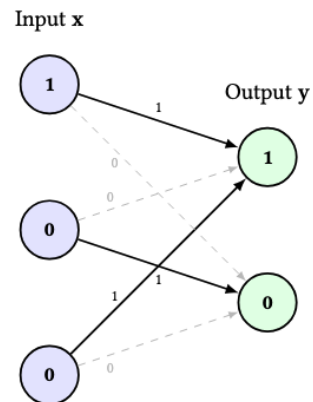
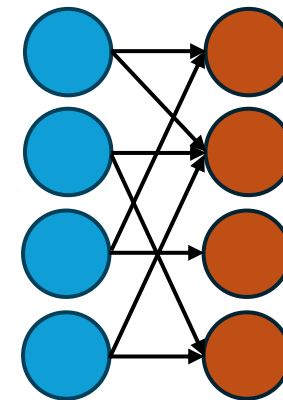


Bridging Neuromorphic and Traditional Computing Performance: *An Information-Theoretic Approach*

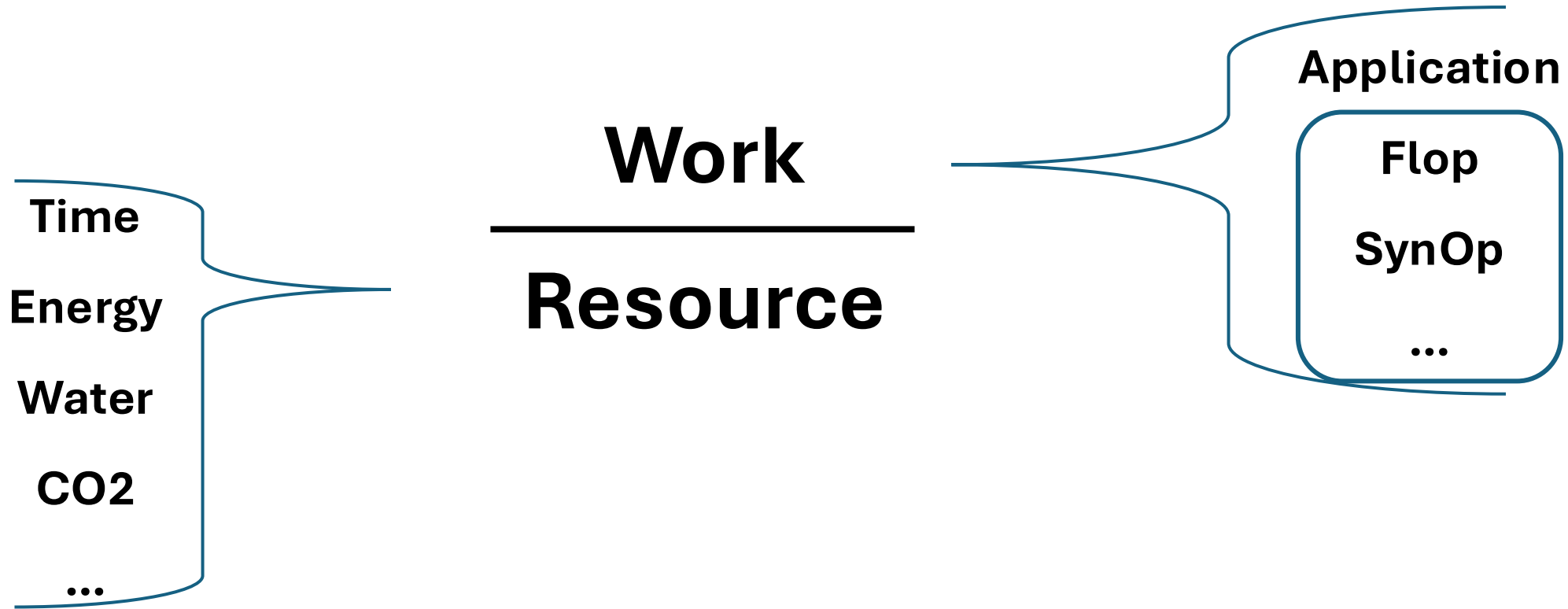
Max Hawkins and Richard Vuduc



$$\begin{matrix} & \mathbf{W} \\ \begin{matrix} 1 & 0 & 1 \\ 0 & 1 & 0 \end{matrix} & \times & \begin{matrix} \mathbf{x} \\ 1 \\ 0 \\ 0 \end{matrix} & = & \begin{matrix} \mathbf{y} \\ 1 \\ 0 \end{matrix} \end{matrix}$$



How do we model and compare
the performance of computers?



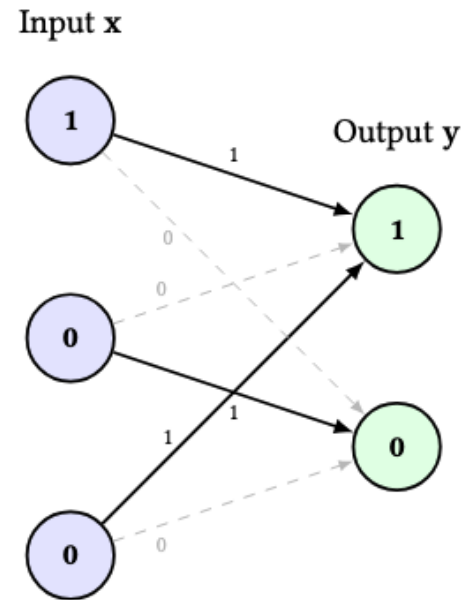
My PhD thesis:
Towards a general model of computational work.

This talk:
Sparsity and Noise

Sparsity

What's the computational work needed for a single pass?

- 3 different answers/frameworks:
 - 6 multiplies – dense/HPL
 - 3 multiplies – sparse/HPCG
 - 1 ‘effective synOp’ – NeuroBench
- No agreement on sparse work!



$$\begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \times \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

(b) Linear Algebra View

Sparse Operation Accounting

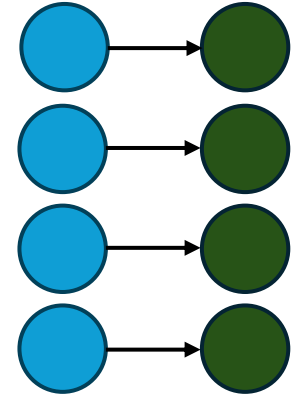
- Nvidia structured sparsity flop accounting:
 - “[We count] the multiply-by-zero operations, even though they are not actually performed in hardware” - <https://forums.developer.nvidia.com/t/dense-vs-sparse-tensor-core-performance-fp16/314261>
- U.S. Gov’t computing export controls:
 - “For integrated circuits ... , the TOPS values are the values for processing of dense matrices (e.g., without sparsity).” - <https://www.federalregister.gov/d/2023-00888/p-126>

**We don’t have a standard treatment of sparsity,
and that prevents fair performance comparison of neuromorphic hardware.**

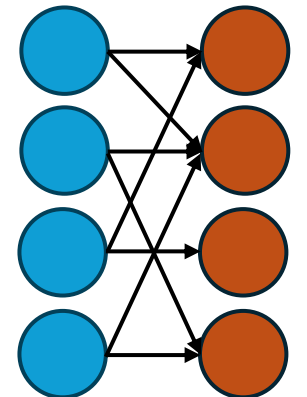
Noise/Determinism

- Is noise a failure mode or something to design around?
 - ‘Traditional’ computing: Failure
 - Neuromorphic: We’ve got this.
- Per-application error tolerance thresholding
 - Not general and binary success/fail

How do you compare the value of two otherwise identical operations: one noisy, one deterministic?



VS



Back to Bits

- HW channel: Maps input states to output states
- Extend Shannon's communication theory to computation
 - Identity function --> Arbitrary function
- Measure of work: Mutual Information (I)
 - How much of the output's uncertainty does the input remove?

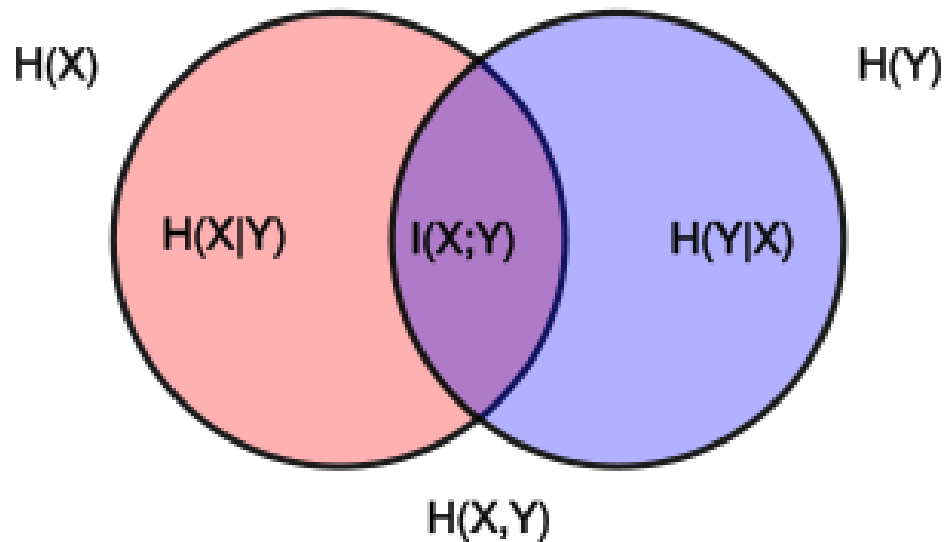
$$I(X; Y) = H(Y) - H(Y|X)$$

- Applies to arbitrary encoding formats
 - Spike trains, FP64, smoke signals, etc

Proposal: Computational work is uncertainty reduction.

Noise Revisited

- Mutual information naturally accommodates noise
- Deterministic ALU: $H(Y|X) = 0$
- Noisy neuron: $H(Y|X) > 0$



$$I(X; Y) = H(X) - H(X|Y)$$

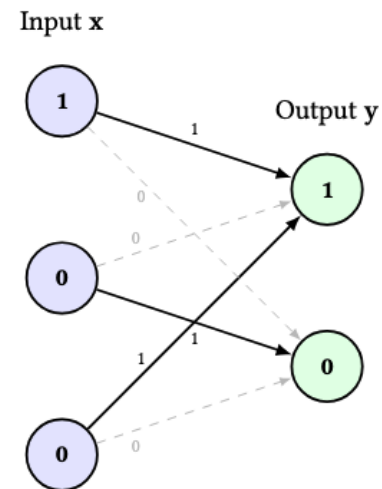
Sparsity Revisited

- Structural Sparsity
 - No a-priori uncertainty in a zero-weight output
 - No benefit of computing it ($H(Y) = 0$, so $I(X; Y) = 0$)
- Distributional Sparsity
 - Considers the reduced uncertainty of a variable

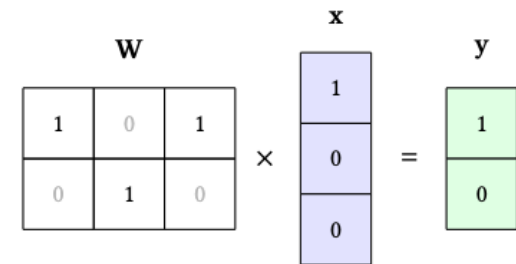
- Resolving our example:

Mutual Information: $3 * 0.811 = 2.43$ bits

effective synOp' – NeuroBench



(a) Network View



(b) Linear Algebra View

*Assuming deterministic hardware and $p(y = 0) = 0.75$

What is your useful model of neuromorphic computing hardware?

Thank You

mhawkins60@gatech.edu

richie@cc.gatech.edu

Back to Bits paper:

<https://arxiv.org/pdf/2508.05621>

