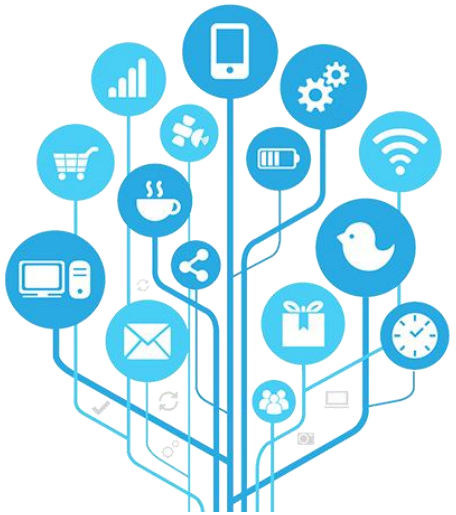
Deep Neural Networks
+
Analog Processing In Memory

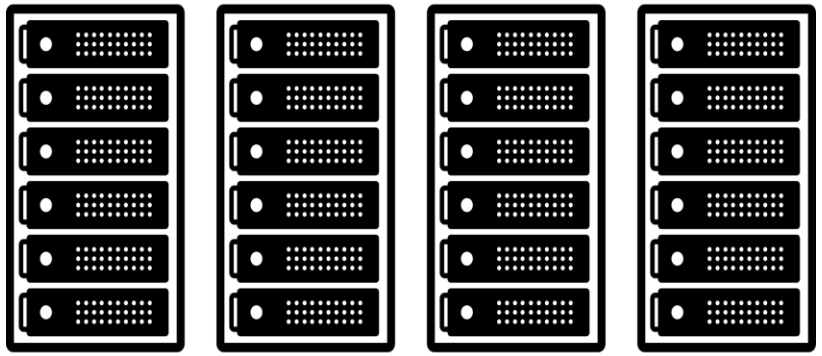
Autonomous Navigation

Internet-Of-Things



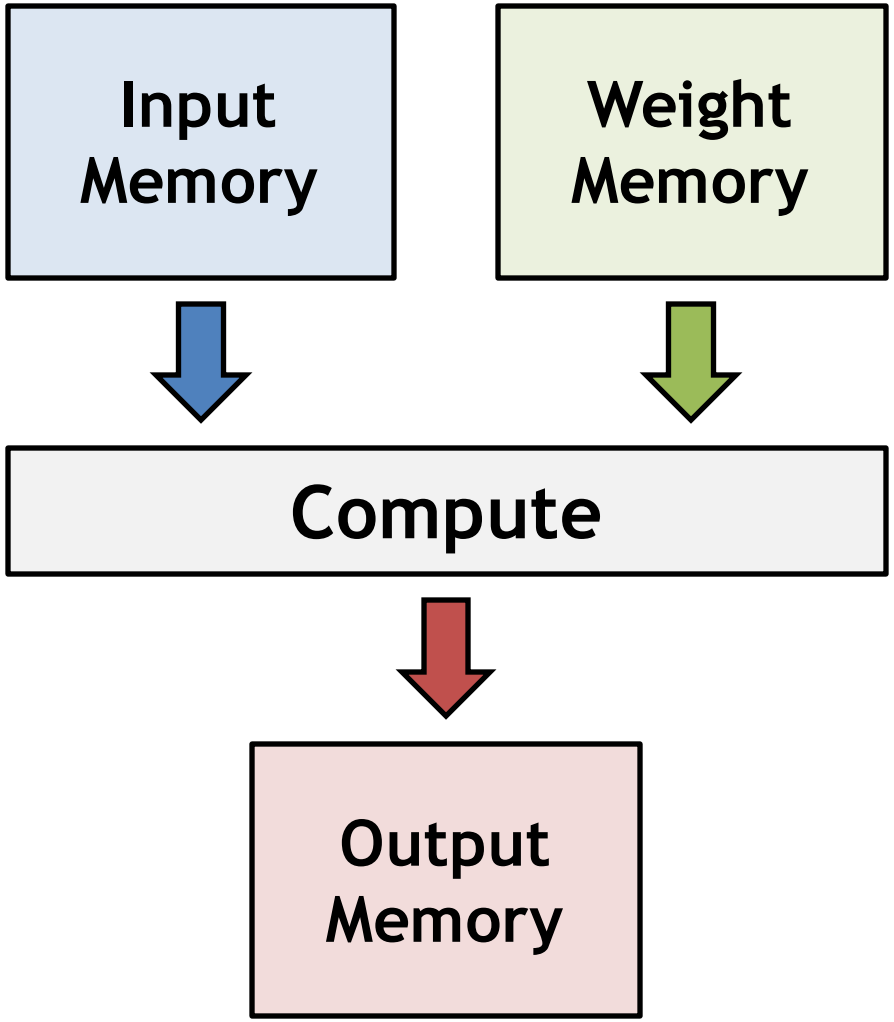
Efficient, High Throughput DNN Inference

Mobile Applications

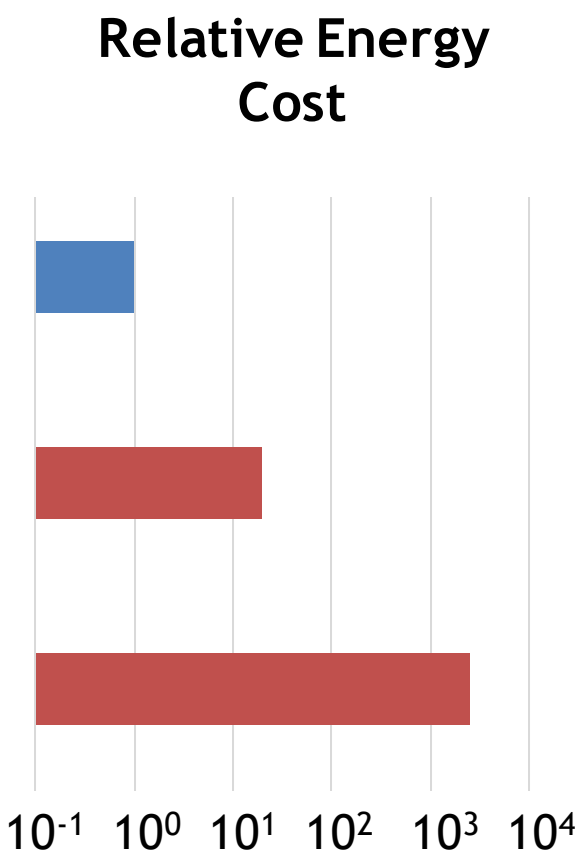


Efficient Datacenters

Conventional DNN Accelerator

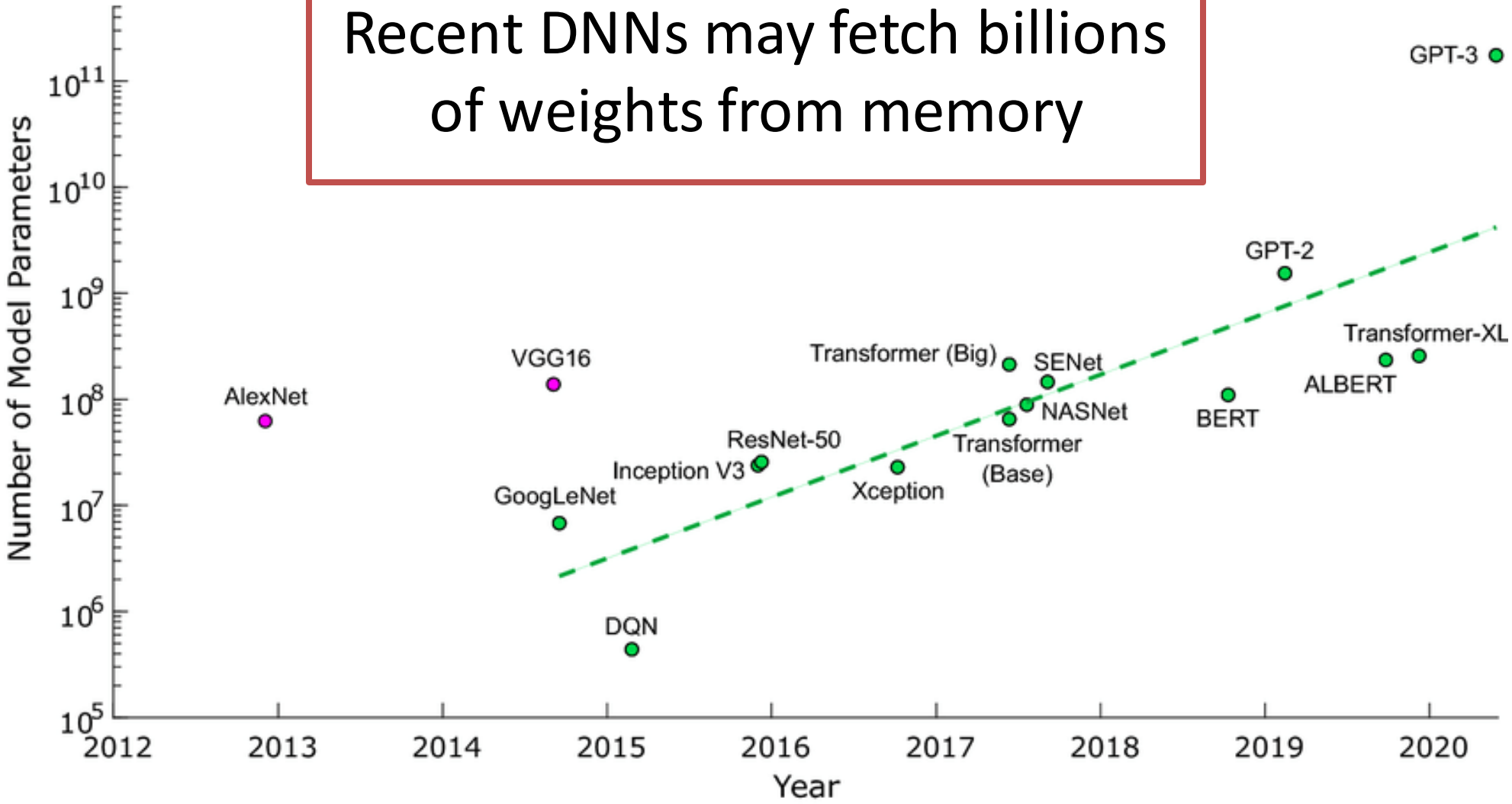


Operation	Energy (pJ)
8b MAC (1 Computation)	0.25
32b SRAM Read (8kB)	5
32b DRAM Read	640



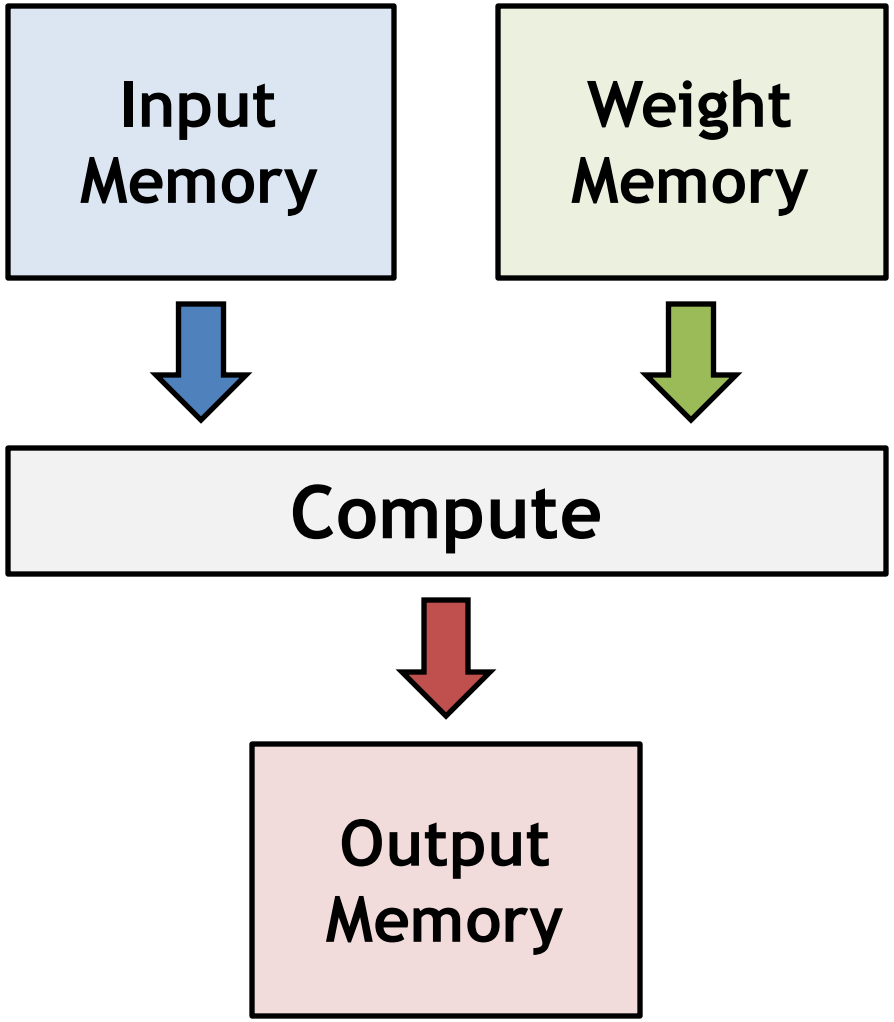
[Horowitz, ISSCC 2014]

Recent DNNs may fetch billions of weights from memory

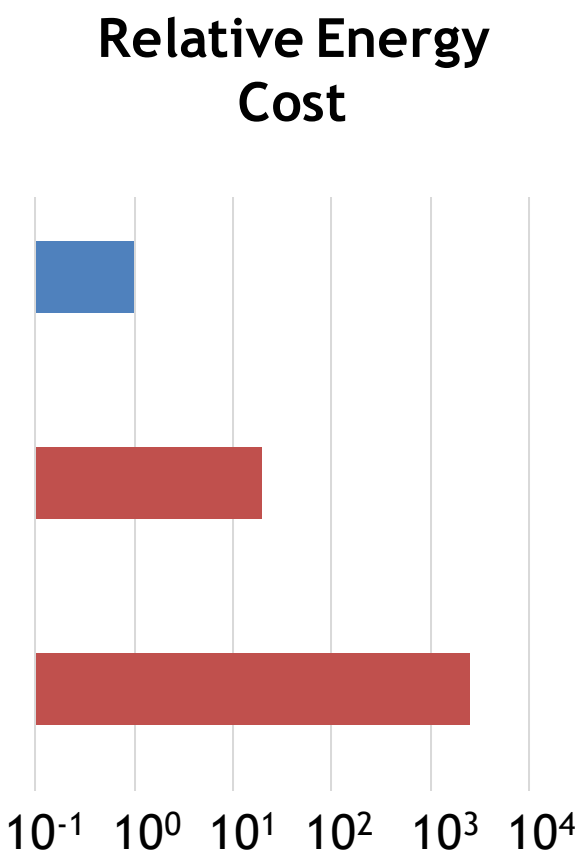


[**Bernstein et al.**, Scientific Reports 2021]

Conventional Accelerator

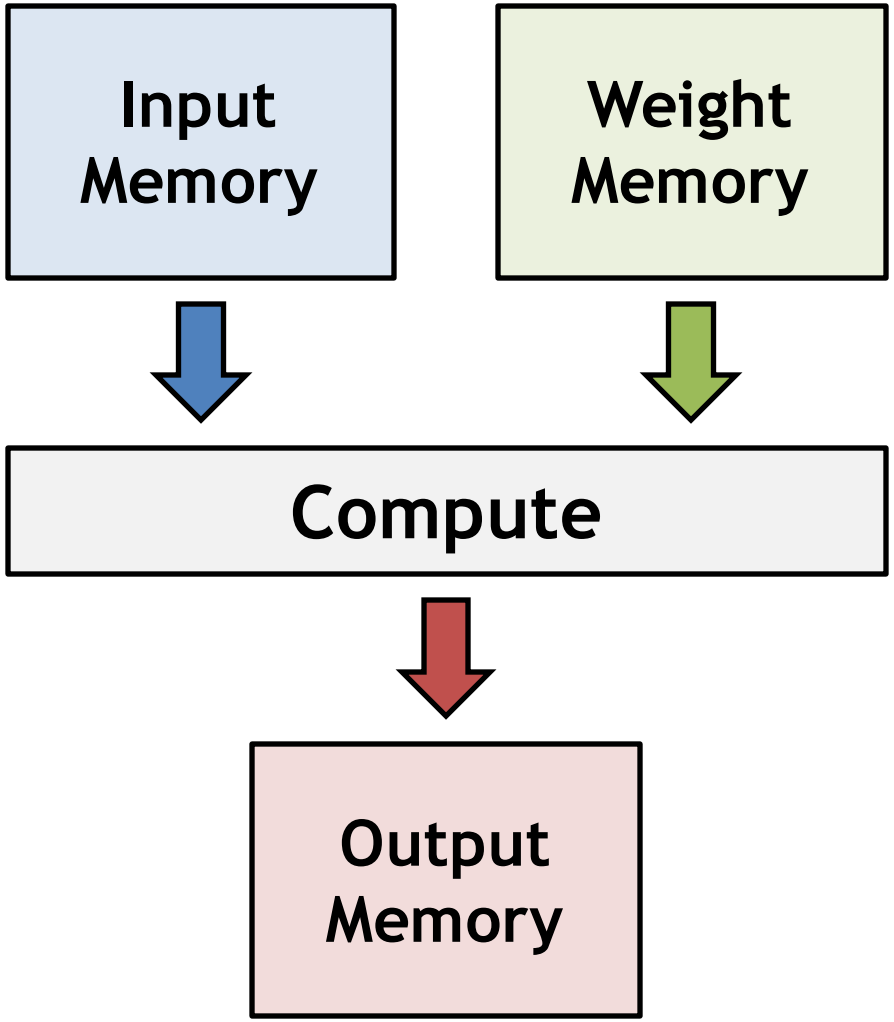


Operation	Energy (pJ)
8b MAC (1 Computation)	0.25
32b SRAM Read (8kB)	5
32b DRAM Read	640

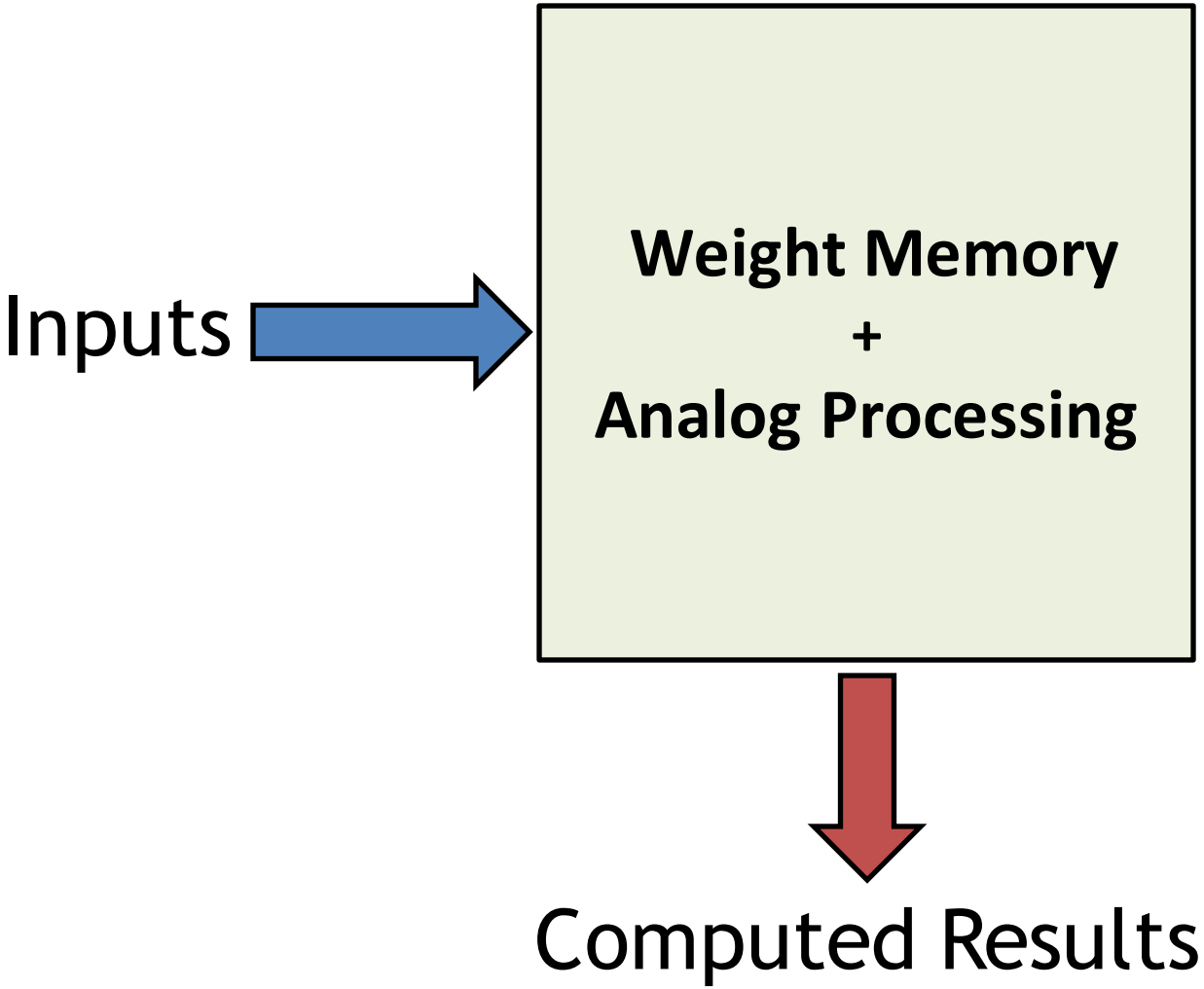


[Horowitz, ISSCC 2014]

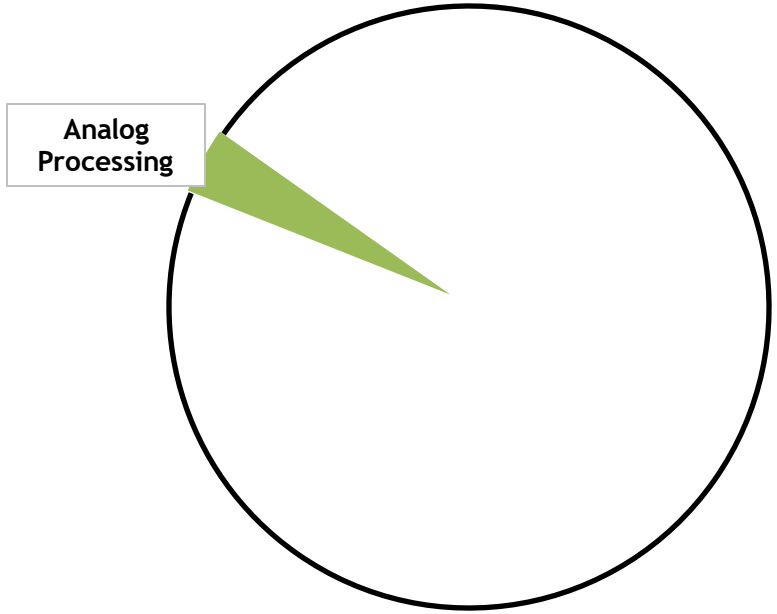
Conventional Accelerator



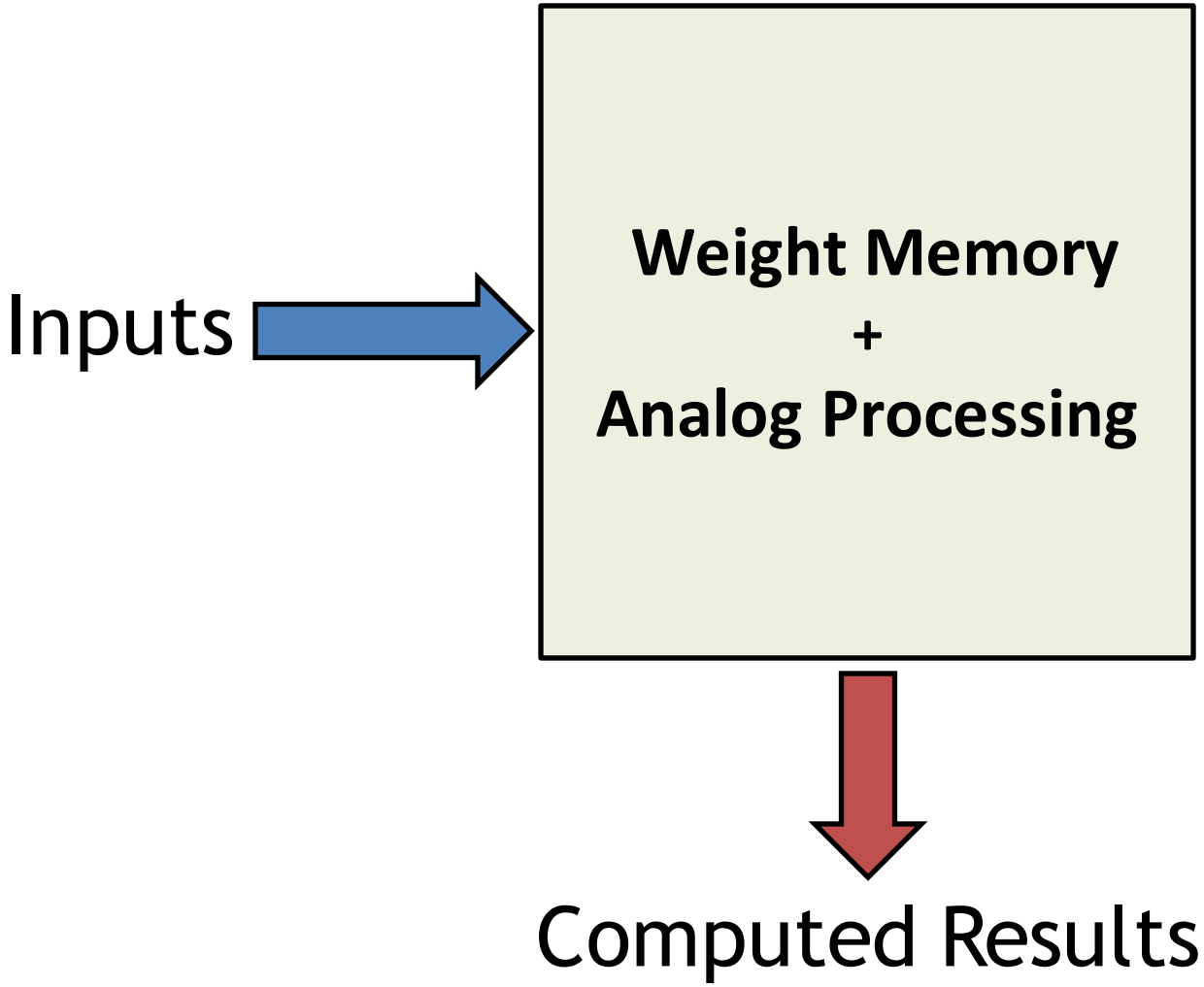
Processing In Memory Accelerator



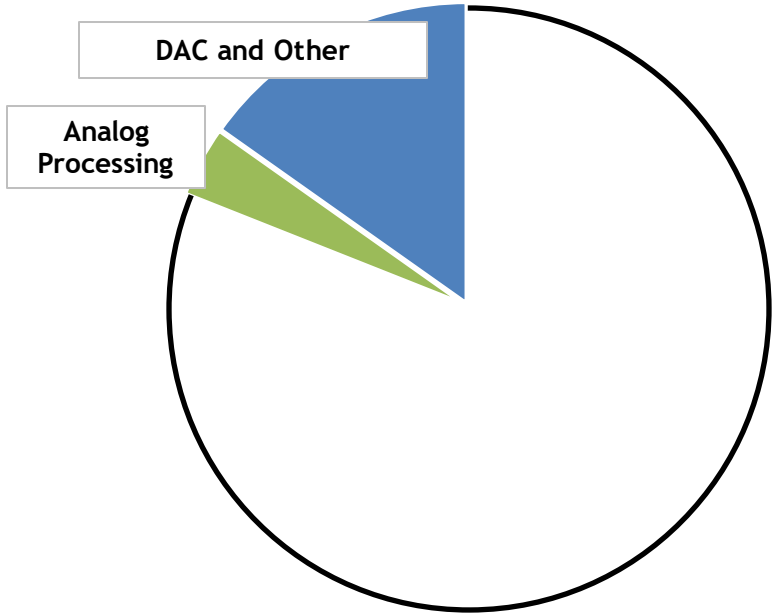
Energy Breakdown



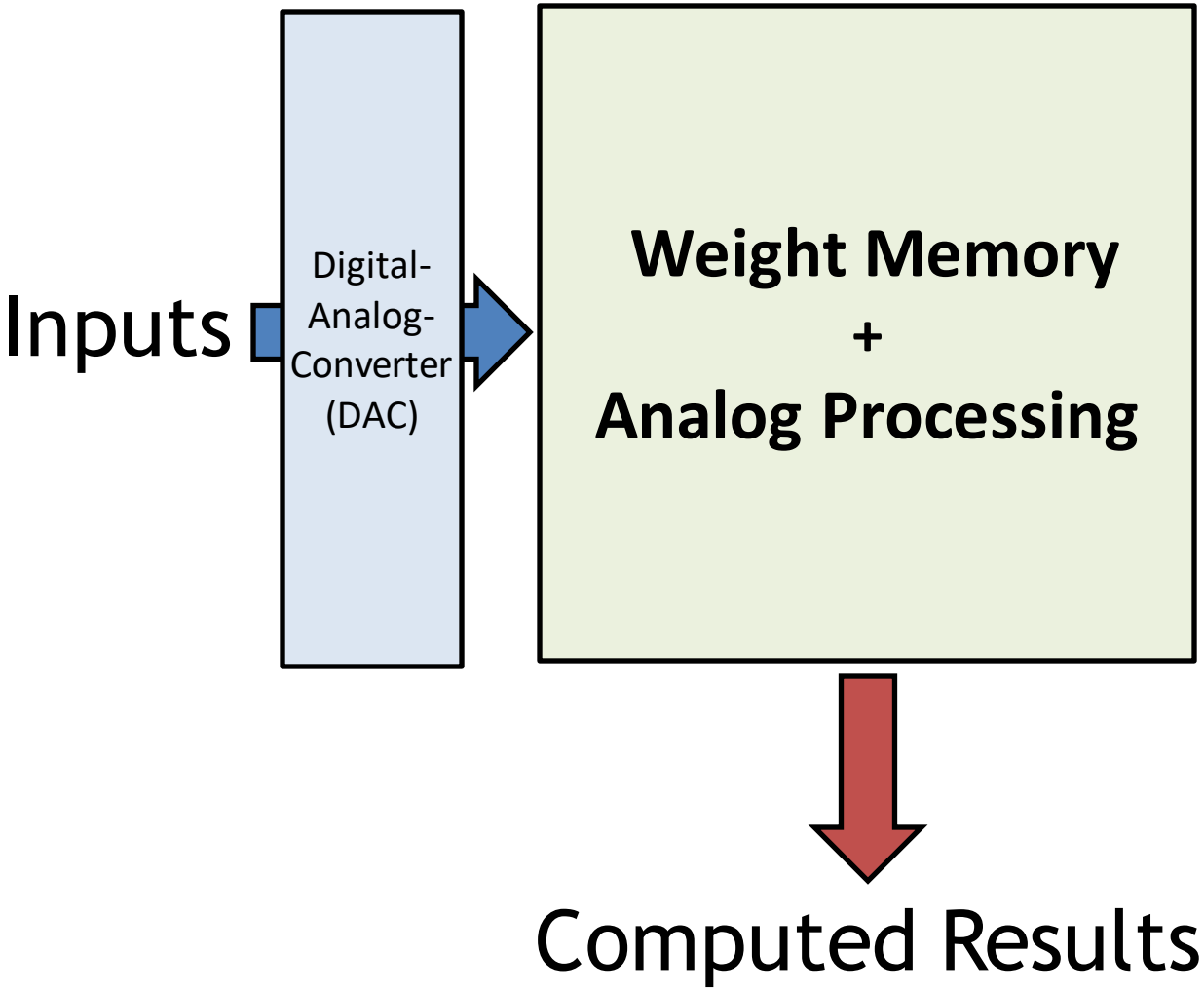
Processing In Memory Accelerator



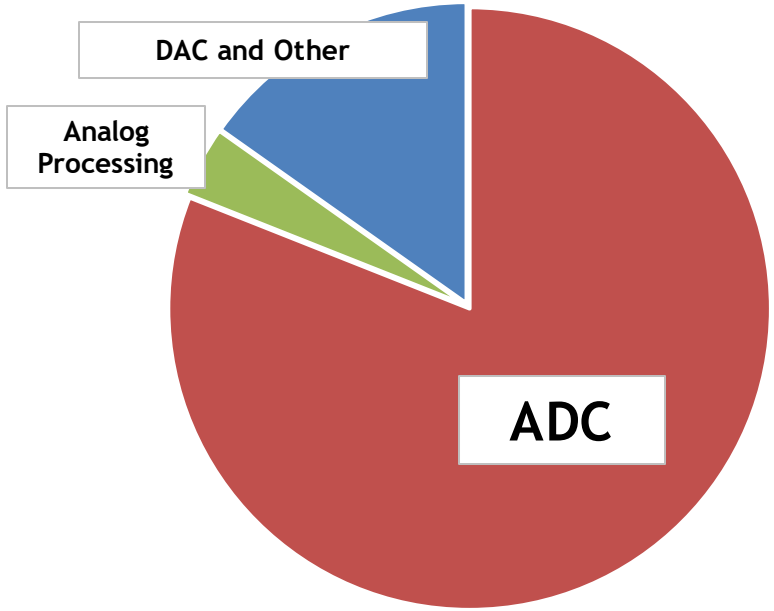
Energy Breakdown



Processing In Memory Accelerator

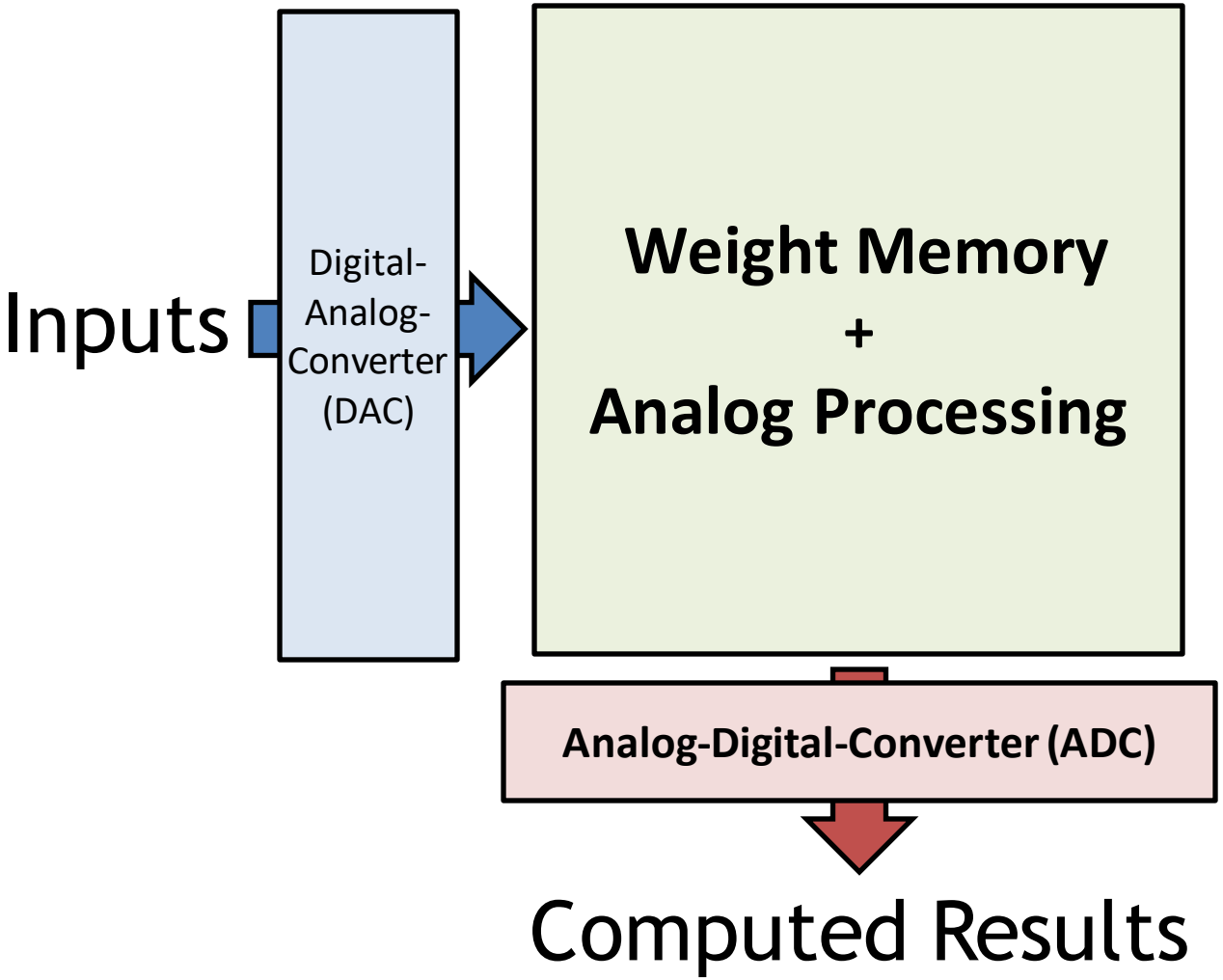


Energy Breakdown

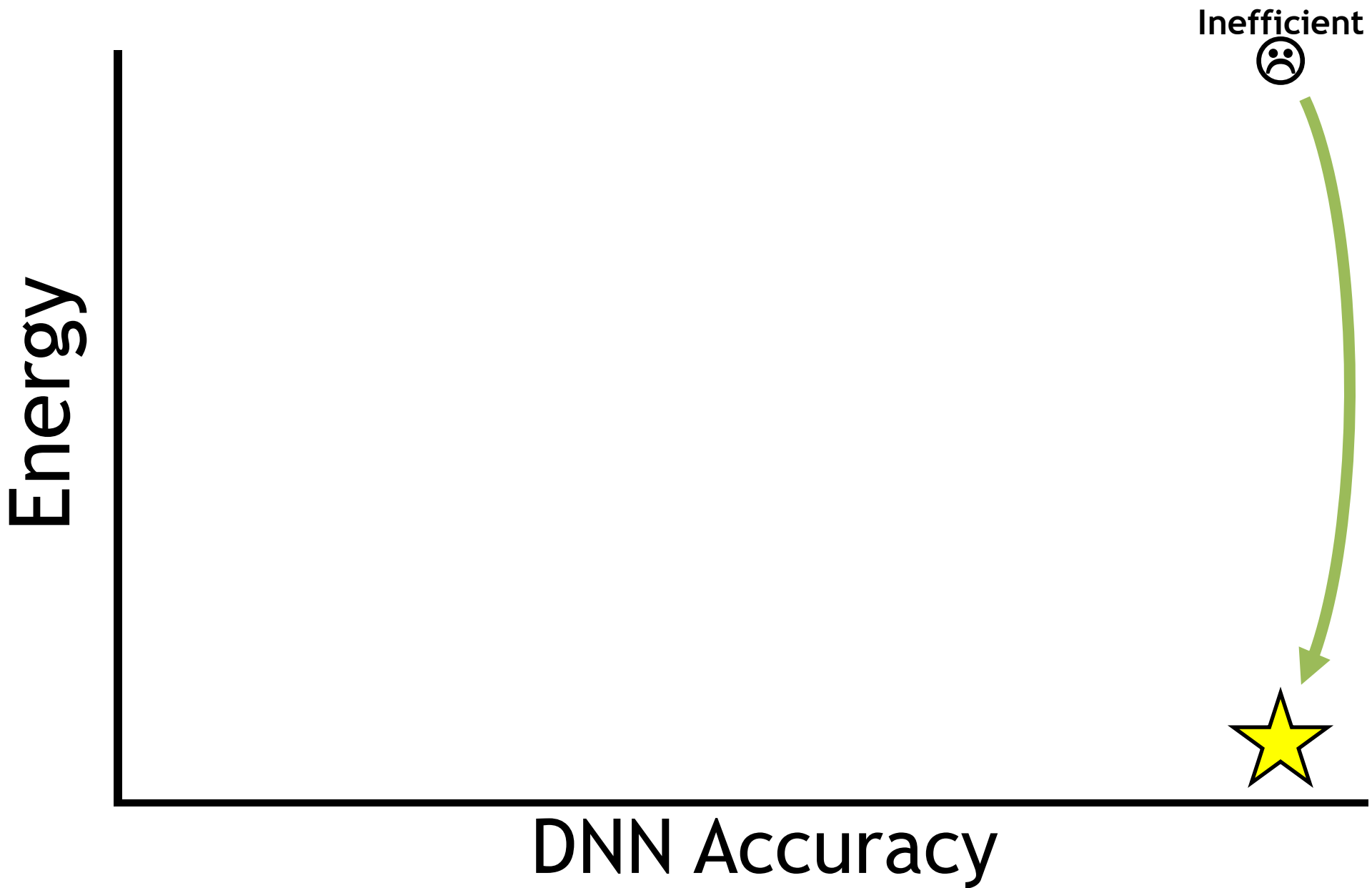


ADC consumes significant energy

Processing In Memory Accelerator

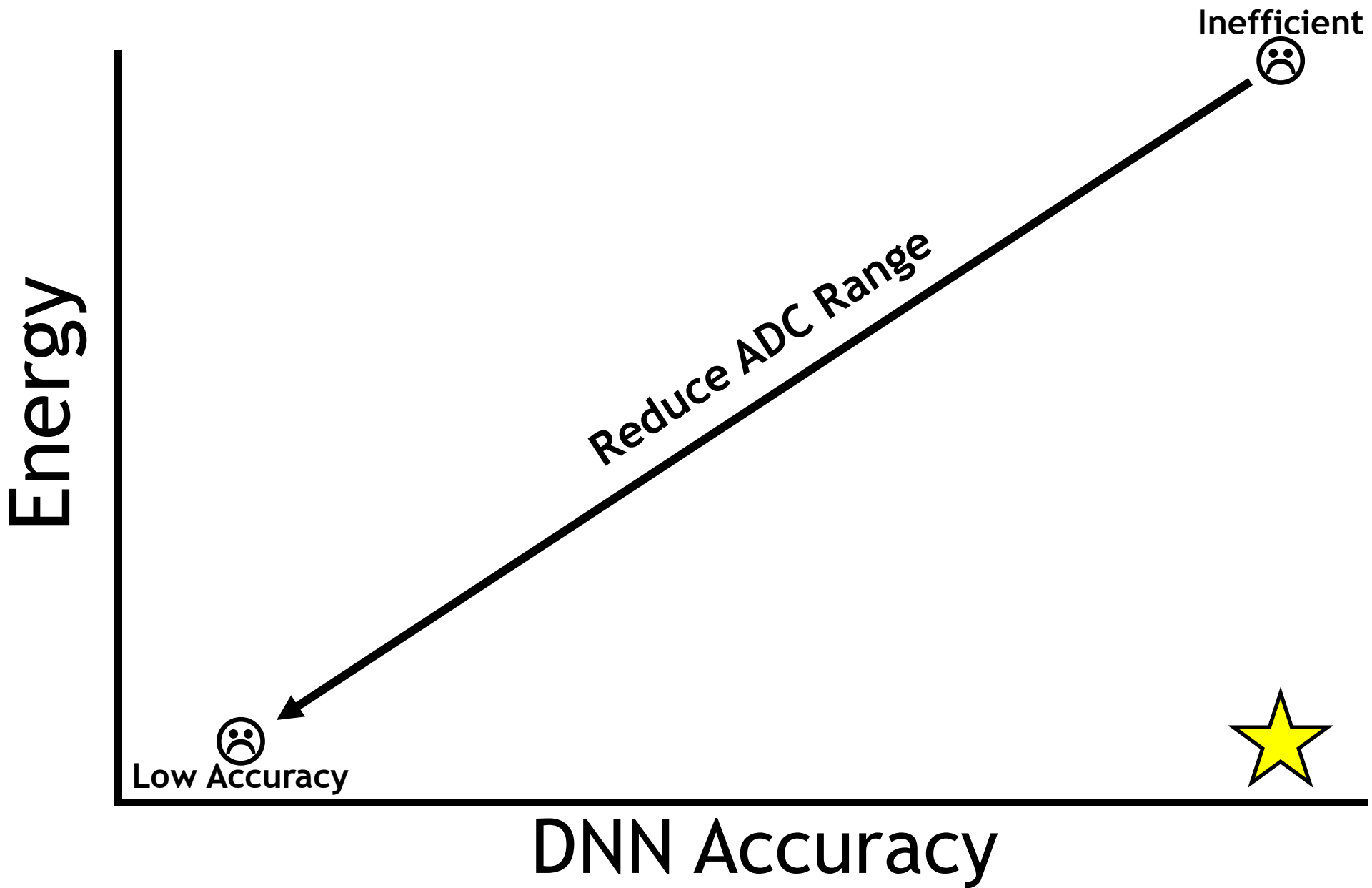


Reducing ADC Energy

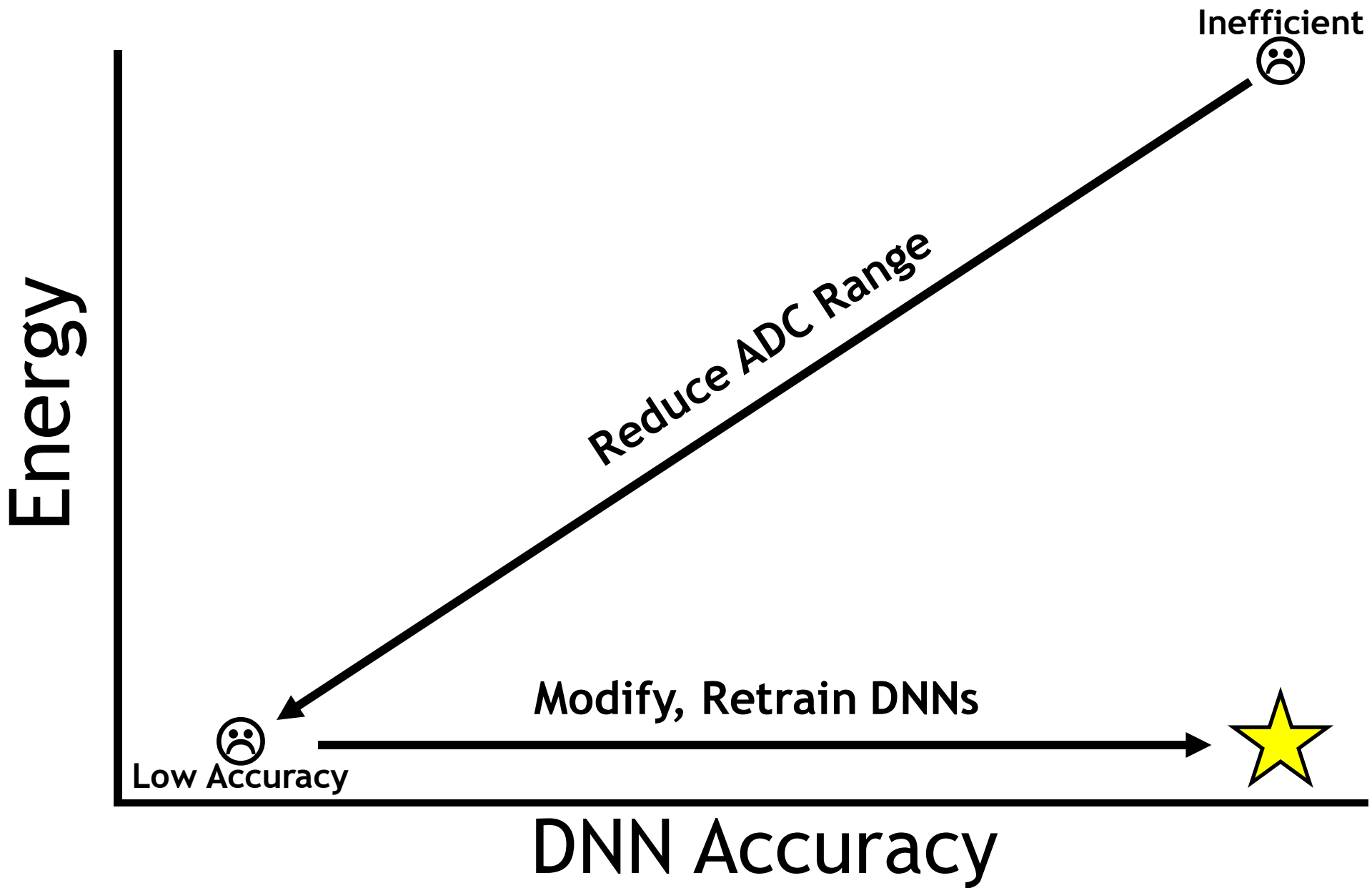


How do we reduce ADC energy?

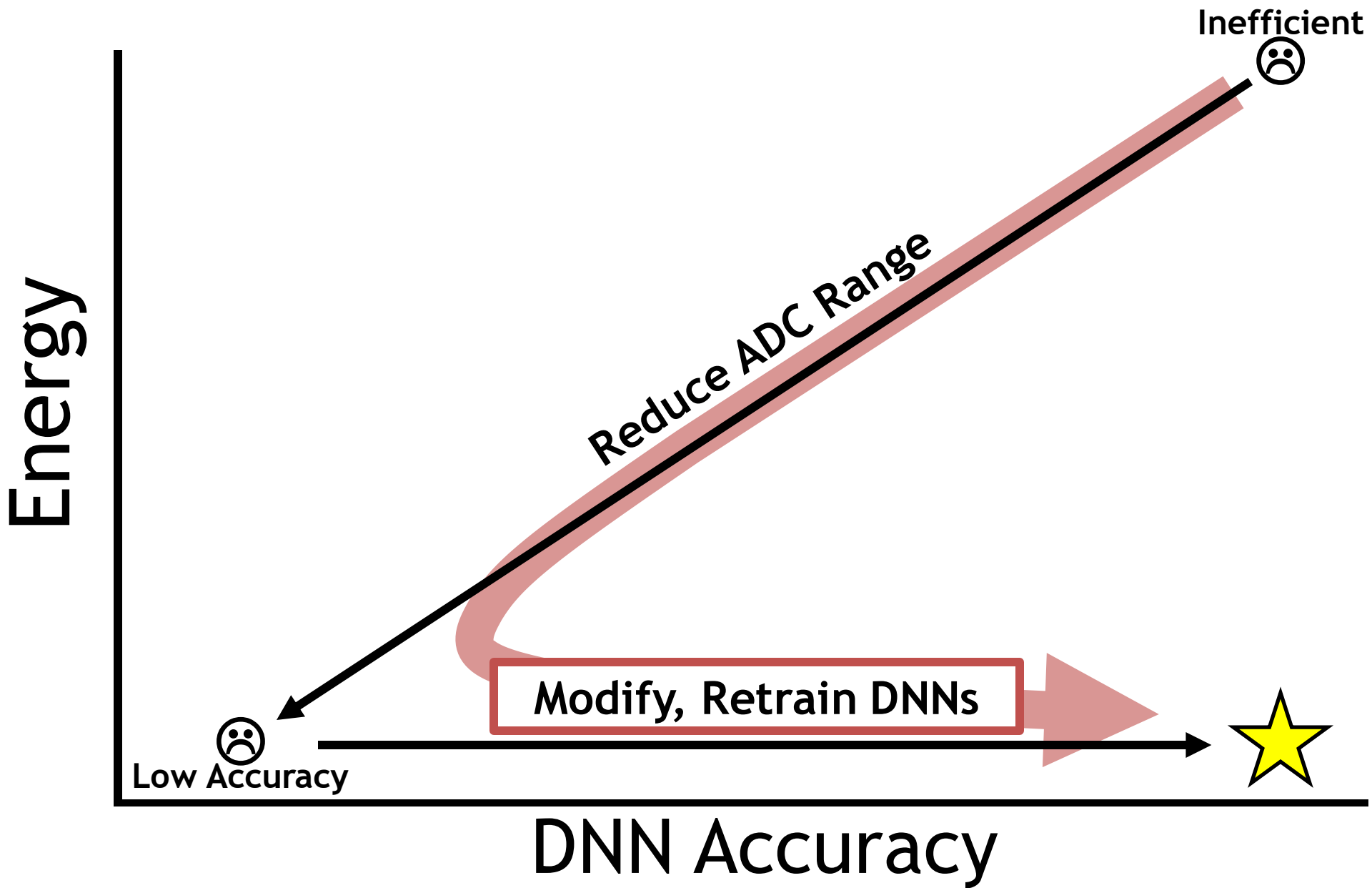
Reducing ADC Energy



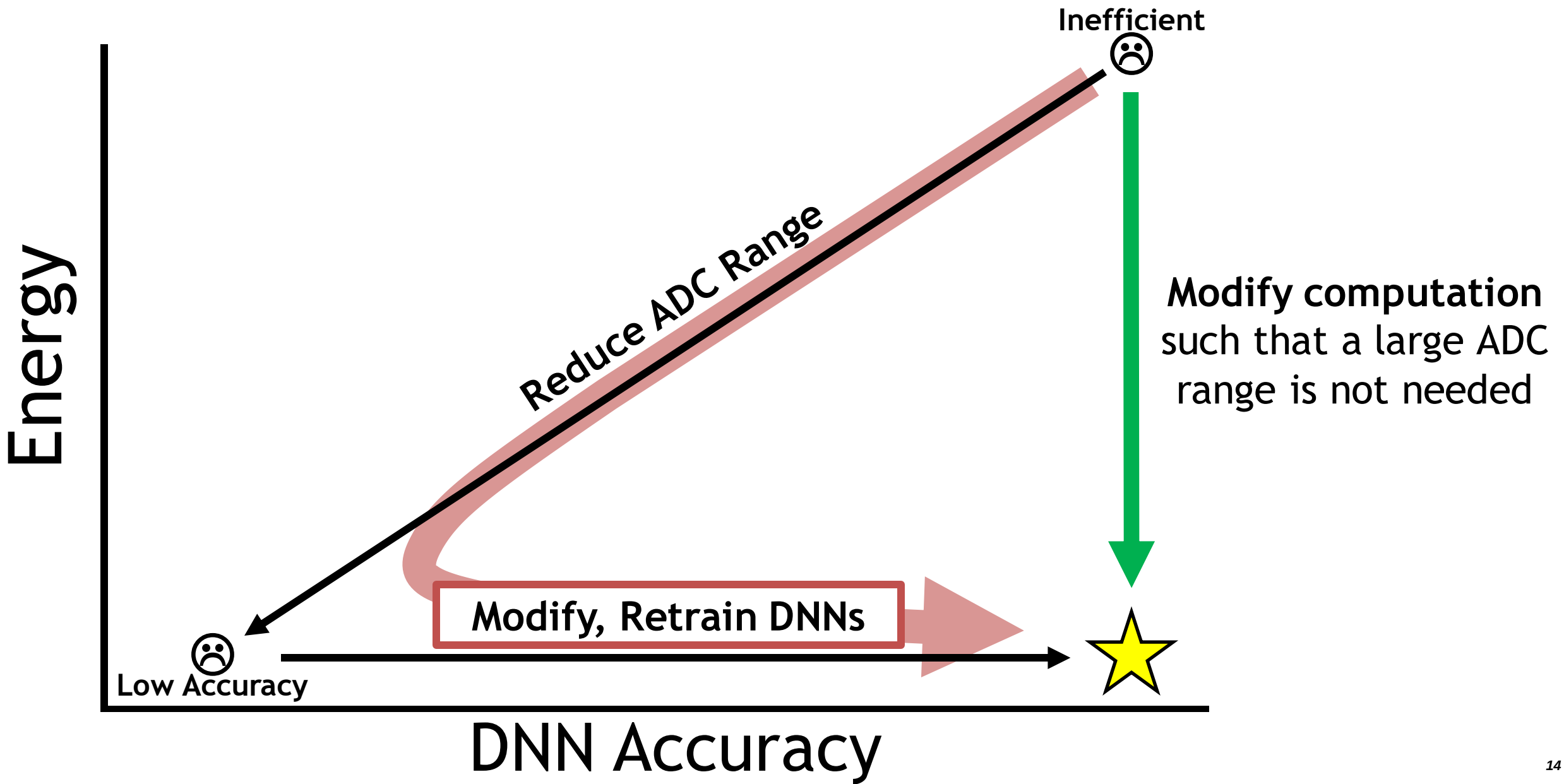
Reducing ADC Energy



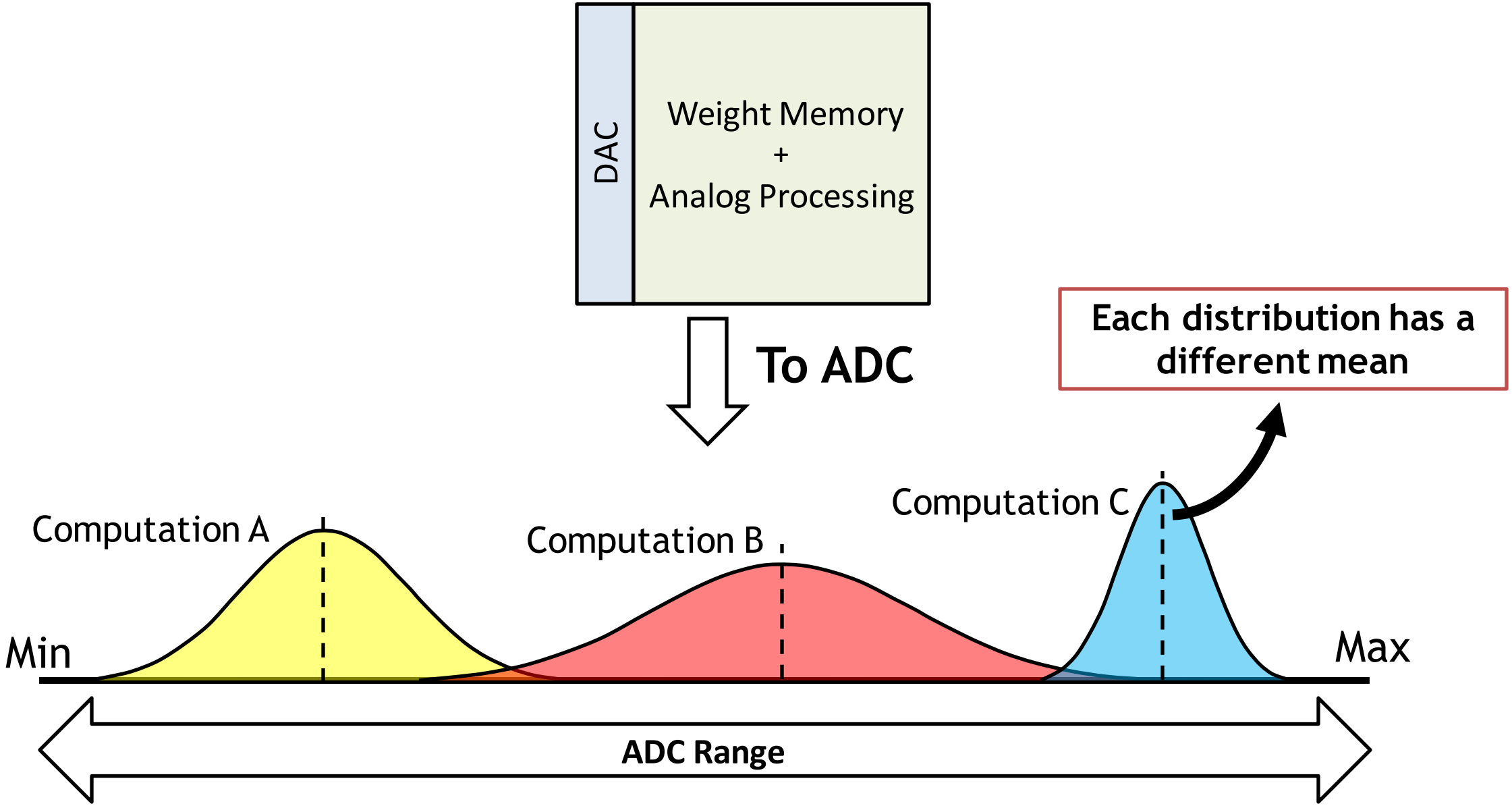
Reducing ADC Energy



Reducing ADC Energy



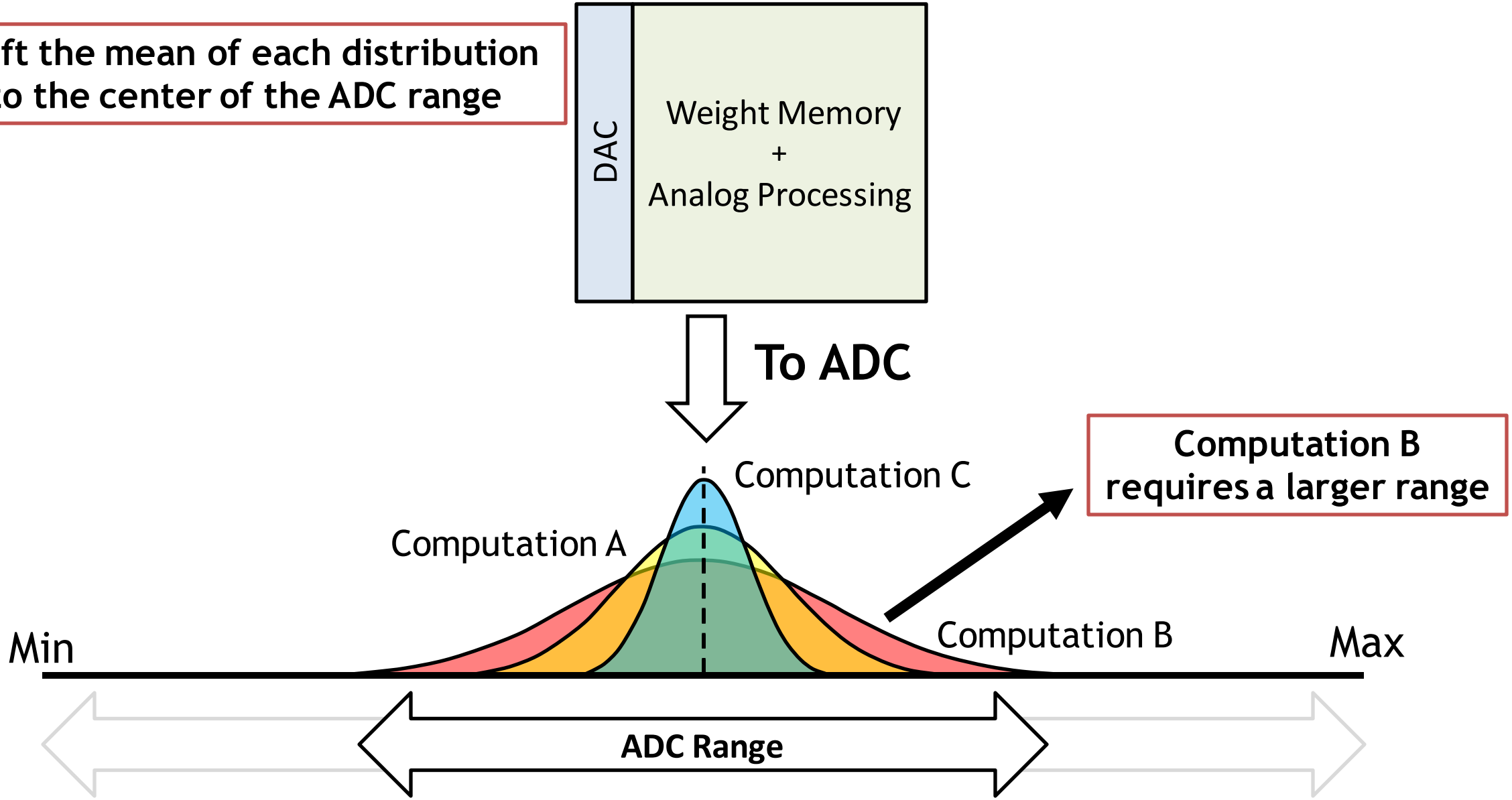
Computing in a Narrow Range



[Andrulis, Emer, Sze, ISCA 2023]

Computing in a Narrow Range

1. **Shift the mean of each distribution to the center of the ADC range**



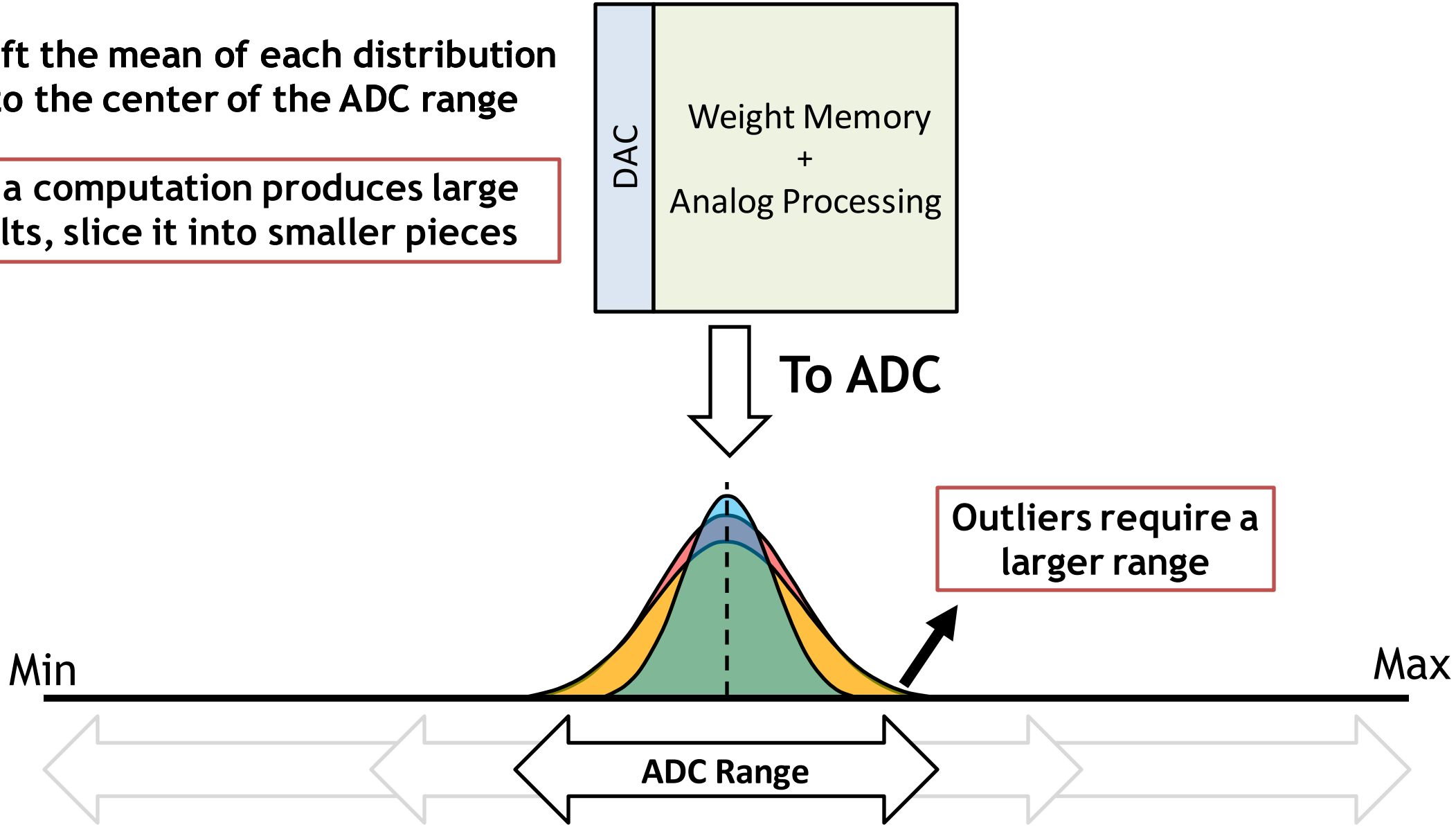
Computation B requires a larger range

[Andrulis, Emer, Sze, ISCA 2023]

Computing in a Narrow Range

1. Shift the mean of each distribution to the center of the ADC range

2. If a computation produces large results, slice it into smaller pieces



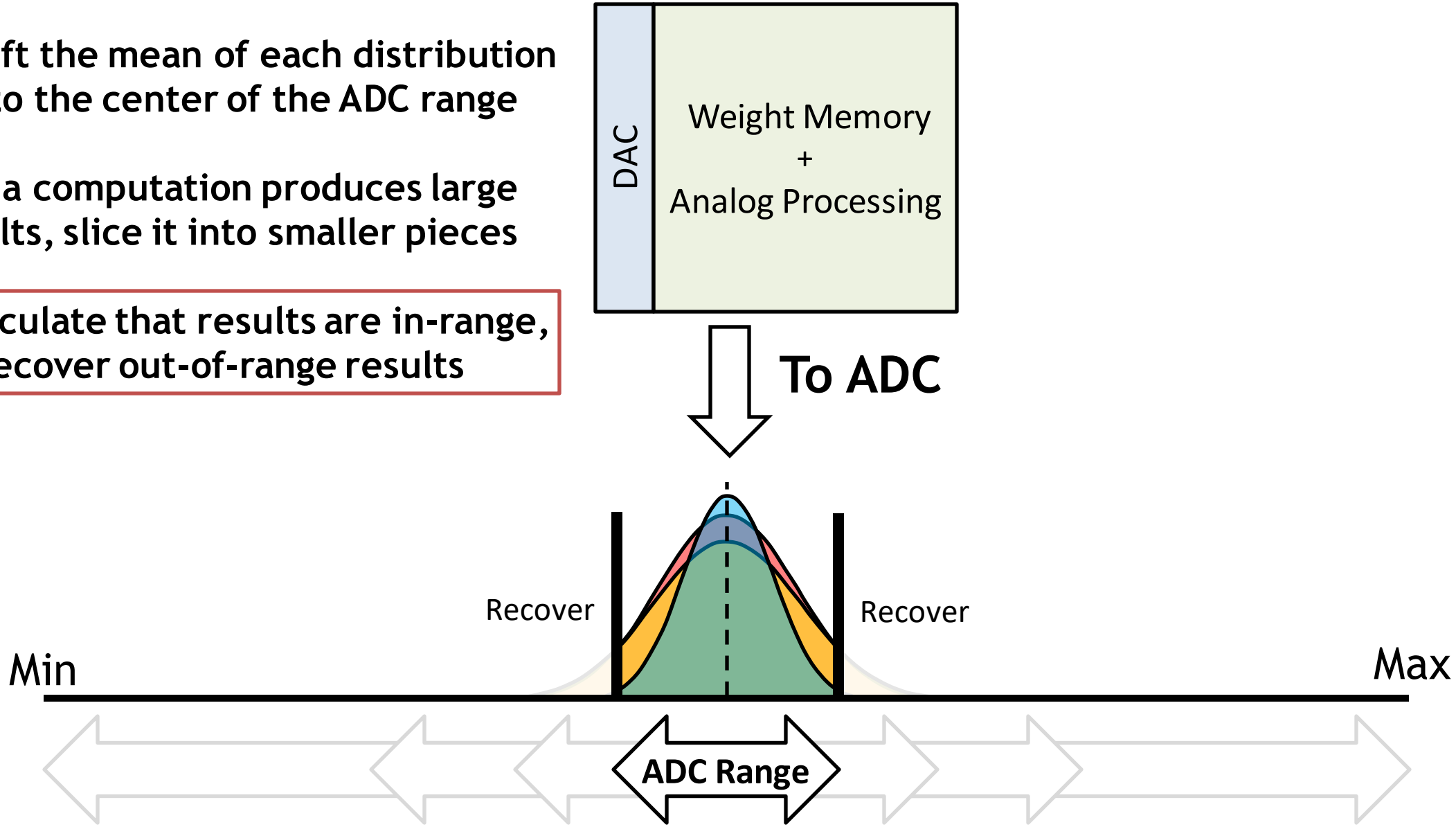
[Andrulis, Emer, Sze, ISCA 2023]

Computing in a Narrow Range

1. Shift the mean of each distribution to the center of the ADC range

2. If a computation produces large results, slice it into smaller pieces

3. Speculate that results are in-range, recover out-of-range results



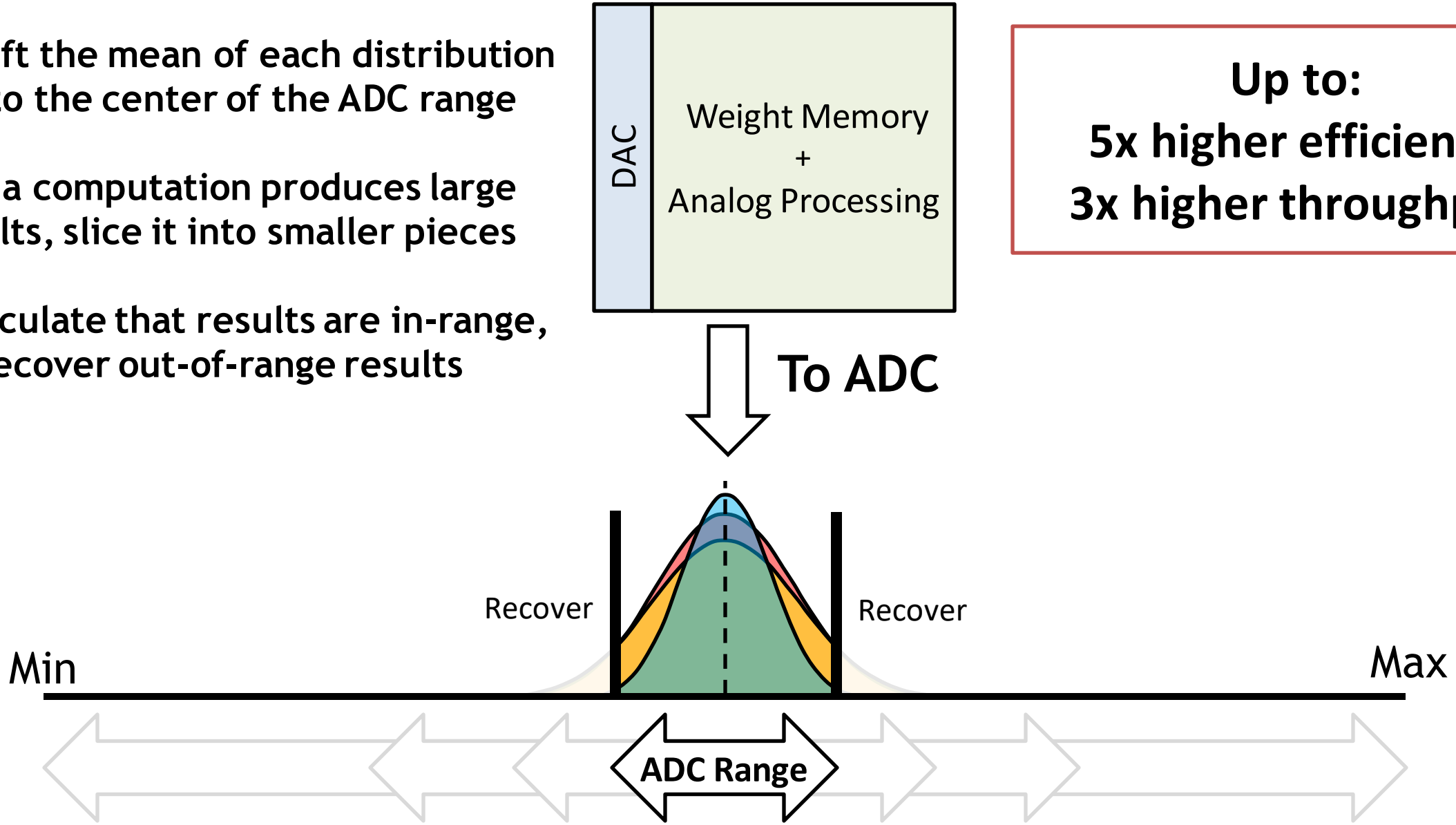
[Andrulis, Emer, Sze, ISCA 2023]

Computing in a Narrow Range

1. Shift the mean of each distribution to the center of the ADC range

2. If a computation produces large results, slice it into smaller pieces

3. Speculate that results are in-range, recover out-of-range results



Up to:
5x higher efficiency
3x higher throughput

[Andrulis, Emer, Sze, ISCA 2023]

Key Takeaways

Analog Processing-In-Memory can efficiently run Deep Neural Networks

But to use it effectively, we must think about how we compute

What computations does the neural network do?

How do we formulate computations for analog hardware?

Good answers can lead to lower-energy hardware.