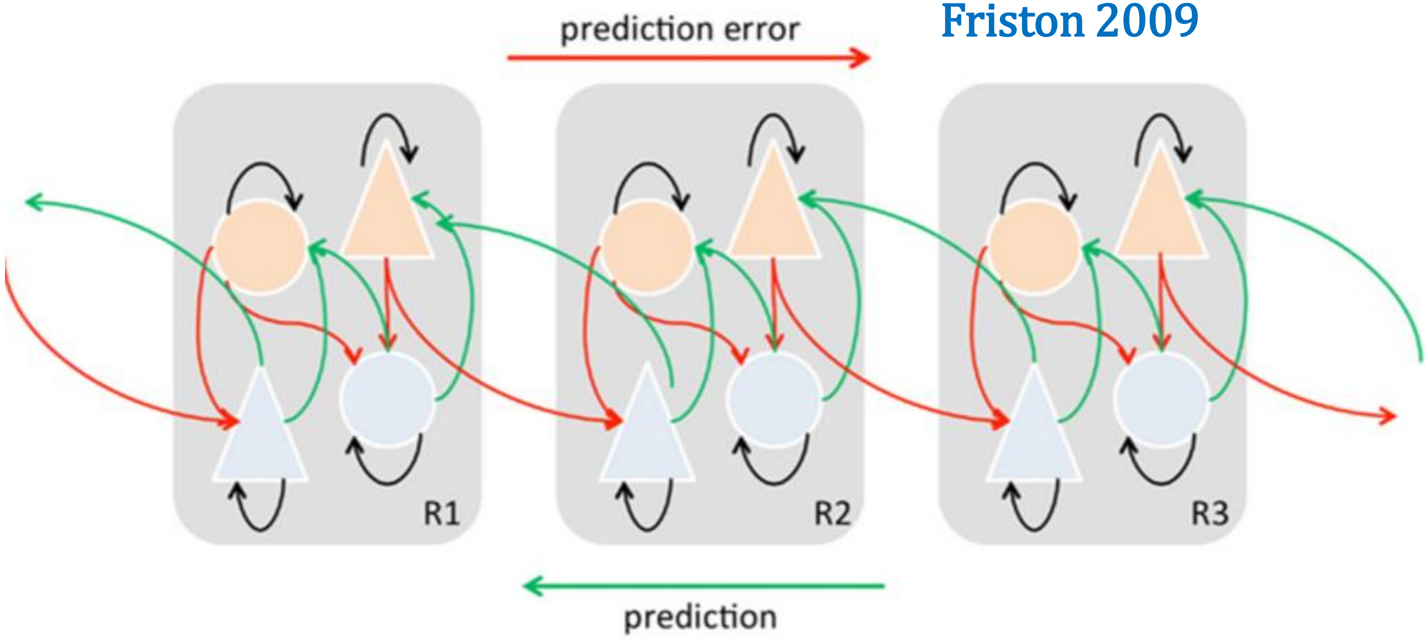
Population Dynamics Biased to Optimize energy while enhancing information processing

Deneve et al 2017



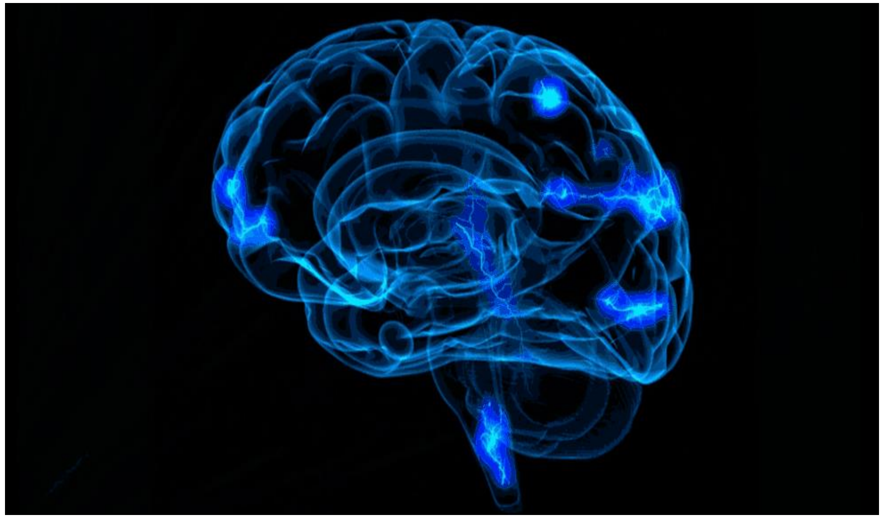
Enforcing a balance between excitation and inhibition in populations of spiking neurons enable efficient information coding

Friston 2009



Reduced encoding of redundant information at lower layers (R1) for energy savings at higher layers (R3)

Massively Parallel & Asynchronous Brain Enables System Level Efficiency during information processing

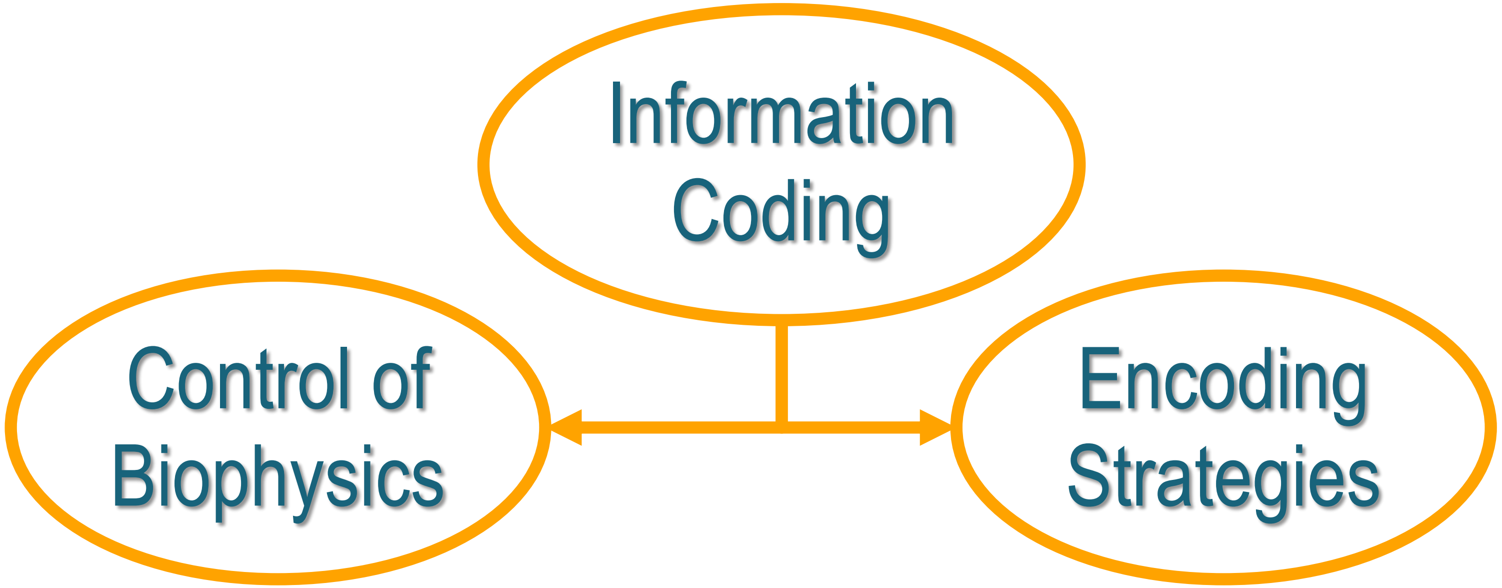


Buzsaki, 2006, Zeki, 2015

- Strong evidence brain is asynchronous – many clocks or rhythms
- This “just in time” mode of operation is one reason for brain efficiency during behaviors (system level)
- Complex operations can take more time than average and simple ones can take less
- Actions can start as soon as prerequisite actions are done

Strong link between energy & information offers a blueprint for the design of neuromorphic systems

Neural/synaptic circuit design (event driven/ asynchronous), mix of analog low voltage circuits with digital circuits, programmable plasticity



Make the signals that carry information energetically cheap

Circuits that require fewer signals to carry information

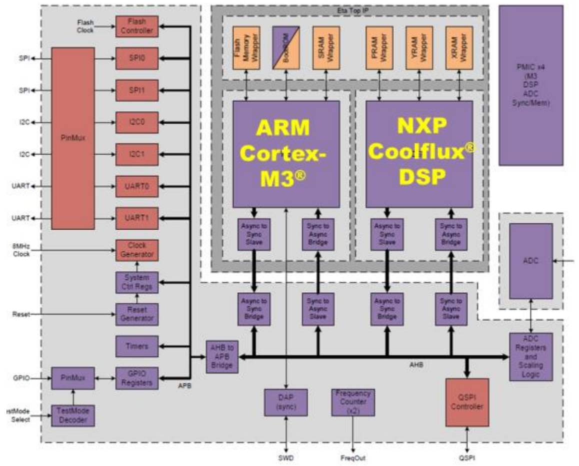
Asynchronous multicore, programmable connectivity, 3D memory structures for density, metal layer specific routing, Memory & compute tightly integrated

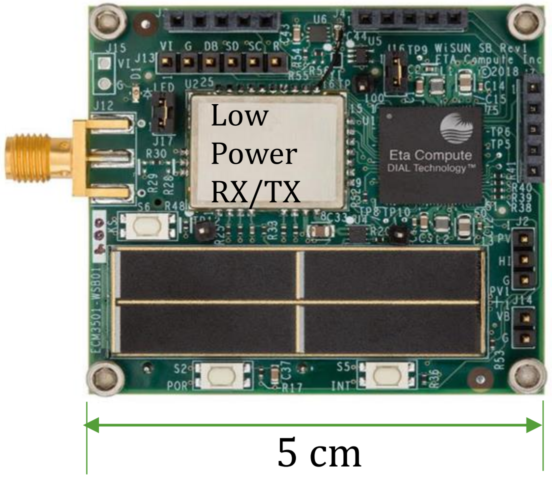
Eta Compute
DISRUPTION AT THE CORE

Eta Compute Asynchronous Single Core

Delay Insensitive Asynchronous logic (DIAL™) core – timing precision without fast clocks

Advantage	Why?
Low power consumption	Due to fine grain clock gating and zero standby power
High operating speed	Operating speed determined by local latencies not global worst case latency
Robustness to variations in supply voltage, temperature and fabrication process parameters	Our innovation where timing is based only on matched delays (and is insensitive to circuit and wire delays)
Formally Verifiable	Our innovation to ensure synchronous and asynchronous operations are equivalent
No new tools are needed	Our methodology is fully implanted using conventional tools





Single Core:
128 KB RAM
512 KB Flash

2 K neurons,
128 K synapses

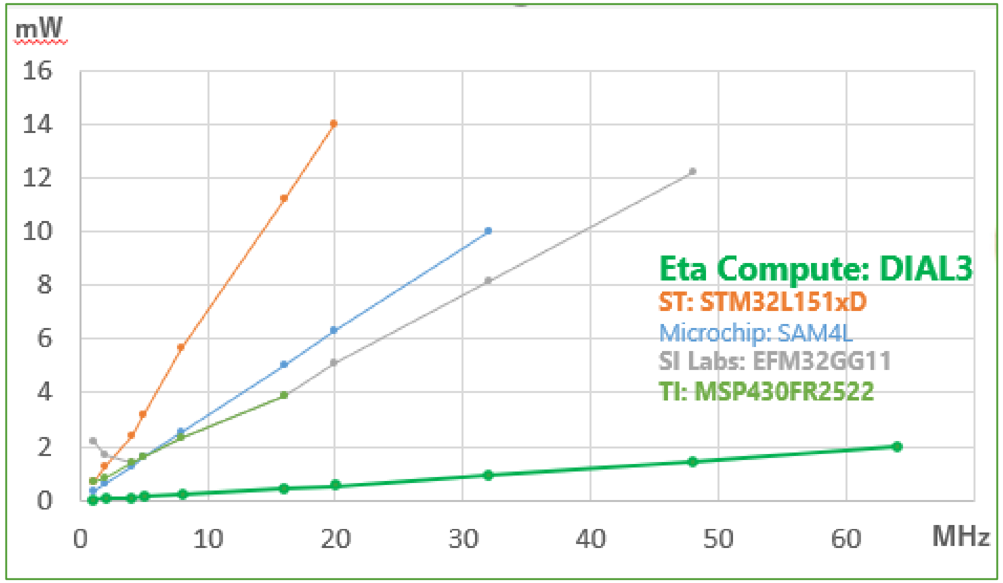
Max speed:
60 MHz, 2 mW

Continuously variable voltage down to 0.2V for low power 2 µW at 100 KHz !

Low cost – 55 nm process

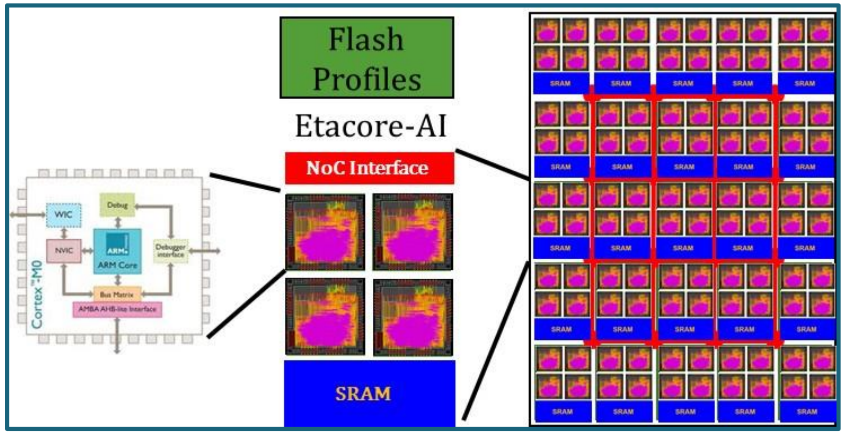
Eta Compute
DISRUPTION AT THE CORE

Eta Compute Asynchronous Multicore Chip



Multicore – tapeout in fall 2018

100K neurons, 10M synapses

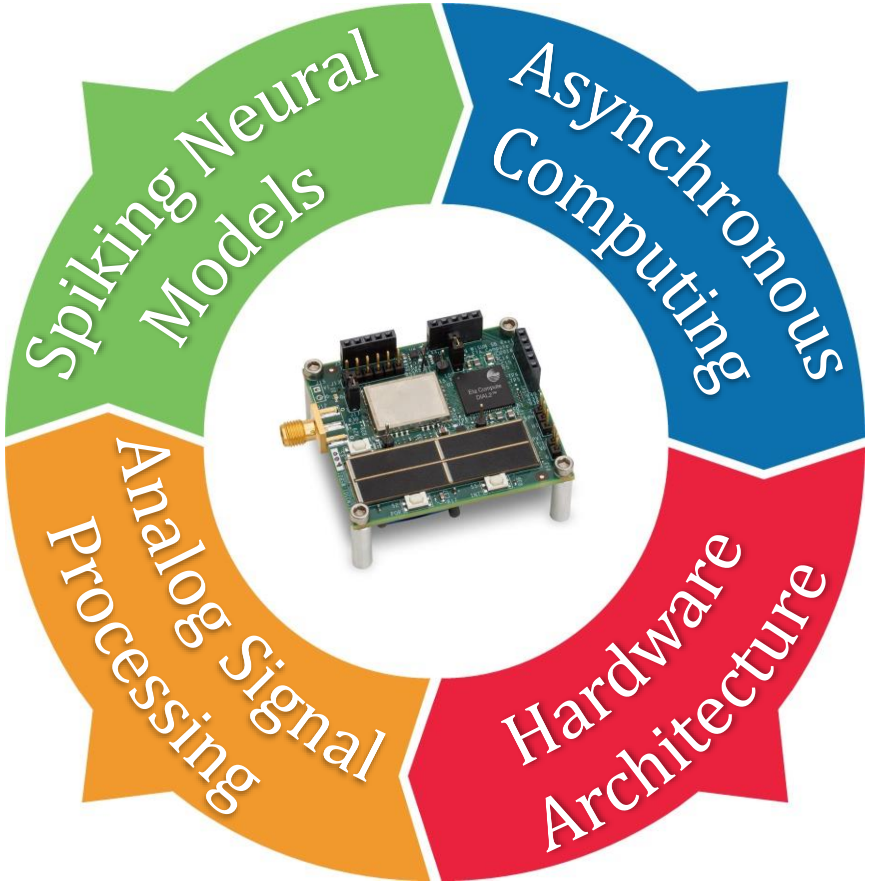


100 cores, < 100 mW, < 100mm^2

- Sparse spiking dynamics that runs at low frequencies ⇒ exploit low voltage operation
- Software programmable in C++ for flexibility in exploring model parameters and dynamics
- Demonstrated voice applications for keyword spotting, continuous speech recognition, etc
- Beginning to look into image processing and health monitoring applications

Eta Compute
DISRUPTION AT THE CORE

Summary



- Brain evolved to strike a balance between information processing and energy consumption
- Asynchrony in computing and information processing is one of the key principles in enabling this balance in neuromorphic systems
- Spiking neural models with synaptic plasticity combined with analog signal processing running on DIAL enables efficient information processing
- Several applications currently being tested seems to hold promise for interesting applications at the edge

Eta Compute
DISRUPTION AT THE CORE