Sandia National Laboratories

SPIKING NEURAL NETWORKS FOR GENERAL PURPOSE COMPUTING

Daniel Puckett and Kaylin Hagopian

MOTIVATION

Moore's Law is slowing down, but compute demands are rising. To increase compute, we are inspired by the brain to replace transistors with neurons. **This transformation is promising because neurons can convey more than a '1' or '0' down a wire** since they are controlled by input charge instead of input voltage (unlike transistors).

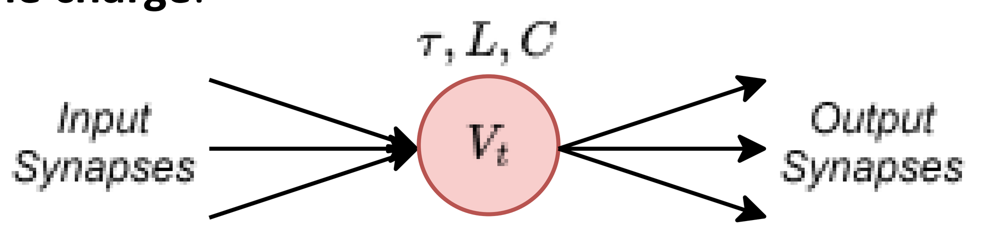
However, neurons are complex, so a neuron will use more power and area than a transistor. Thus, **neurons can only replace transistors if the operation- and system-level benefits from using neurons outweigh their unit-level costs.**

Unfortunately, exhaustively evaluating these tradeoffs is an expensive process in engineering and fabrication costs. We **approximate the operation-level benefits**, identifying if further investigation is prudent, by:

1) adapting an existing neuron model to support variable-charge spikes;
2) encoding integers in spikes and using our model and encoding to design neuron-based adders and memories;
3) analytically comparing our adders and memories to CMOS-based circuits to approximate the minimum area, power, and latency a neuron needs to outperform transistors.

NEURON AND SYNAPSE MODELS

We propose the **Overflow neuron**, which extends the leaky integrate-and-fire (LIF) neuron by **creating spikes with variable charge**:



$$V_t = \begin{cases} 0, & if \hat{V}_t >= \tau \\ \hat{V}_t(1-L), & Otherwise \end{cases} \qquad \hat{V}_t = V_{t-1} + \sum C^i_{spike} \qquad C^o_{spike} = \hat{V}_t - \tau + C$$

Additionally, we extend synapses to support variable charge spikes:

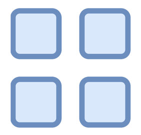
$$C_{out} = C_{in} \times W + C$$

The variables used above are defined in the table below:

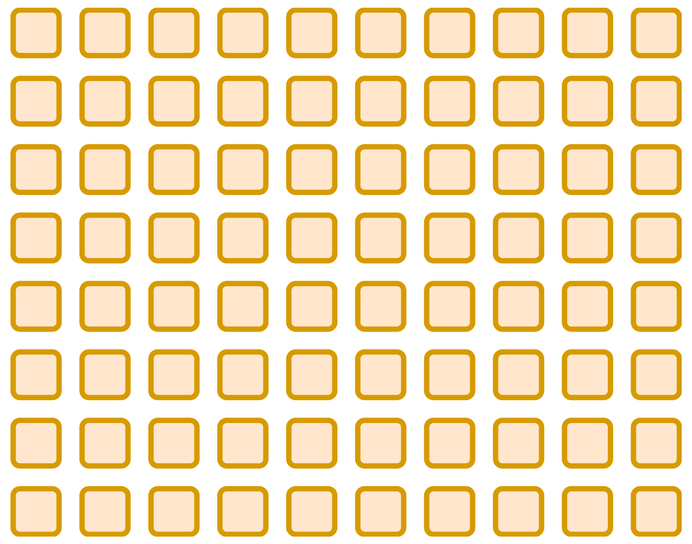
C	Charge boost (neuron or synapse)	C^o_{spike}	Charge of spike leaving neuron
τ	Neuron's threshold	C_{in}	Charge entering synapse
L	Neuron's leakage	Δ	Synapse's delay
V_t	Neuron's voltage	W	Synapse's weight
$\hat{V}_t$	Intermediate value	C_{out}	Charge leaving synapse
C^i_{spike}	Charge of spike entering neuron		

Computers built with **neurons** may be **more efficient than** those built with **transistors**…

NUMBER OF NEURONS IN 2-BIT ADDER

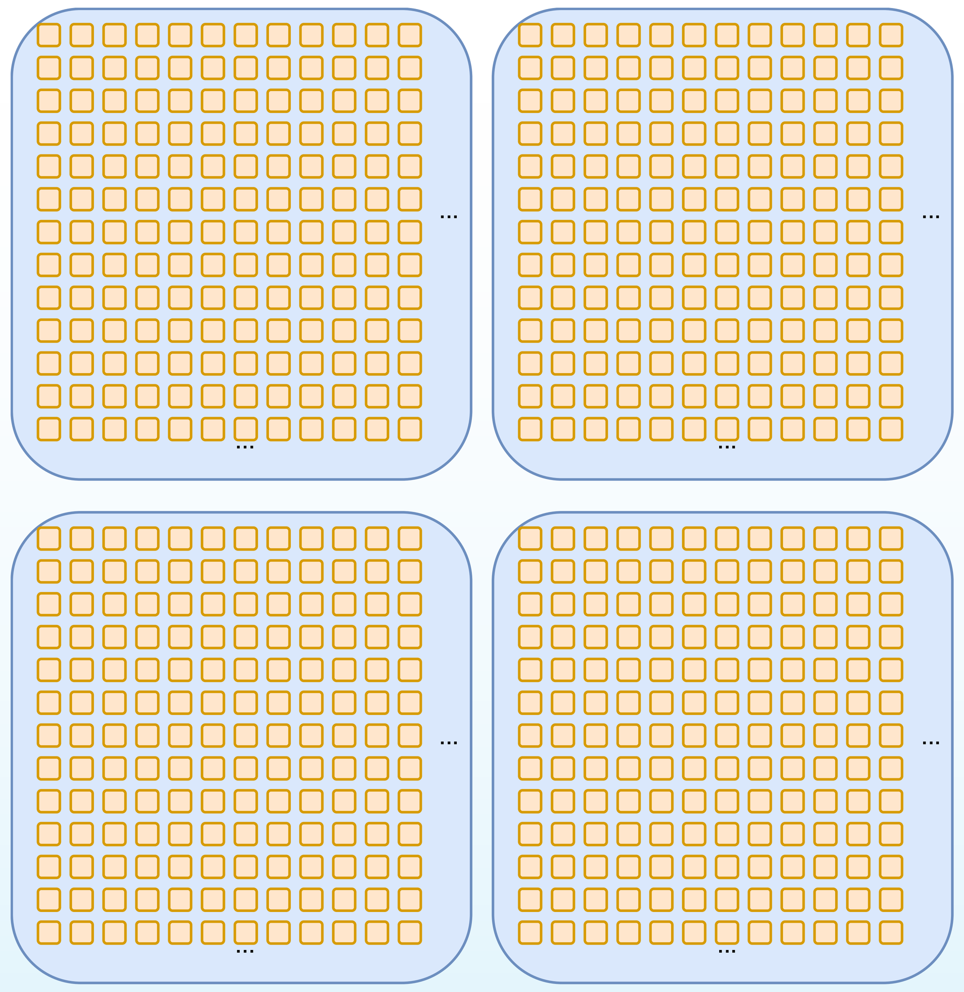


NUMBER OF TRANSISTORS IN 2-BIT ADDER

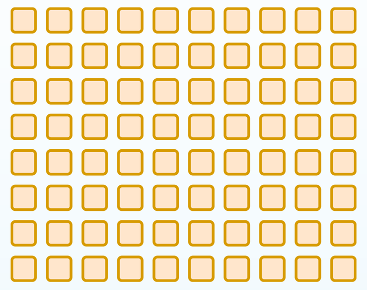


… so long as neurons are **smaller than 20 transistors**

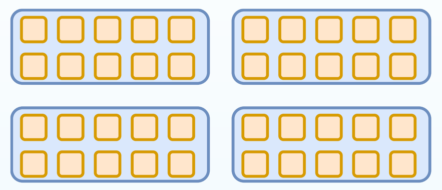
2-BIT ADDER WITH LOIHI



2-BIT ADDER WITH TRANSISTORS



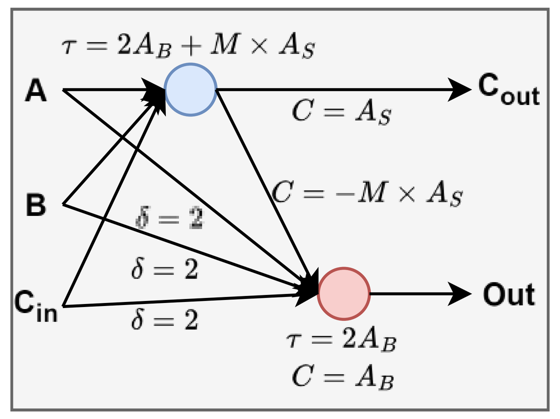
2-BIT ADDER WITH 10-TRANSISTOR NEURONS



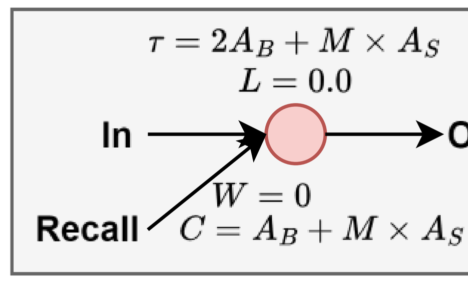
INTEGER ENCODING

We construct a new integer encoding that utilizes **variable-charge spikes** and implement addition and memory operations in this encoding.

The encoding is a **positional number system** where each spike represents a digit and the spike's charge represents the value of the digit. The spikes are arranged spatially such that a number with *S* digits is conveyed in a single timestep using *S* synapses and *S* spikes. The single-digit addition and memory operations for this encoding are shown below.



Single-Digit Adder



Single-Digit Memory

M is the number of values each spike can represent, A_B is the base charge of a spike and A_S is the charge between two adjacent numbers. The blue circle represents a LIF neuron and the red circles represent Overflow neurons. Unless otherwise specified, δ is 1, W is 1, C is 0, and L is 1.

COMPARISON TO TRANSISTORS

To identify if neurons' operation-level benefits outweigh their unit-level costs, **we compare our neuron-based adder and memory to 32-bit CMOS adders and memories.**

We set *M*, the number of values each digit can represent, to three. We set *S*, the number of digits, to 21 for a total range greater than 10 billion, which is 2.5x larger than the range of a 32-bit integer (4 billion).

Despite this, the neuron-based operations analytically require **less area** (fewer neurons than transistors), **less power** (fewer spikes than transistor flips), and **less latency** (fewer synapse than transistor delays), than CMOS-based operations, as shown below and in the center pane.

Operation	Area	Power	Latency
Addition	21.3x	21.2x	2.9x
Memory	9.0x	3.0x (read) / 4.6x (write)	1.0x

CONCLUSION

Our research finds that neurons' operation-level benefits may outweigh their unit-level costs if neurons can be fabricated smaller than 20 transistors. While this indicates **computers built with neurons may be more efficient than those built with transistors**, additional work on fabricating neurons, building accurate neuron models, and understanding system-level benefits and challenges is needed.

U.S. DEPARTMENT OF ENERGY
NNSA National Nuclear Security Administration

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

This work was funded by NNSA's Office of Defense Nuclear Nonproliferation.

TEXAS A&M UNIVERSITY
Department of Electrical & Computer Engineering

Sandia National Laboratories