Lecture: Spiking Networks Architecture

- Topics: TrueNorth design, projects discussion

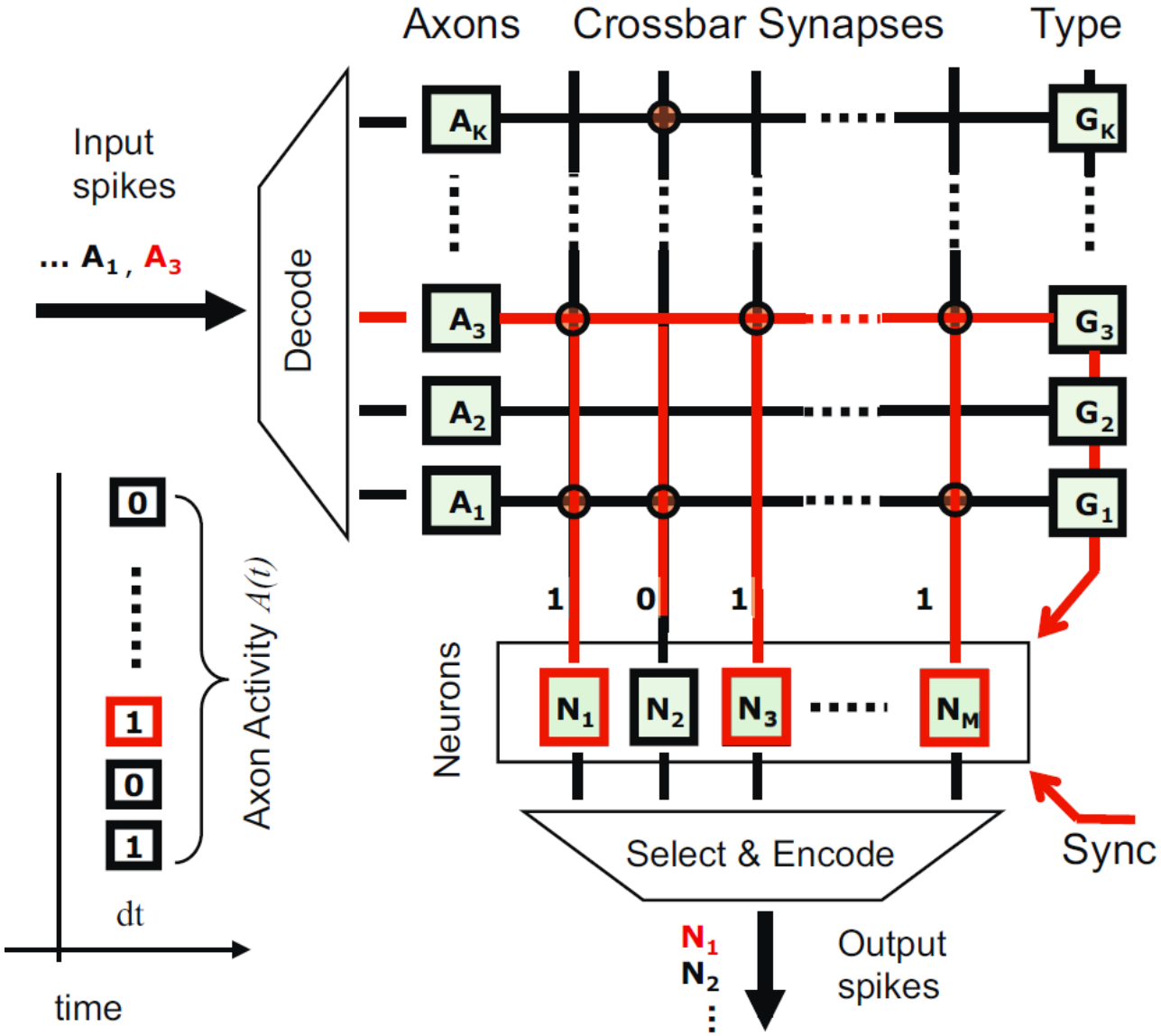
The Spiking Approach

- Low energy for computation: only adds, no multiplies
- Low energy for communication: depends on spikes per signal
- Neurons have state, inputs arrive asynchronously, info in relative timing of spikes, other biological phenomena, …

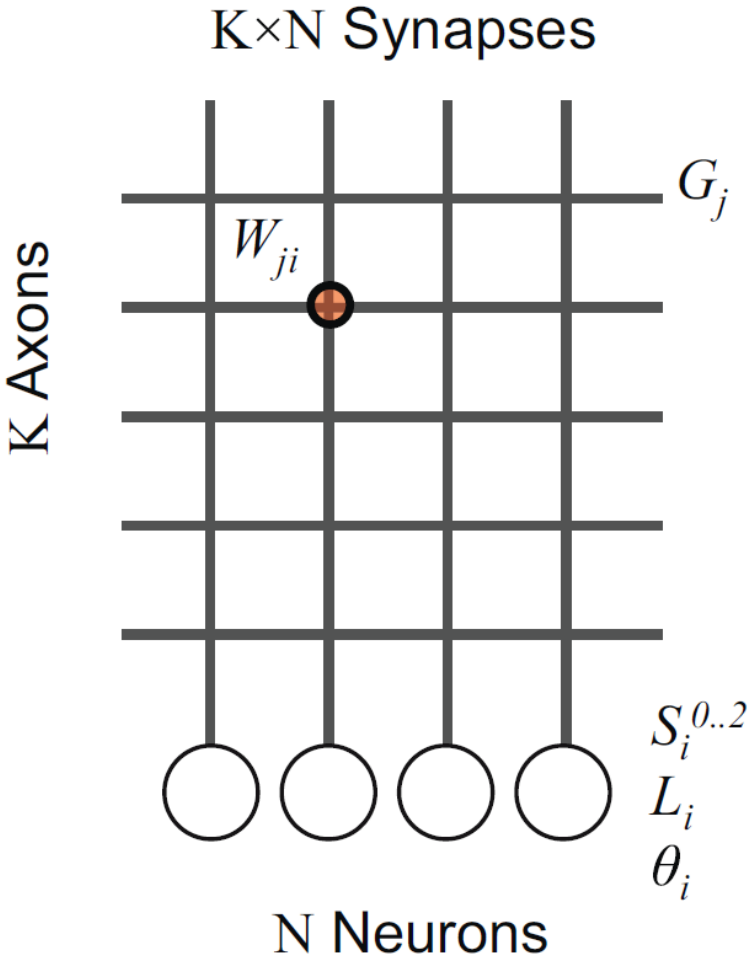
IBM TrueNorth

- Product of DARPA's SyNAPSE project
- Largest chip made by IBM (5.4 billion transistors)
- Based on LLIF neuron model
- Lots of on-going projects that use TrueNorth to execute new apps
- Lots of limitations as well – all done purposely to reduce area/power/complexity

TrueNorth Core

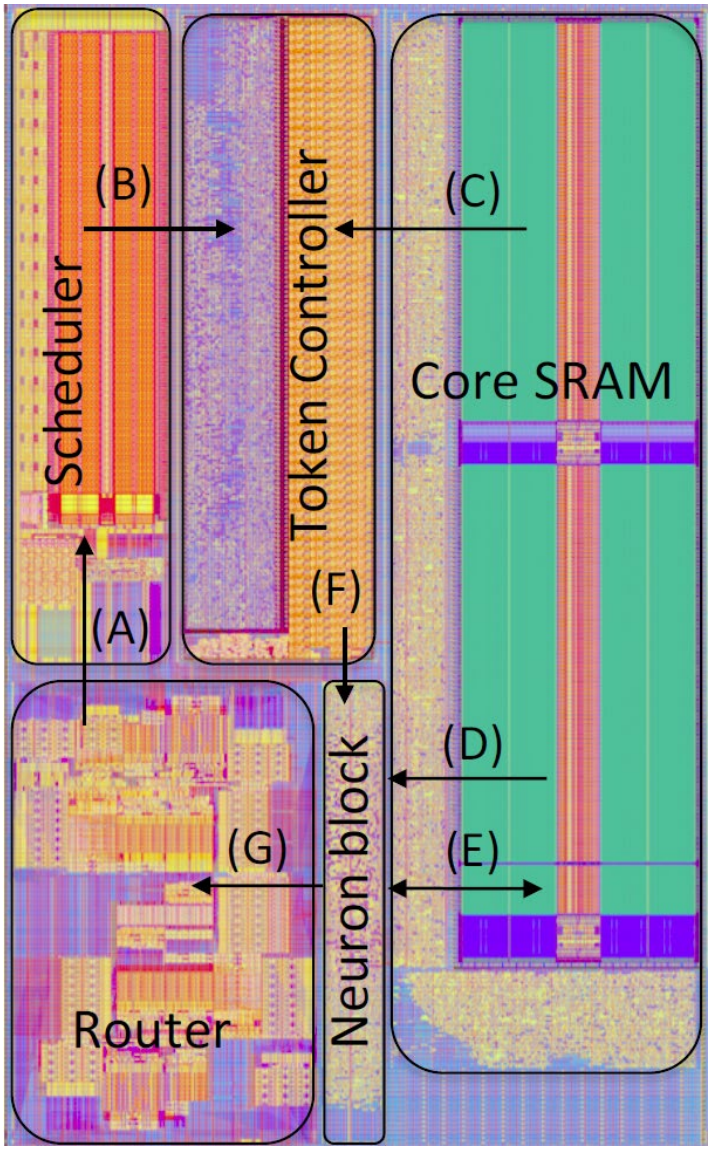


TrueNorth Core (Axonal Approach)



Name	Description	Range
W_{ji}	Connection between axon j and neuron i	0,1
G_j	Axon type	0,1, 2
$S_i^{0..2}$	Synapse values	-256 to 255
L_i	Leak	-256 to 255
θ_i	Threshold	1 to 256

TrueNorth Core (Dendritic Approach)



Principles

- Focus on low power (speed is secondary)
- Low power with asynchronous circuits
- Low power with high-Vt circuits and SOI process
- A dendritic approach leads to fewer SRAM reads/updates and is more deterministic
- Also uses an under-provisioned network (giving rise to the concept of “future” ticks)

Receiving Spikes

- Spikes arrive at the *Scheduler*; stored in a 256x16 SRAM grid to indicate axon and time of the spike
- The *Token Controller* receives spikes in a "*tick*" (and their types – exc/inh); it then sequentially walks through 256 neurons – this dendritic approach makes the latency and SRAM accesses more deterministic
- It reads a 410-bit word from SRAM for that neuron (this word has connectivity info used by Token Controller and neuron parameters used by *Neuron block*
- Based on connectivity and input spikes, instructions are sent to Neuron block

Neuron Block

- Neuron computations are performed here (these are non-trivial because of stochastic behaviors – more later)
- A leak is introduced every tick for every neuron
- After thresholding, a spike may be triggered
- The neuron block uses synchronous circuits, but it is active only when it receives instructions from the Token Controller; the Token Controller, Scheduler, and network Router are all asynchronous to exploit the low power inherent in spike sparsity

More Details

- The 410 bit word: 256 synaptic connections, 4 9-bit weights, 9-bit leak, 8-bit threshold, 26-bit destination axon, 4-bit delivery time, 9-bit potential, … (62 more bits)

- The network may occasionally not handle a burst of spikes; if this is likely, then spikes must have a future delivery time; allows a spike to take multiple ticks to reach its destination

- 256x256 core; 4K cores; million neurons; 430 mm2; 65 mW on average; 28nm Samsung process; 1ms tick

Limitations

- A neuron can only have 256 inputs (no provision for partial sums)
- A neuron can only have one destination axon output of a specific type (seen by 256 neurons)
- Weight quantization (4 weights per neuron)

Measurement Data

- Because of the synch/asynch approach, the chip works correctly even at 0.7V supply voltage (low power!); 0.7-1.05V
- 42 mW at 0.7V, 0 Hz firing rate, 0 synapses/neuron
  323 mW at 1.05V, 200 Hz firing rate, 256 synapses/neuron
- At 0.75 V, the max throughput (giga synaptic ops per second) is 58 GSOP/s, the energy efficiency is 400 GSOP/sW
- At 0.75 V, complex recurrent network, 20 Hz avg firing rate, 128 syn/neuron, 65 mW and 46 GSOP/sW
- At low activity levels, the ticks can be sped up to 21 KHz

More Results

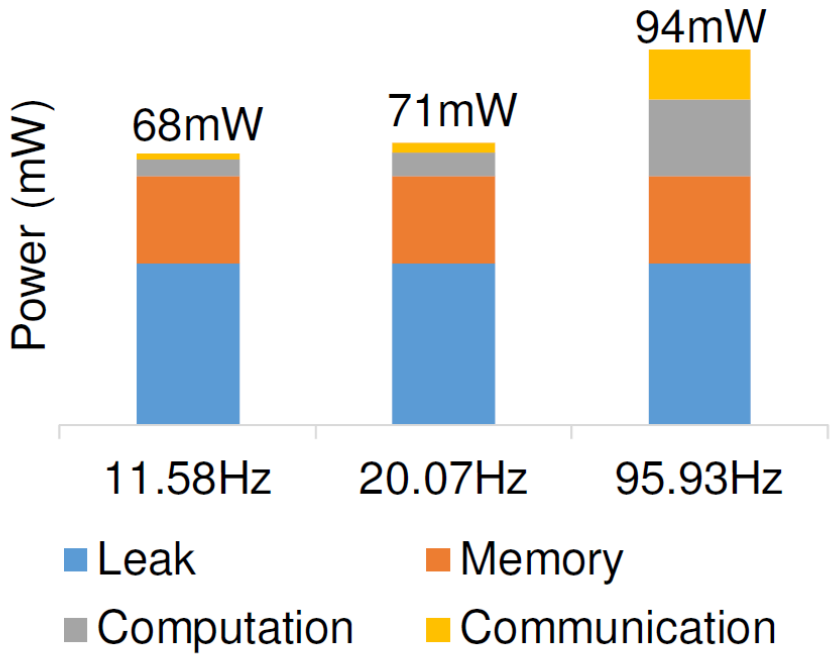


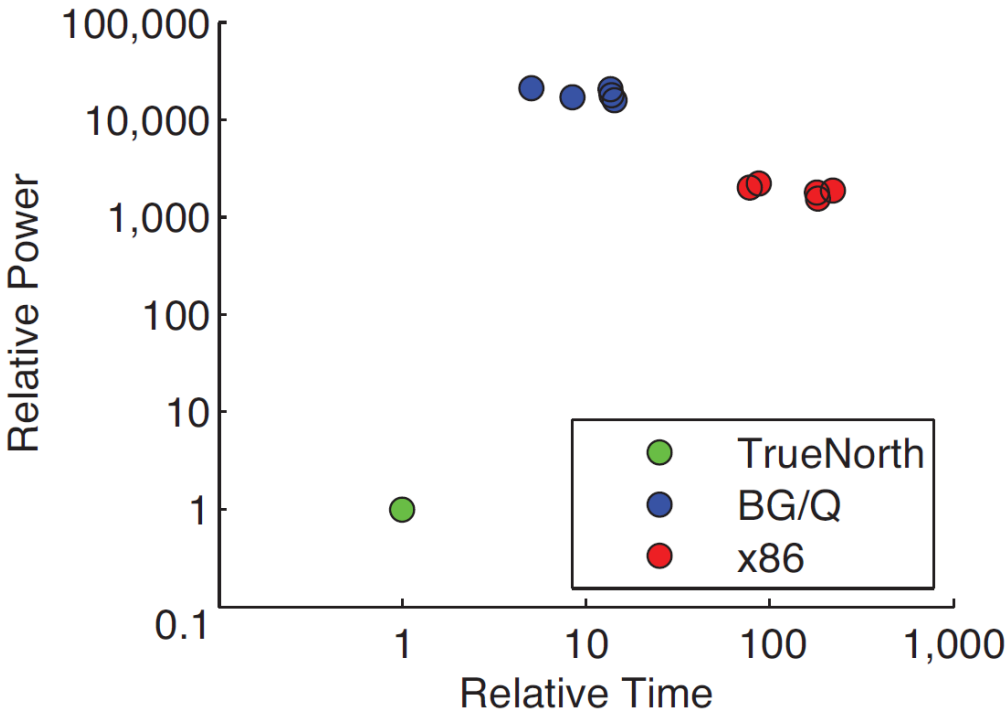
Fig. 17: Total TrueNorth chip power breakdown (@ 0.8V) for three complex recurrent networks with 128 synapses per neuron average, and three different average firing rates.

Improvements

- 4K chips on a rack
  ➔ 4 B neurons, 300 W for TN chips, plus more for communication/FPGAs
  -- 4 KWs (SC'14)

- 96 racks ➔ 412 B neurons, 10^{14} synapses (human brain scale), 29 kW for TN chips, few hundred kW overall

- Power and time, relative to the Compass simulator, running on the BG/Q supercomputer and a dual-socket Intel system



Stochastic Neurons

- The TrueNorth neurons have many stochastic parameters
- Helps emulate many biological and non-linear behaviors
- Continues to be low cost; applies the RISC argument
- Stochasticity in input, state, and output
- Different modes for leak, threshold, reset
- Helps create a library of 50+ neuron models, including the 20 defined by Izhikevich (using 1-3 neurons)

Neuron Parameters

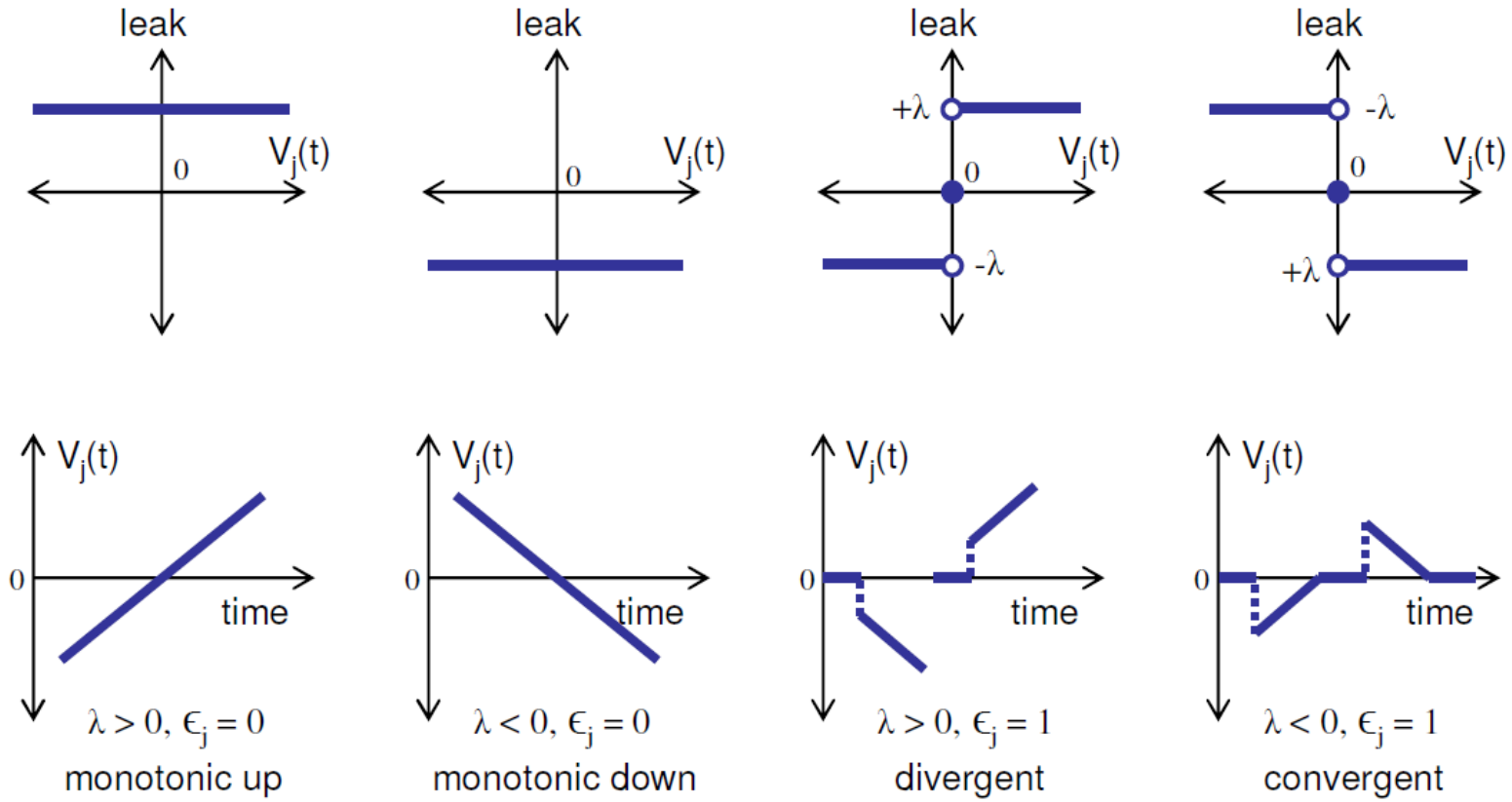
Variables and Parameters	Symbol	Format	
membrane potential	$V_j(t)$	signed int	9 bits
local timestep	t	unsigned int	
input spikes on i^{th} axon	$A_i(t)$	$\{0,1\}$	
synaptic PRN	$\rho_{i,j}$	unsigned int	
leak PRN	ρ_j^{λ}	unsigned int	
threshold PRN (drawn)	ρ_j^{T}	unsigned int	
threshold PRN (masked)	η_j	unsigned int	
leak direction variable	Ω	$\{-1,0,+1\}$	
synapse (i^{th} axon, j^{th} neuron)	$w_{i,j}$	$\{0,1\}$	256 bits
type of i^{th} axon	G_i	$\{0,1,2,3\}$	
synaptic weight/probability	$s_j^{G_i}$	signed int	36 bits
synaptic weight/probability select	$b_j^{G_i}$	$\{0,1\}$	4 bits
leak-reversal flag	ϵ_j	$\{0,1\}$	1 bit
leak weight/probability	λ_j	signed int	9 bits
leak weight/probability select	c_j^{λ}	$\{0,1\}$	1 bit
positive $V_j(t)$ threshold	α_j	unsigned int	8 bits
negative $V_j(t)$ threshold/floor	β_j	unsigned int	8 bits
threshold PRN mask	M_j	unsigned int	8 bits
reset voltage	R_j	signed int	9 bits
negative thresh: reset or saturate	κ_j	$\{0,1\}$	1 bit
$V_j(t)$ reset mode	γ_j	$\{0,1,2\}$	2 bits
PRNG initial seed value	ρ_j^{seed}	unsigned int	8 bits

Synaptic and Leak Probabilities

- If the input is stochastic, the potential change is either -1, 0, or +1. The probability of the incr/decr is determined by the synaptic weight (compare the uniformly distributed pseudorandom number PRN to the |weight|). The incr/decr is determined by the sign of the weight.
- Same for leak
- Have to draw many random numbers for each neuron: threshold, leak, each connected input axon

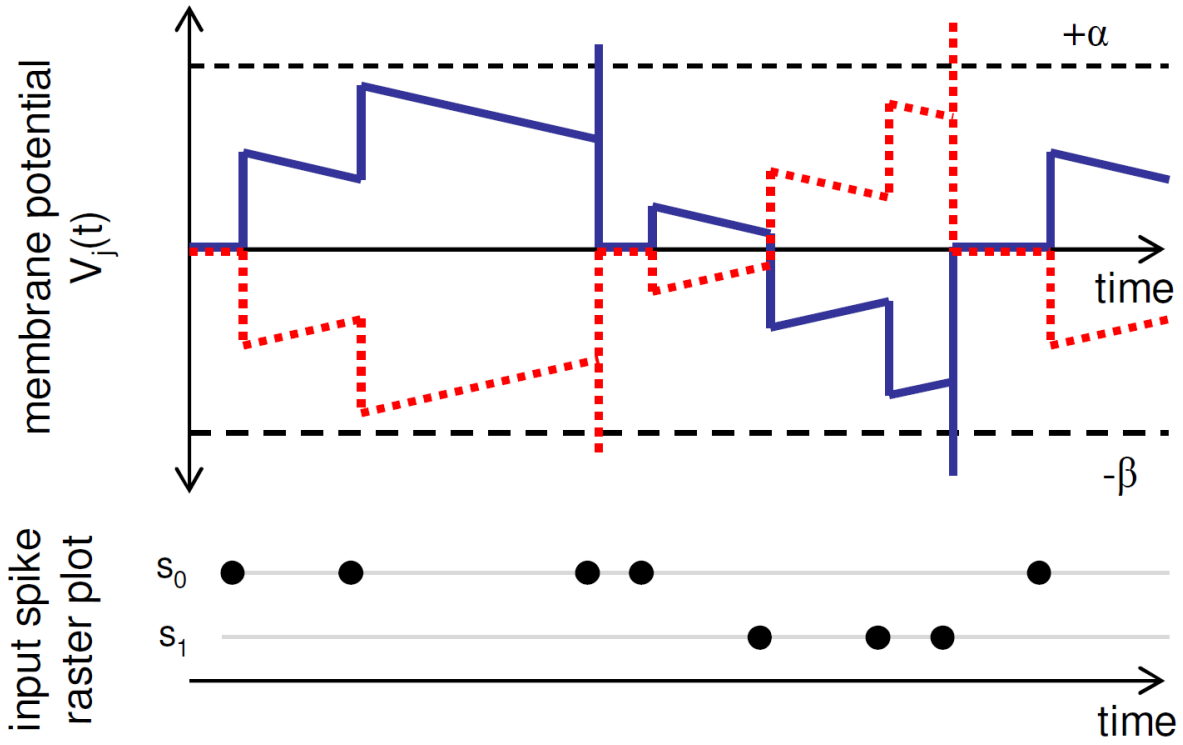
Leak Modes

- Standard leak operation: constant incr/decr of leak
- Leak-Reversal mode: the leak sign flips when the neuron potential flips
- No leak at zero potential
- Creates a convergent leak and a divergent leak



Negative Threshold

- Crossing the positive threshold always produces a spike
- There's also a negative threshold that does not produce a spike
- The negative threshold may either act as a floor, or as a "bounce" to the negative reset voltage
- The bounce allows us to create a neuron ON-OFF pair; they mirror each other



Stochastic Threshold and Reset

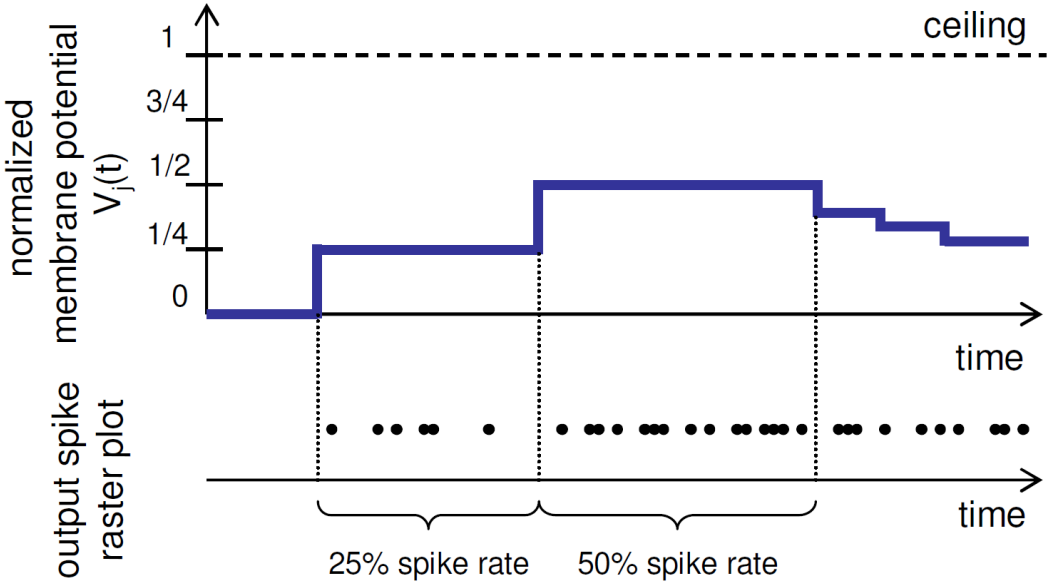
- Can grow the threshold by a random number
- Can set the max of this delta with a mask parameter
- Three reset modes:
  - Normal: after firing, the new potential = reset voltage
    (residual is discarded)
  - Linear: after firing, the new potential = residual voltage
    (reset voltage is not used)
  - Non-reset mode: no reset. Relies on leak and –ve synapses to lower the potential

Synthetic Functions

- Library of 50+ synthetic neuron behaviors (compatible with many coding techniques)

- Example: various arithmetic ops, rate store neuron

Rate store neuron: acts as a memory that is constantly being read (e.g., a state holding recurrent neuron); output freq a potential; uses a stochastic threshold and non-reset mode;



Arithmetic Ops

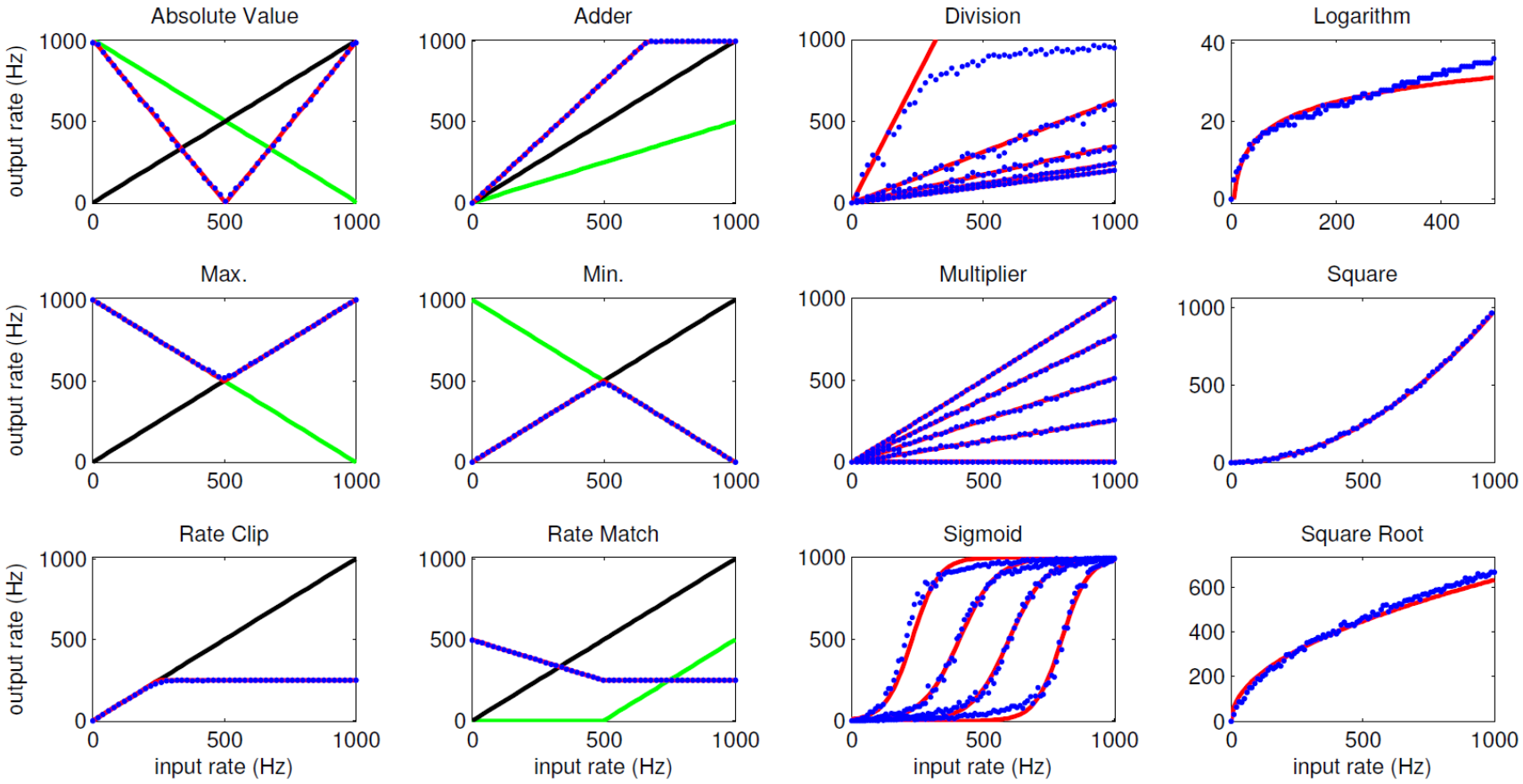
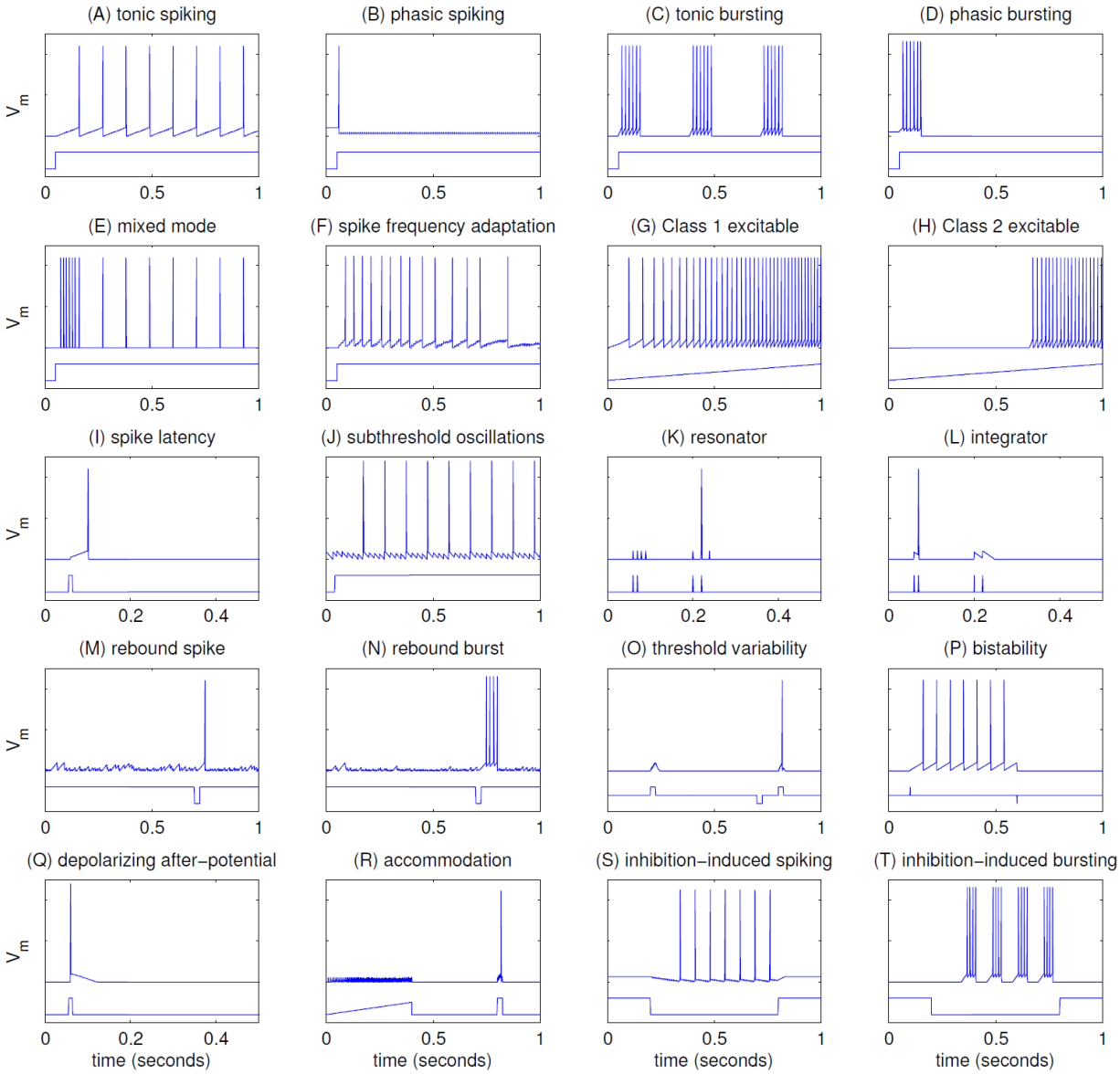


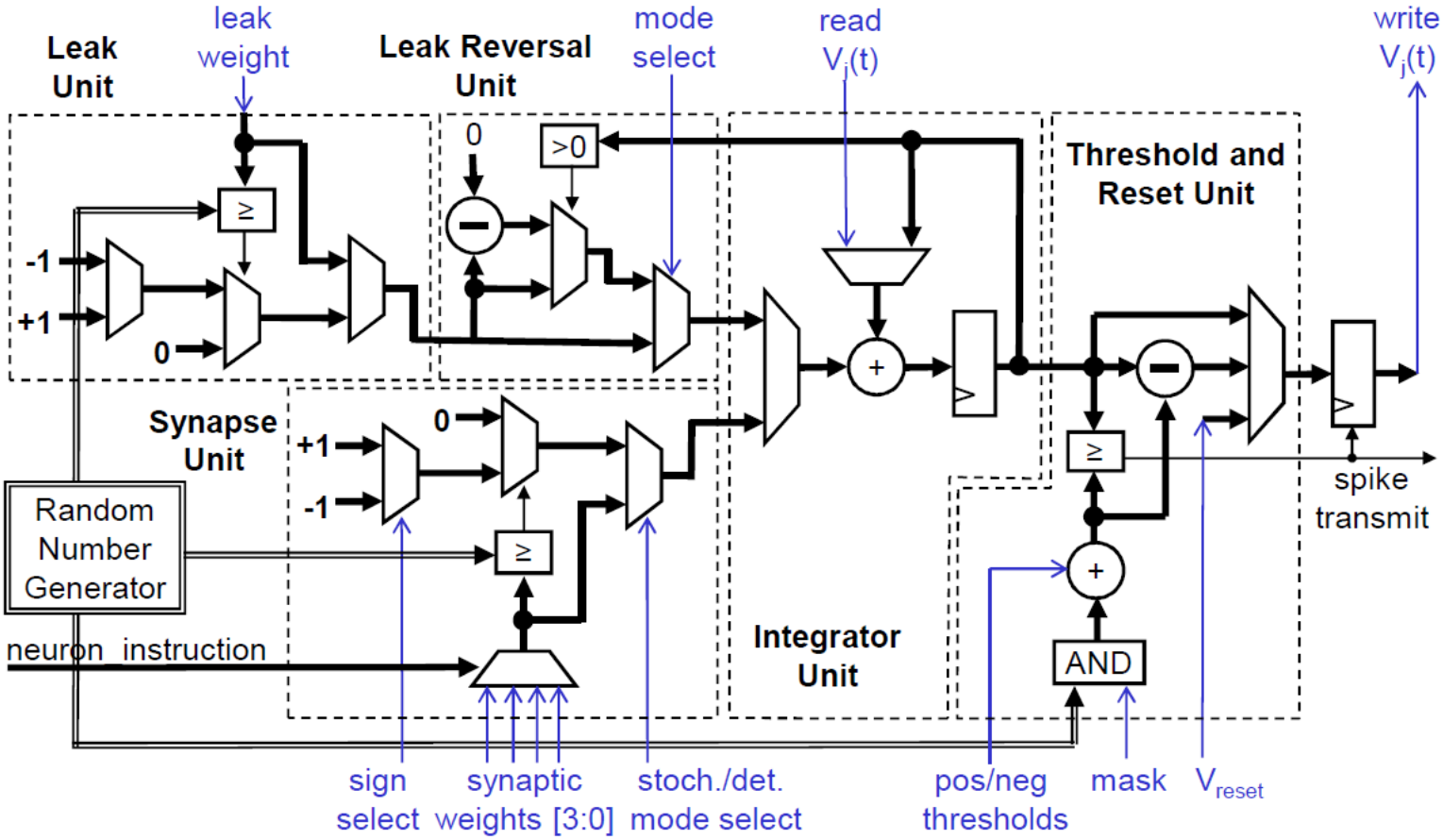
Fig. 8. Arithmetic functions (left to right, top to bottom): absolute value = abs(green-black), adder = (green+black), division, logarithm, maximum = max(green,black), minimum = min(green,black), multiplication, square, rate clip = min(black,250), rate match = 500-abs(green-black)/2, sigmoid, and square root. Actual output response is shown in blue, expected response in red. Green and black points are inputs.

Izhikevich Neuron Models



Neuron Model

- Need 924 gates for neuron computation and 348 for PRNG



TrueNorth Networks

- Could use rate coding and STDP/back-prop to implement networks on TrueNorth (most early TrueNorth papers)

- Many hardware constraints; must therefore use either train-then-constrain or constrain-then-train (the latter is usually better)
  - Only 4 synaptic weights
  - A neuron can only receive 256 inputs (narrow receptive field)
  - A neuron can only have one destination axon/type (narrow projective field, although one can replicate a neuron to fix this)

MNIST – NeurIPS'15

- Spikes are discrete, but the probability of spiking is a continuous function learnable through back-prop
- Limited connectivity in each layer, but over a few layers, a neuron can see every input
- Inputs are provided in a single tick (spike with a probability that corresponds to the input value)
- With an ensemble of networks, the outputs correspond to the expected value; accuracy of 99.42% on MNIST

Implementing CNNs on TrueNorth

- Small kernels: a neuron can only have 256 inputs
- Conv layer has to replicate the shared weights
- Can’t have more than 256 kernels
- They use ternary weights and activations
- A single spike (or not) for every pixel in every feature map

Results

	State of the art		TrueNorth best accuracy	
Dataset	Approach	Accuracy	Accuracy	#cores
CIFAR10	CNN (11)	91.73%	**89.32%**	31492
CIFAR100	CNN (34)	65.43%	**65.48%**	31492
SVHN	CNN (34)	98.08%	**97.46%**	31492
GTSRB	CNN (35)	99.46%	**97.21%**	31492
LOGO32	CNN	93.70%	**90.39%**	13606
VAD	MLP (36)	95.00%	**97.00%**	1758
TIMIT Class.	HGMM (37)	83.30%	**82.18%**	8802
TIMIT Frames	BLSTM (38)	72.10%	**73.46%**	20038

Most are with 8 TN chips

References

- "Cognitive Computing Building Block", A. Cassidy et al., IJCNN'13
- "A Digital Neurosynaptic Core Using Embedded Crossbar Memory with 45 pJ per Spike in 45nm", P. Merolla et al., CICC, 2011
- "TrueNorth: Design and Tool Flow of a 65mW 1 Million Neuron Programmable Neurosynaptic Chip", F. Akopyan et al., IEEE TCAD, 2015
- "Real-Time Scalable Cortical Computing…", A. Cassidy et al., SC'14
- "Spiking Neuron Models", W. Gerstner and W. Kistler, Cambridge University Press, 2002