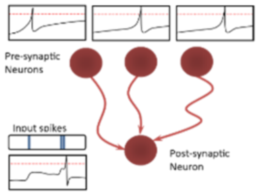
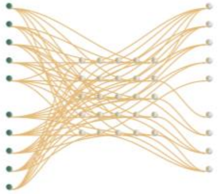
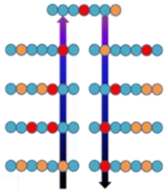
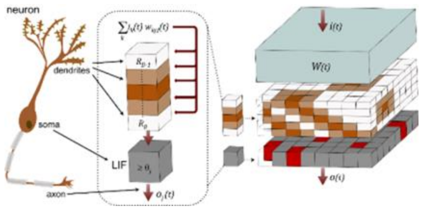
Comparing Neural Accelerators & Neuromorphic Architectures – The False Idol of Operations



Craig M. Vineyard, Mark Plagge, Sam Green, & William M. Severa

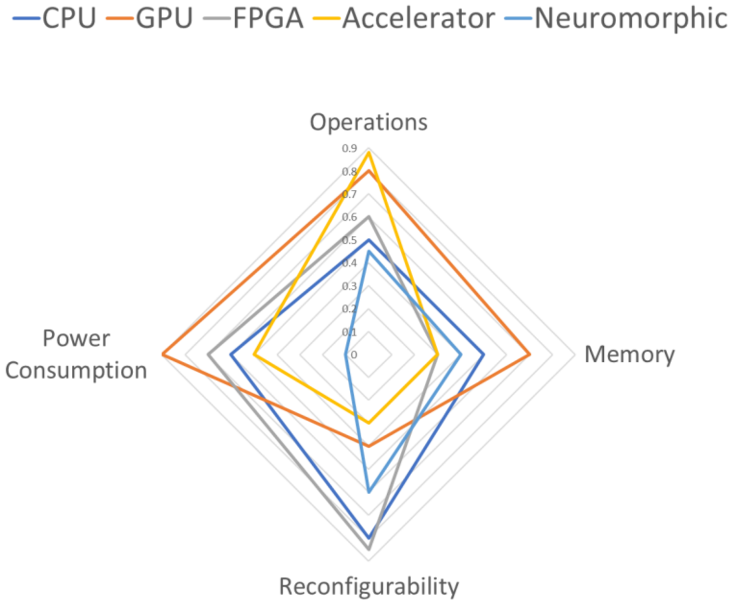
cmviney@sandia.gov

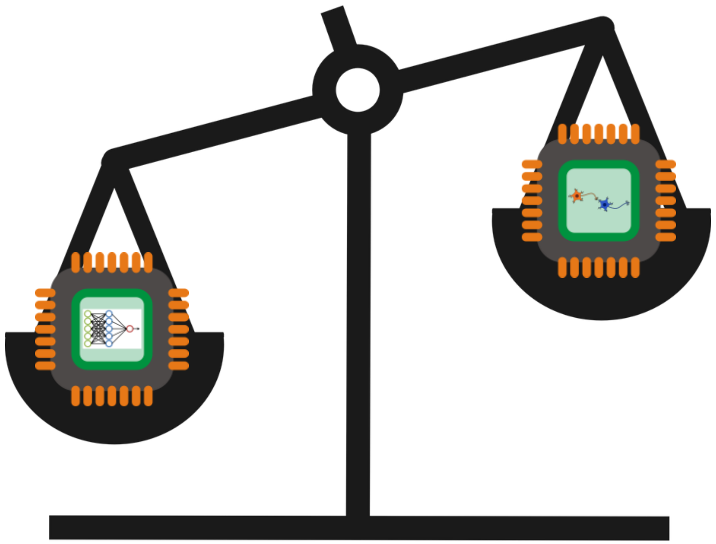
Architectural Comparison

- Neural accelerators & neuromorphic approaches are emerging at different scales, resource requirements, and enabling capabilities

- Beyond the similarity of executing neural network workloads, these two paradigms exhibit significant differences

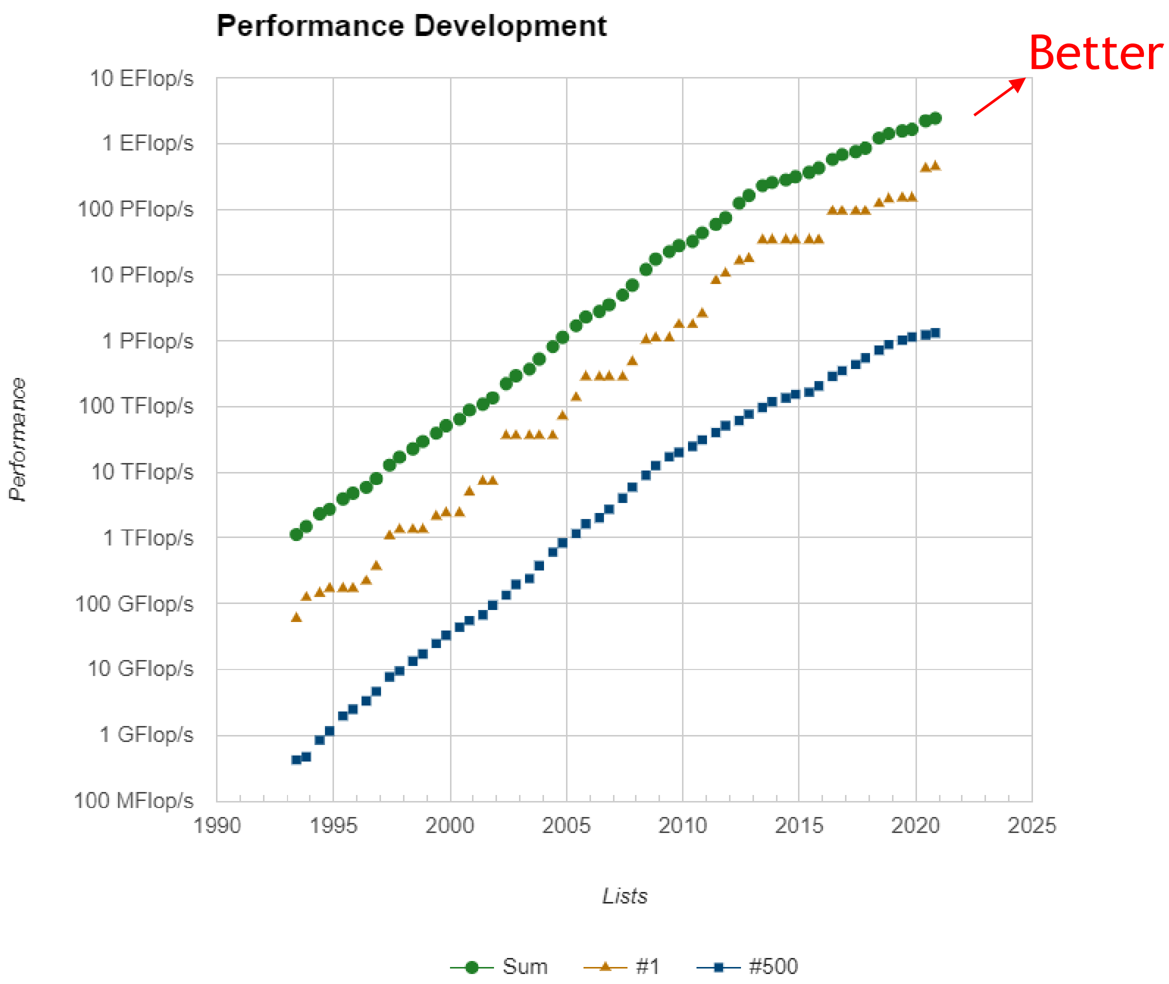
- As processing, memory, and communication are the core tenets of computing, here we compare architectures of neural accelerators and neuromorphic in these terms





Operations

There has been a trend of measuring "better" by the amount of operations

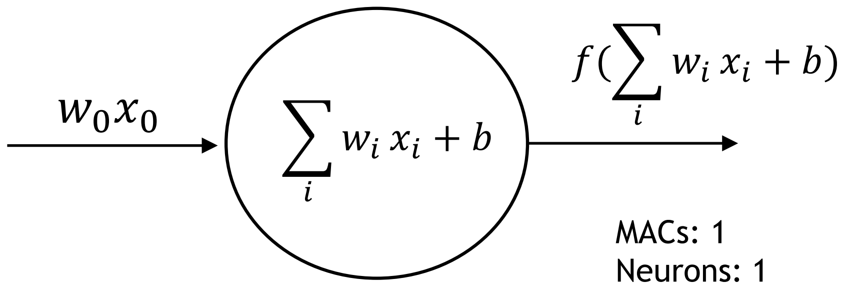


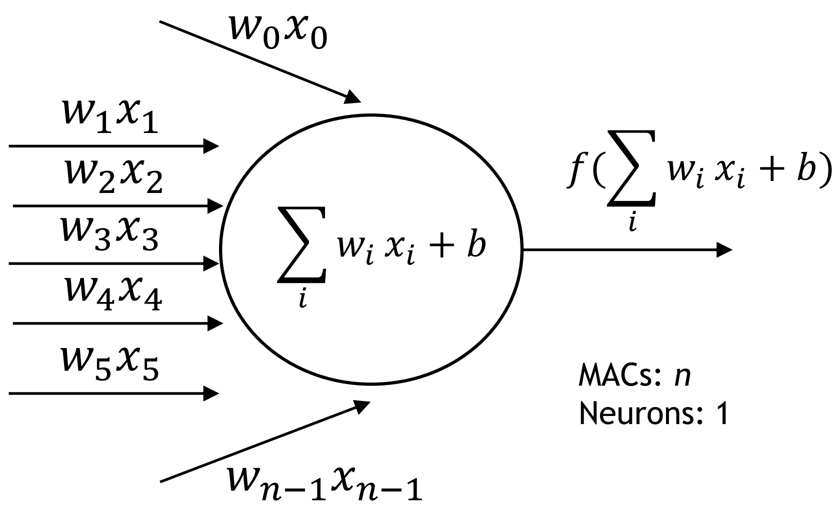
https://www.top500.org/statistics/perfdevel/

Operations

Operation counts alone can be misleading

- In neural networks do not guarantee how accurate your answer will be
- Do not measure how fast your problem will be solved

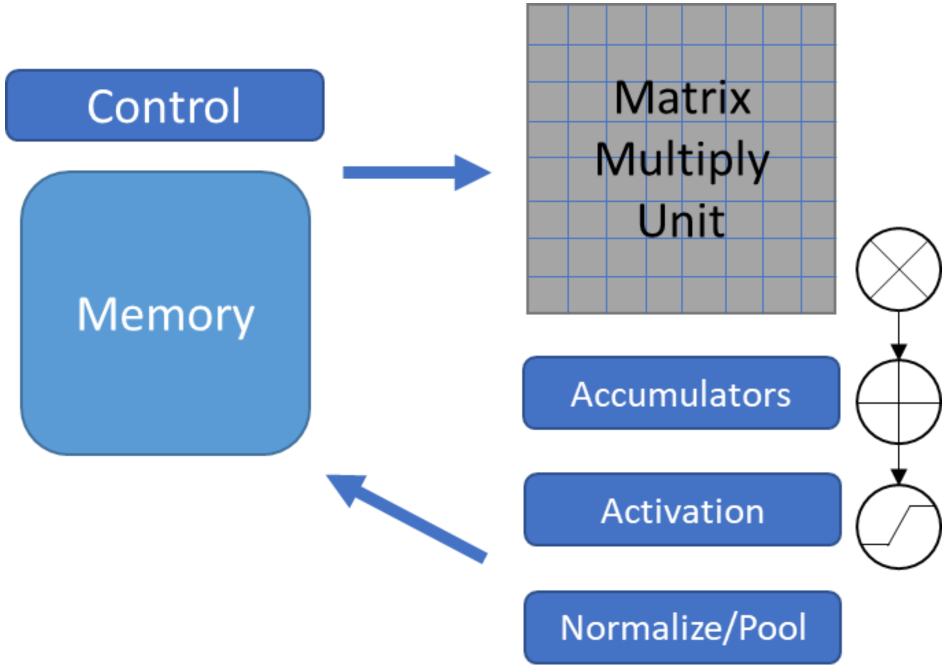


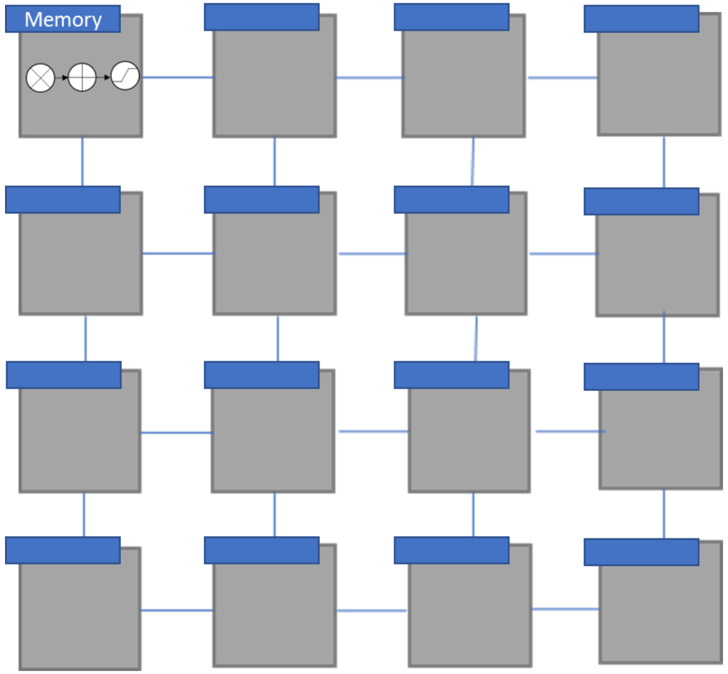


Operations

Emphasis on operation counts has impacted some architectural design choices

- Which furthermore impacts algorithm design choices
- Easy to follow the mindset of more



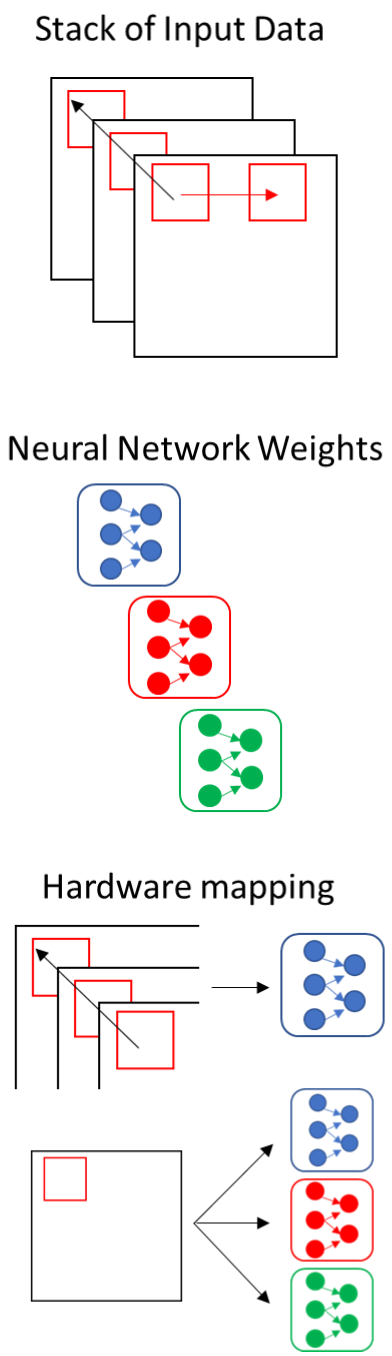


Dataflow

- Dataflow architecture executes computations as data is received
  - Ideologically similar to neural network computation flow
  - Broadly encompasses input data, intermediate computation data, as well as parameter data such as weights and biases

- A dataflow then describes how these various components are moved around in an architecture to perform computation
  - Importantly this matters because data movement from memory access requires more energy than performing computation

- Central to the analysis of how dataflows can bridge computational workflows and architectural execution through the most efficient data movement are the assumptions that data must be moved & that there are limited resources which are being scheduled



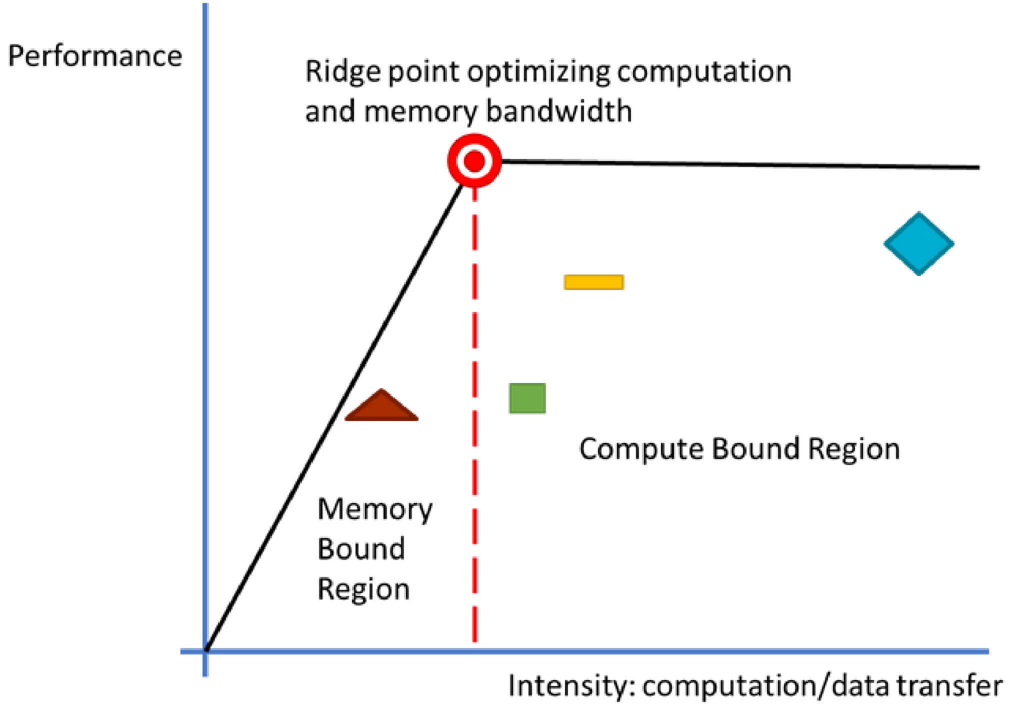
Roofline Model

A roofline model articulates the performance of the interplay between memory and processing for a computational architecture

- Traditionally, the ridge point targets the minimum intensity needed to attain maximum performance

However, we argue alternative computing paradigms can alter the intuition and structure of the roofline model

- While the target traditionally is to optimize towards the ridgepoint, it is possible to be either computer bound or memory bound for neural network computation and still be advantageous



Computational Objectives

Quantitatively assessing aspects of computer architectures has provided an analytical means of exploring the impact of various design choices

- Comparing classes of architectures has often ***relied upon optimizing a shared objective*** despite pursuing different approaches

Comparing neural accelerators and neuromorphic architectures is not as straightforward

- Neural accelerators share design goals of the more traditional computational architectures but focus upon enabling the execution of neural network workloads
- Conversely, neuromorphic approaches strive to enable neural computation but do so by employing design principles of how brains function

Computational Objectives

Instead of assuming more operations is better, ***neuromorphic event-driven computation explores what is the minimum compute needed***

- Analogous to the minimax decision rule from game theory which strives to minimize a maximum cost
- In this context - minimizing the amount of computation needed bounds the maximum cost of computation

This is a fundamentally different paradigm than the converse, maximin which aspires to maximize a minimum gain

- In this context - the objective is to maximize the amount of computation performed to advance the minimal amount of computational progress attained

➢ The best decision is not the same for these two paradigms as they are optimizing for different objectives

Conclusions

As we look to the brain for computing inspiration -

- We know from neuroscience that neuron counts alone are an insufficient measure of cognitive ability
  - For example, the human brain has approximately 86 billion neurons compared with larger brains in elephants consisting of approximately 250 billion neurons
  - Cognitive abilities in biological brains are dependent upon many factors including size, connectivity, surface area, quantity of neurons, support cells, etc.

- Understand the analytical alure to relate architectures based upon operations – BUT novel approaches require understanding their unique benefits

- While the dominant motivational analogy is to compare brains with the power consumption of an ever more efficient lightbulb
  - We should also remember not every neuron fires all the time & aspire to pursue computations not operations

Thank you

Questions?