RAELLA

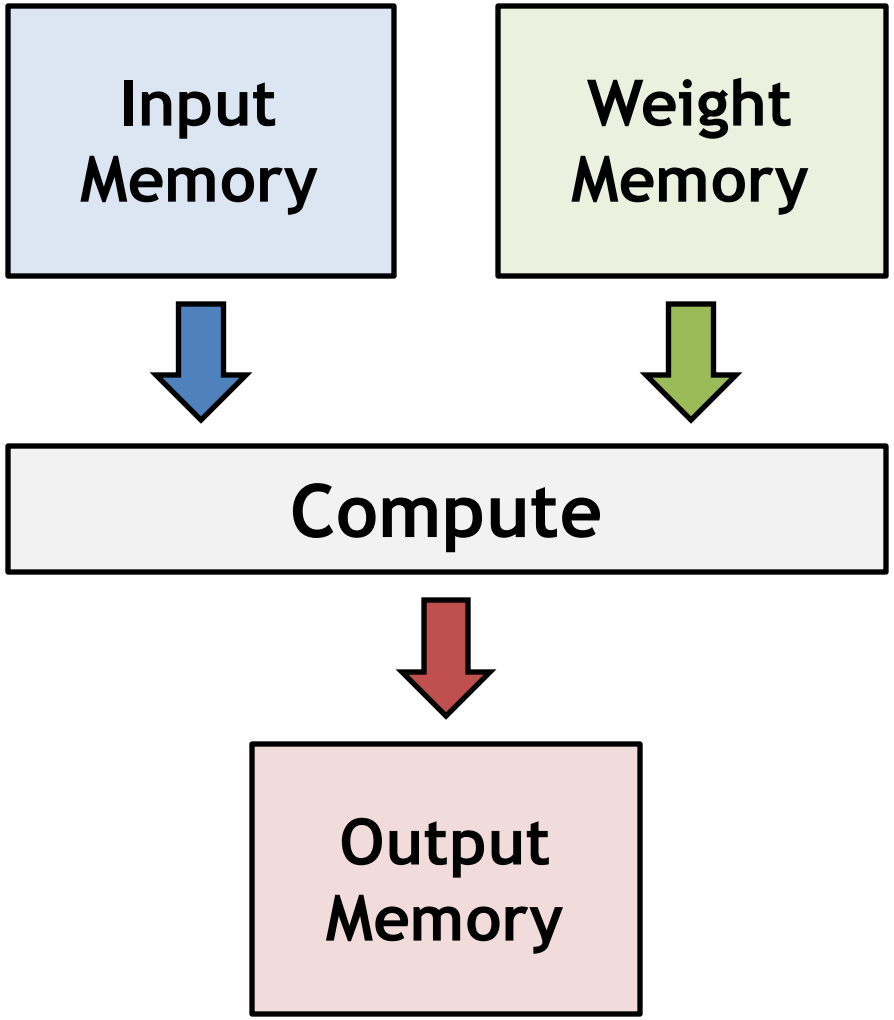
Reforming the Arithmetic for Efficient, Low-Resolution, and Low-Loss Analog PIM: No Retraining Required!

Tanner Andrulis, Joel S. Emer, Vivienne Sze

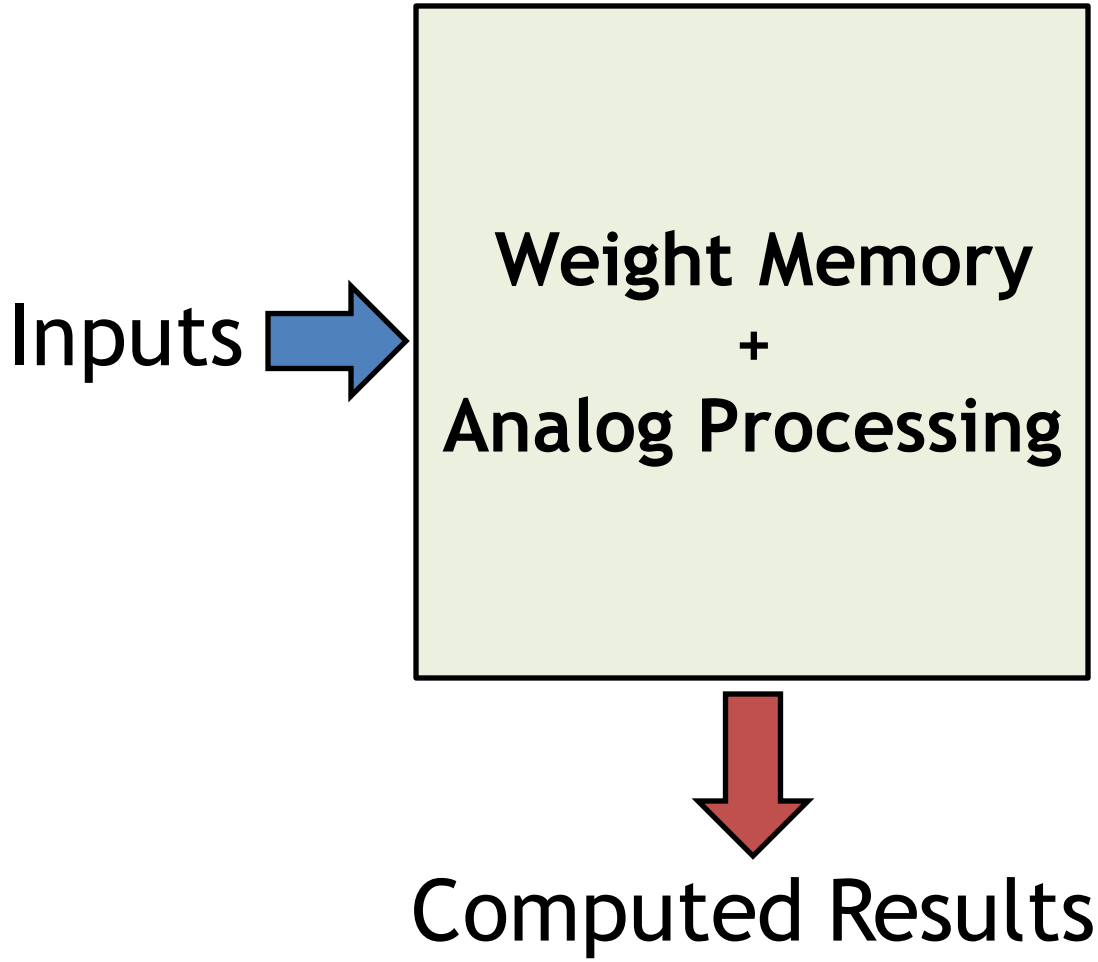
International Symposium on Computer Architecture (ISCA) 2023

Processing In Memory (PIM) Accelerators

Conventional



Processing In Memory



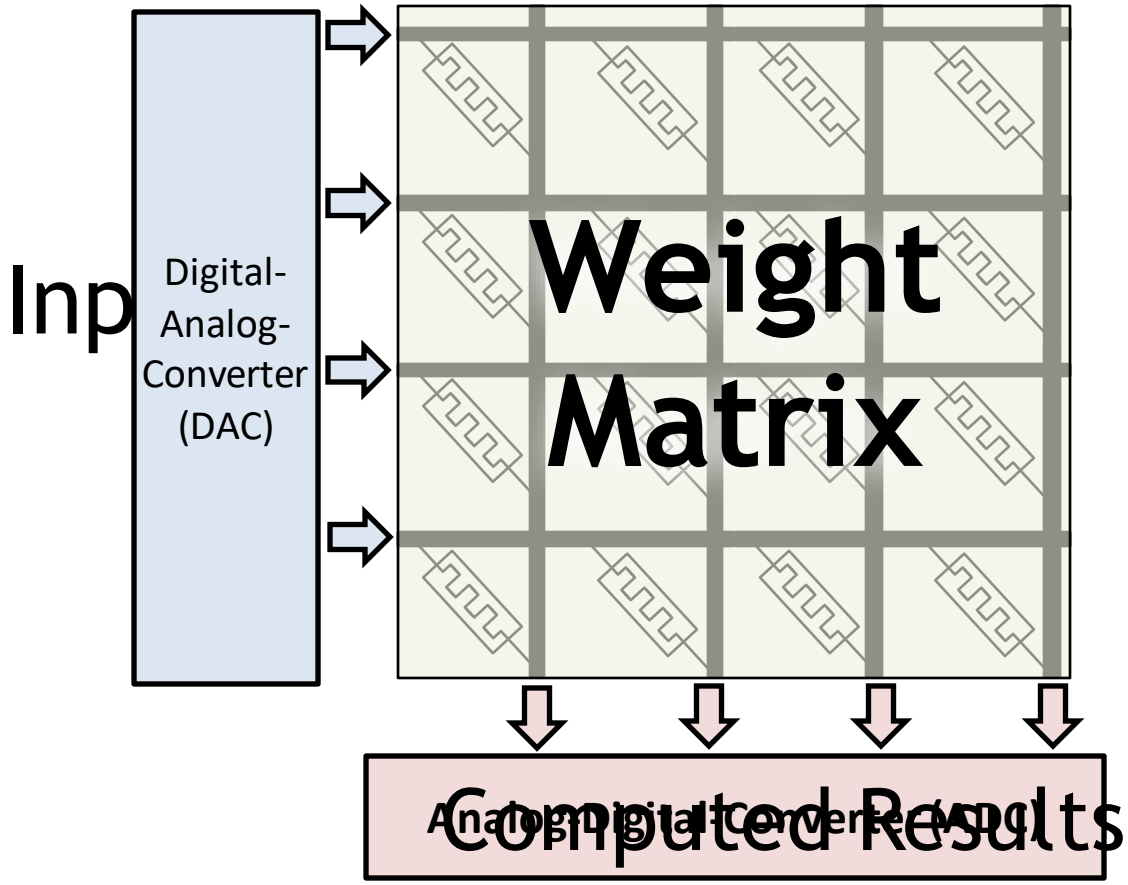
Processing In Memory (PIM) Accelerators

Weight matrix stored in crossbar
as analog conductance values

Input vector applied to rows
as analog temporal values

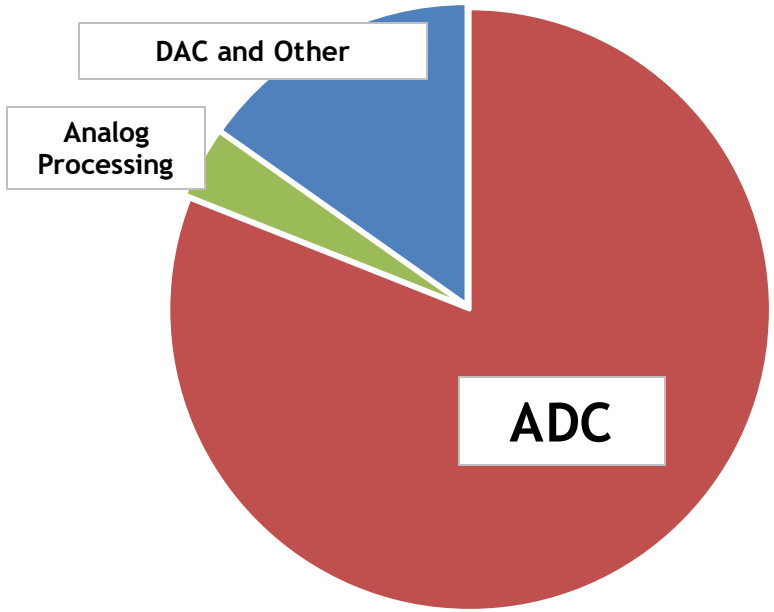
Analog matrix-vector multiply
Charge ~ Conductance × Time

Results appear on columns
As analog charge values

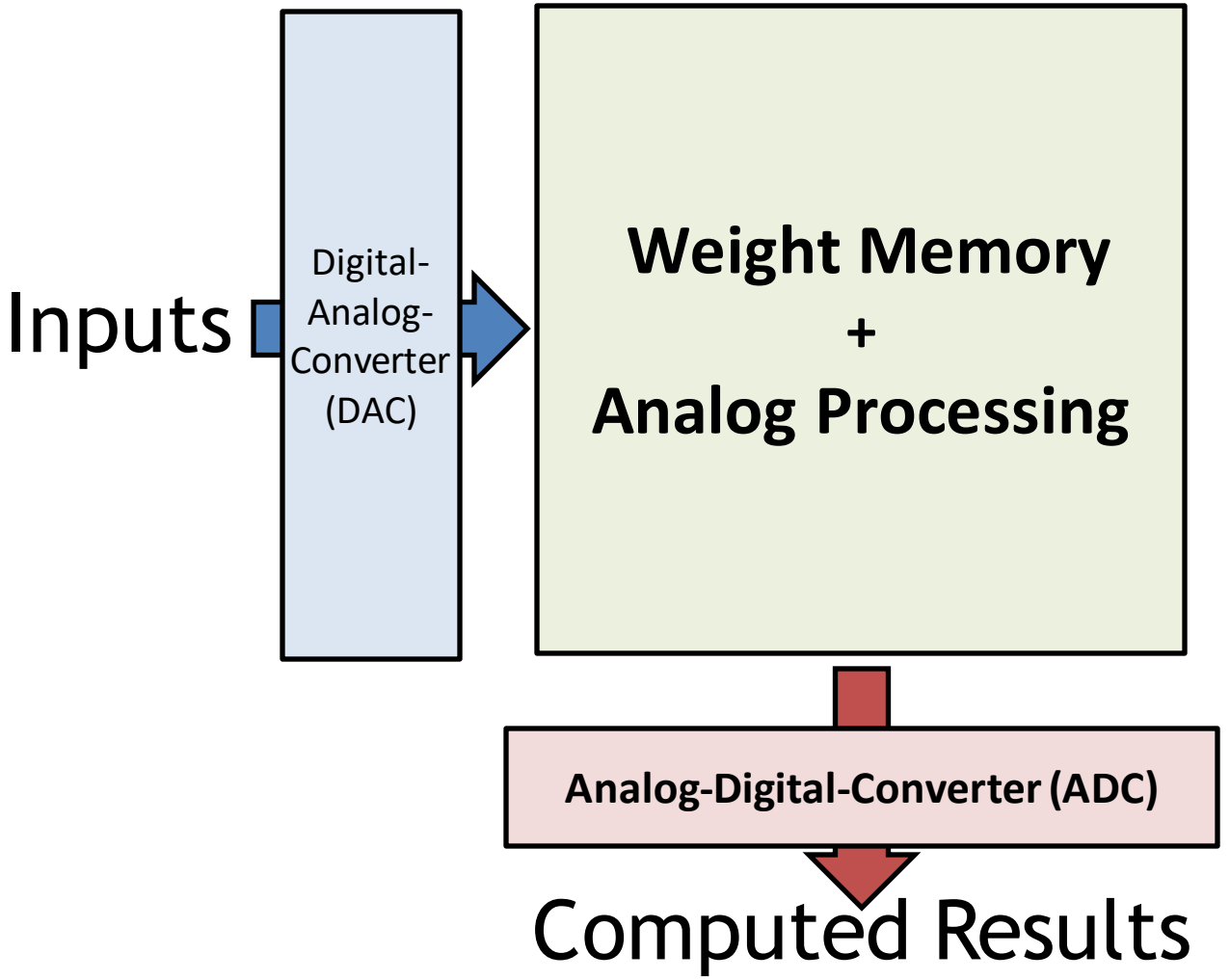


Processing In Memory (PIM) Accelerators

Energy Breakdown



ADCs can consume a significant amount of energy



The Titanium Law of ADC Energy

Titanium Law Helps Us Understand ADC Energy Tradeoffs

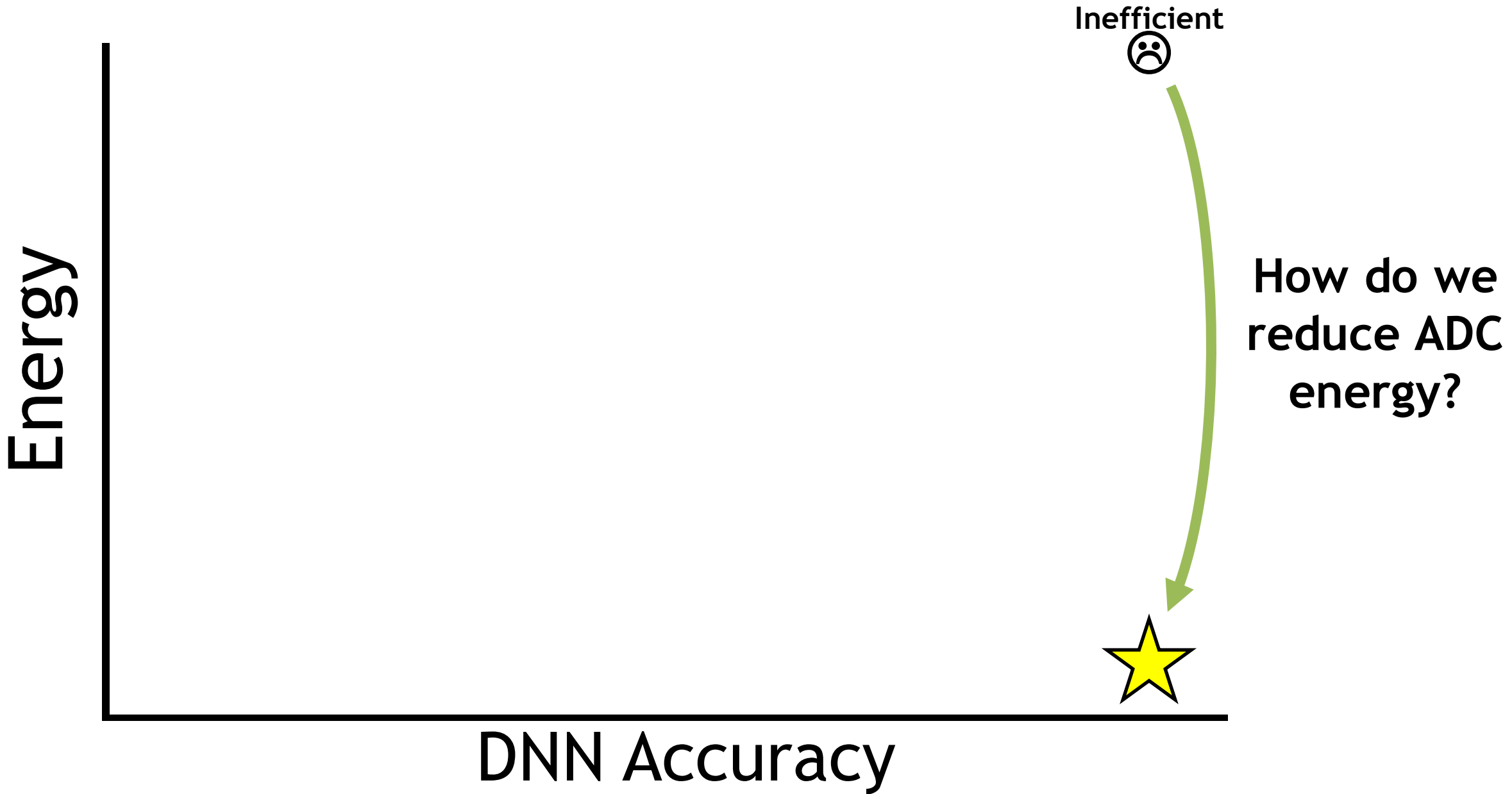
$$\frac{\boldsymbol{ADC\ Energy}}{\boldsymbol{DNN}} =$$

Idea: Break computation into smaller pieces

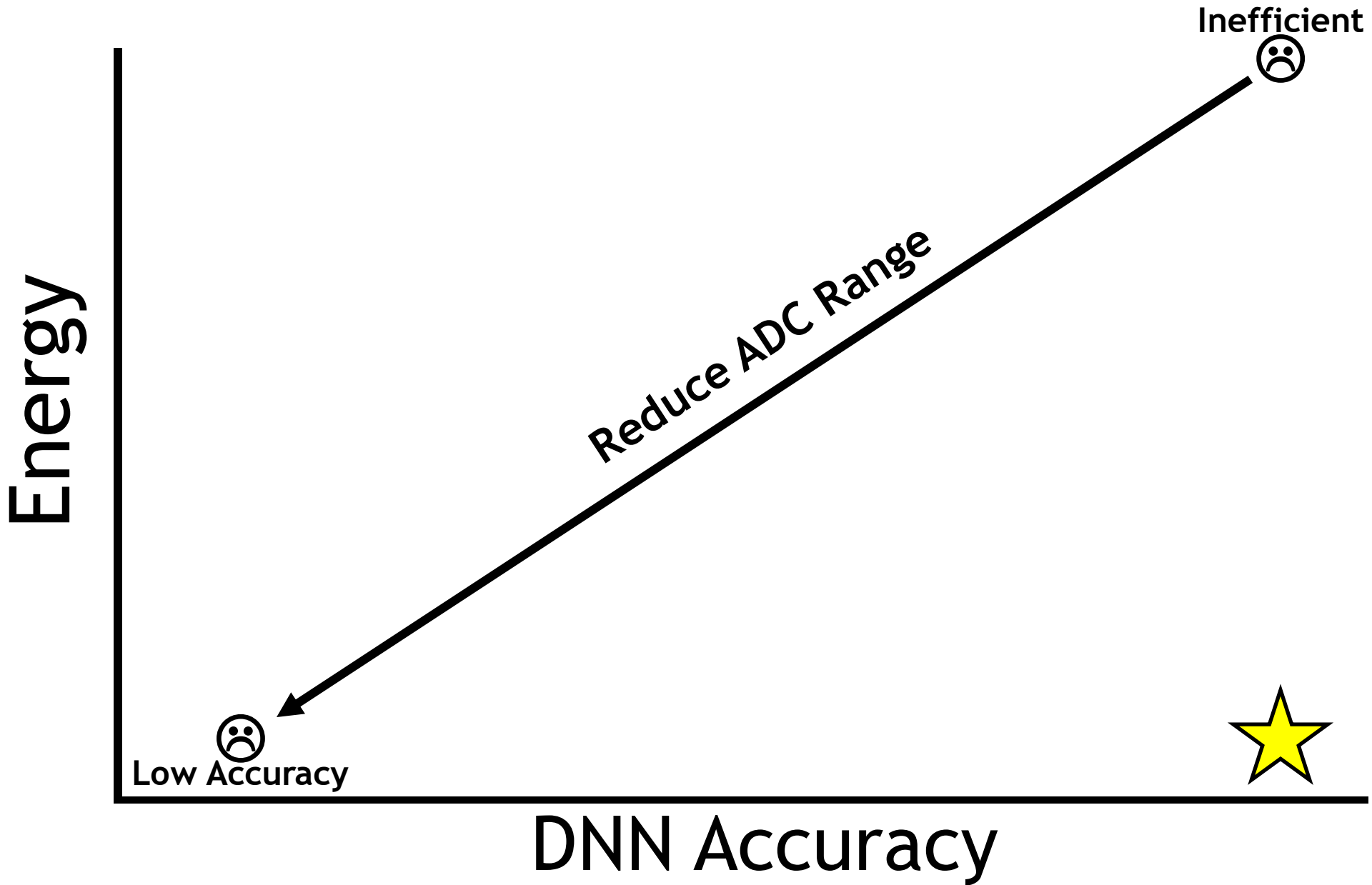
Benefit: Smaller result from each piece, ↓ **Energy/Convert**

Tradeoff: More pieces to process, ↑ **Converts/MAC**

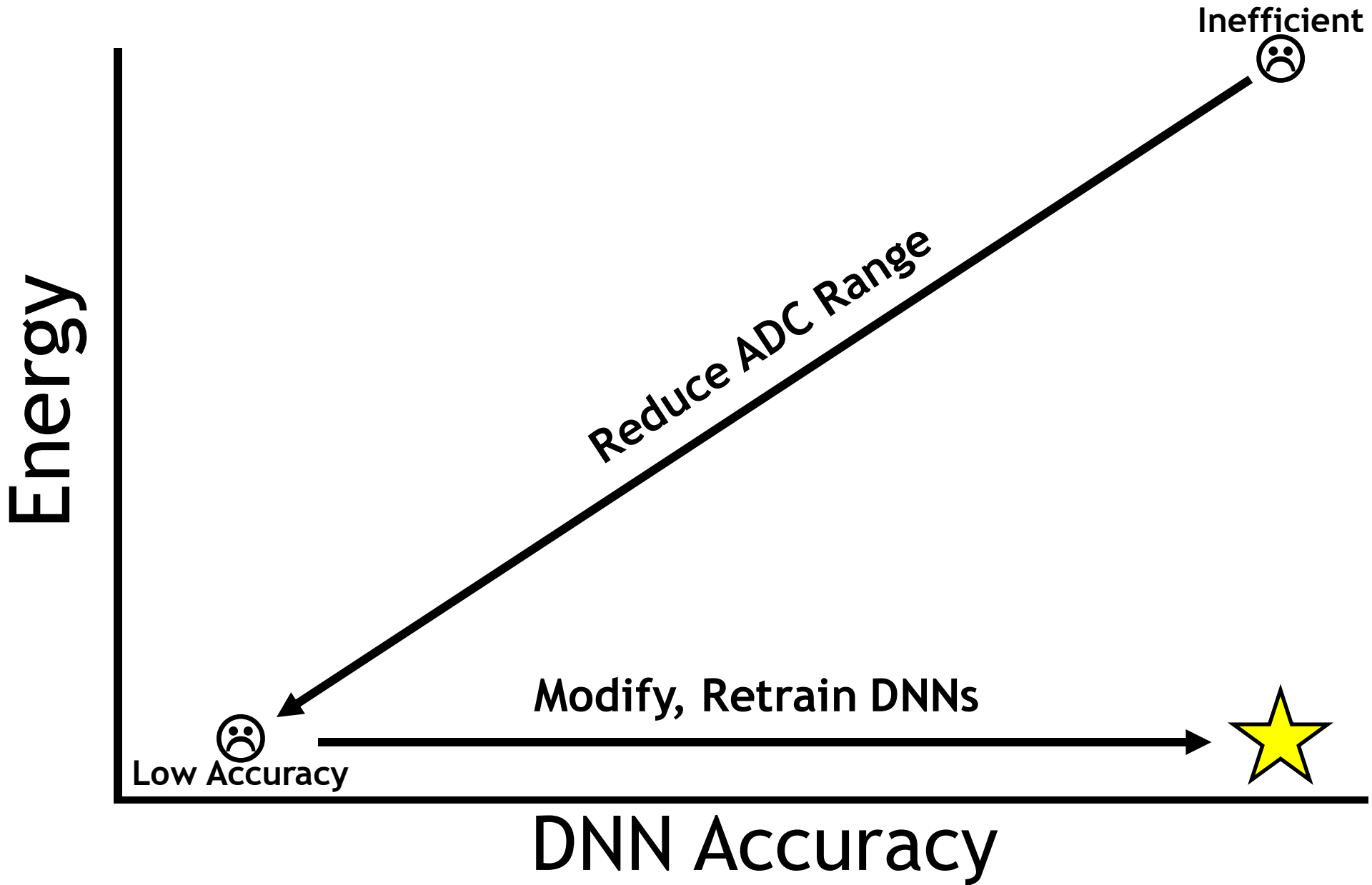
Reducing ADC Energy



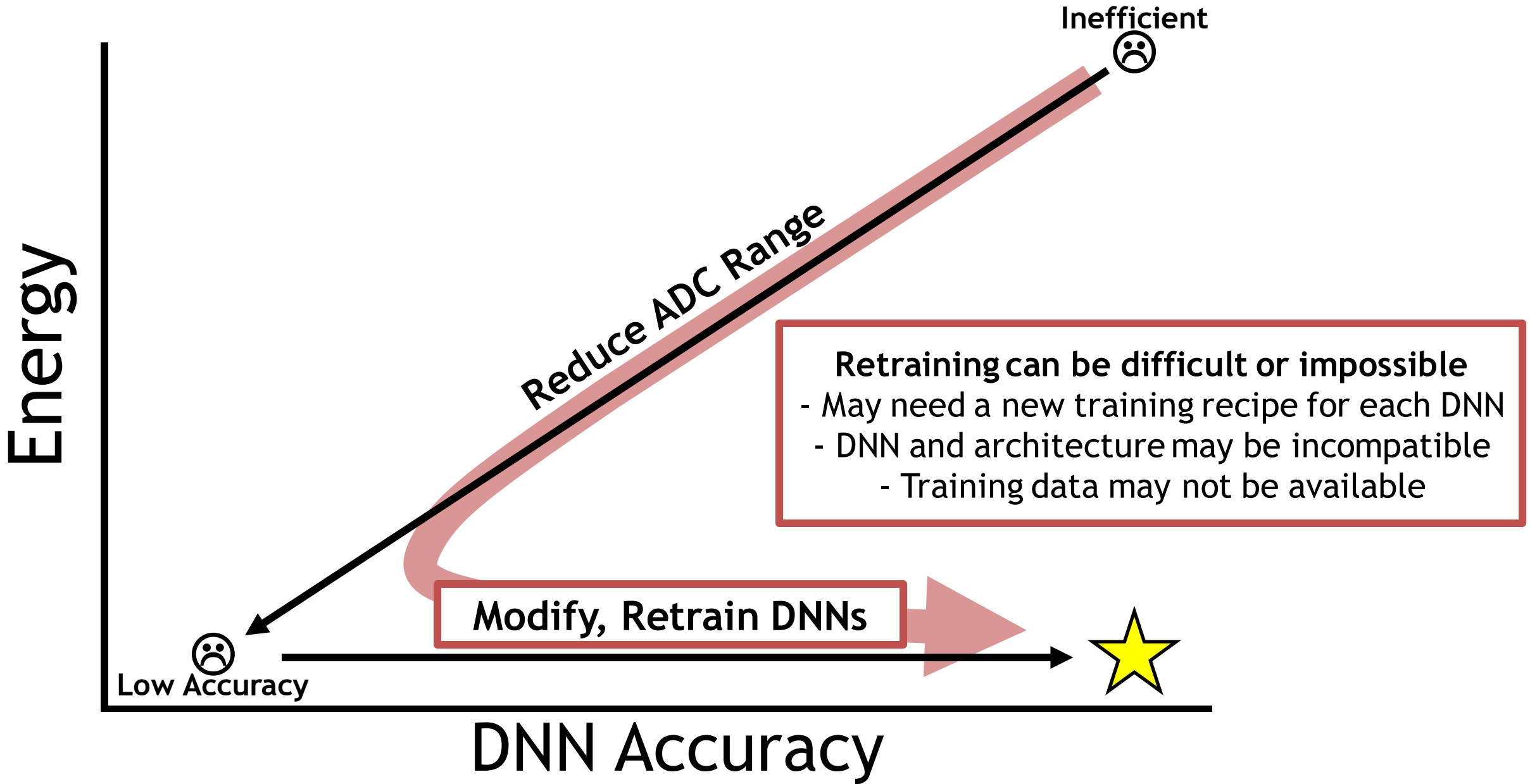
Reducing ADC Energy



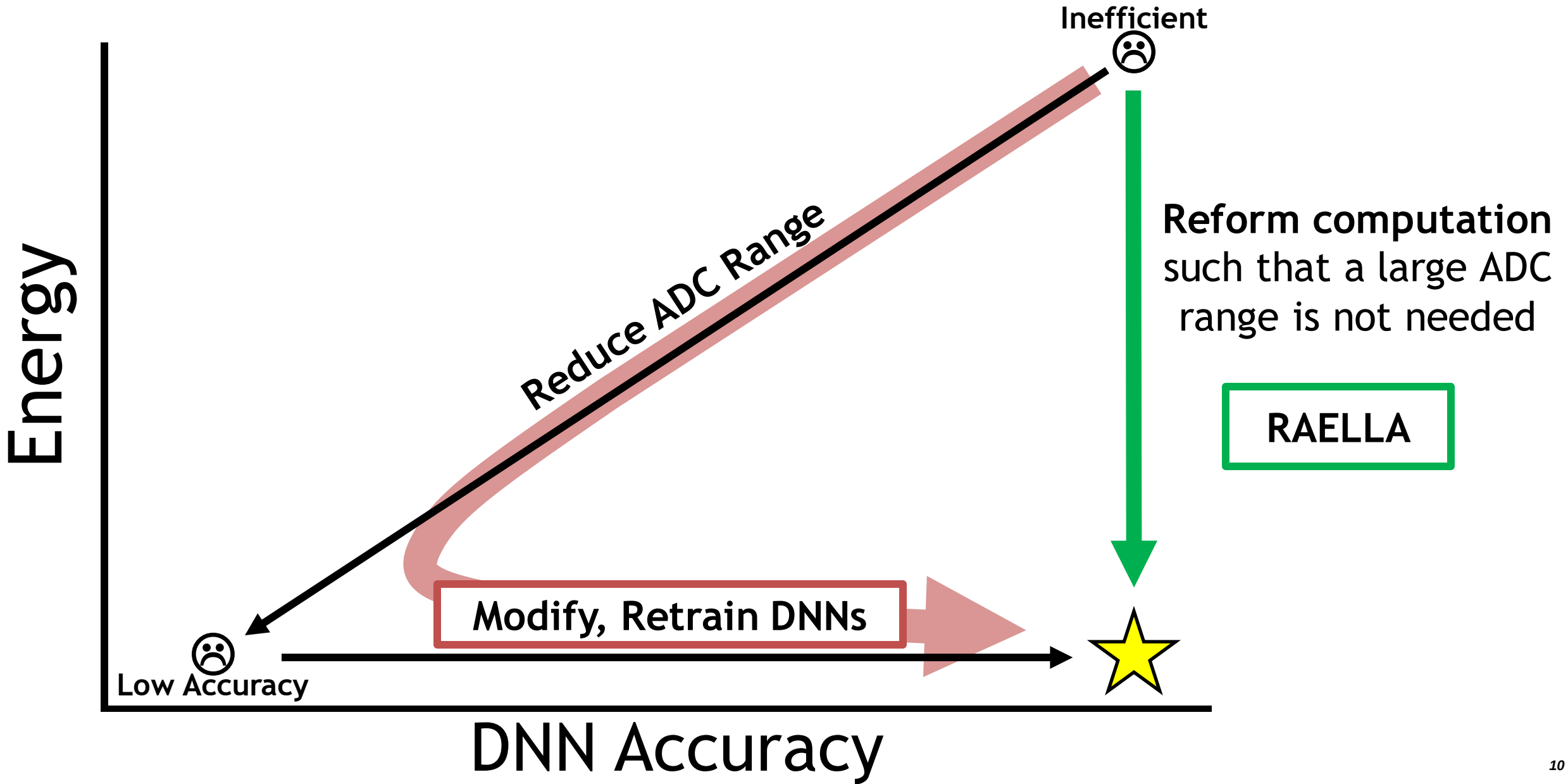
Reducing ADC Energy



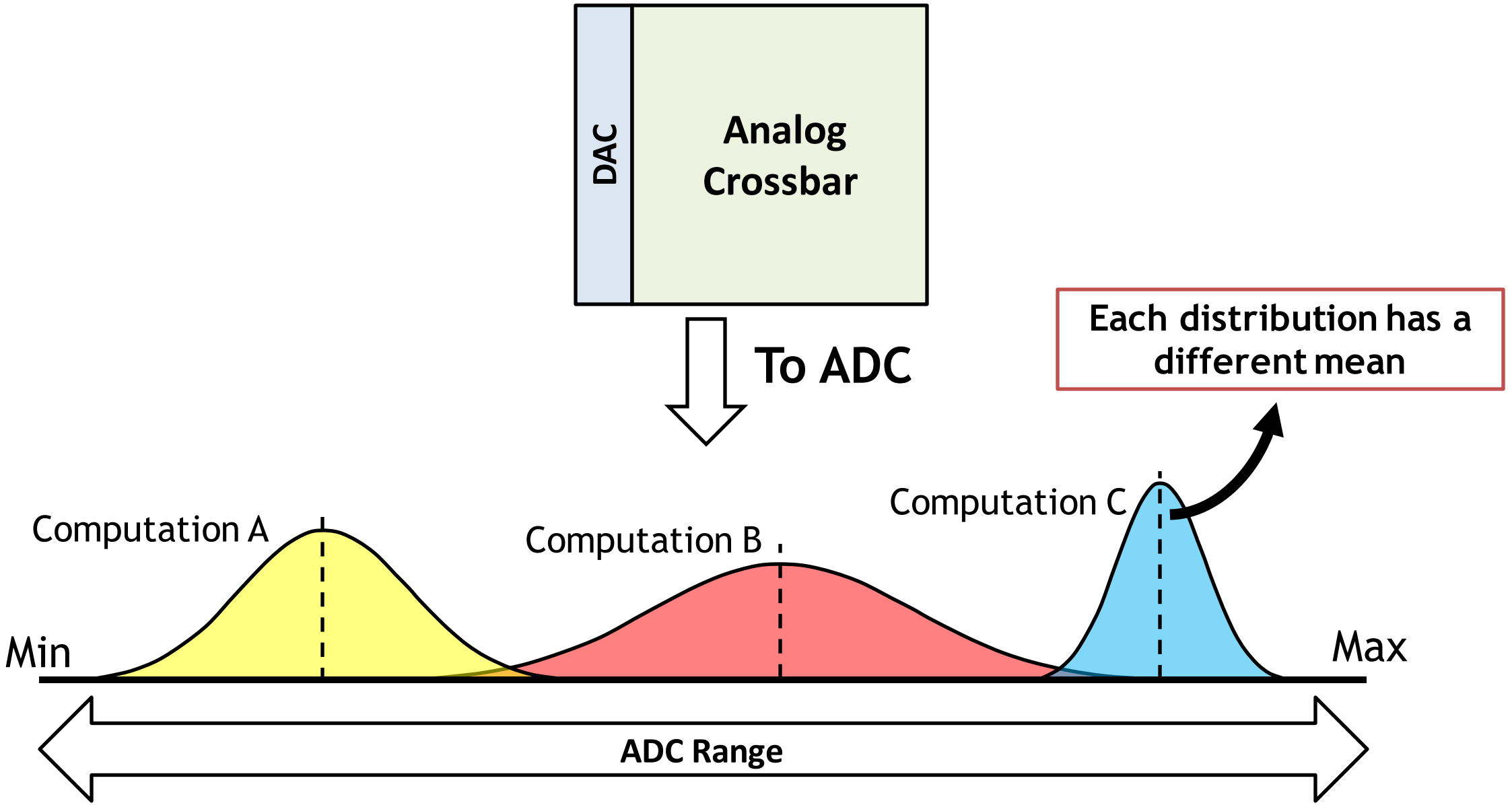
Reducing ADC Energy



Reducing ADC Energy

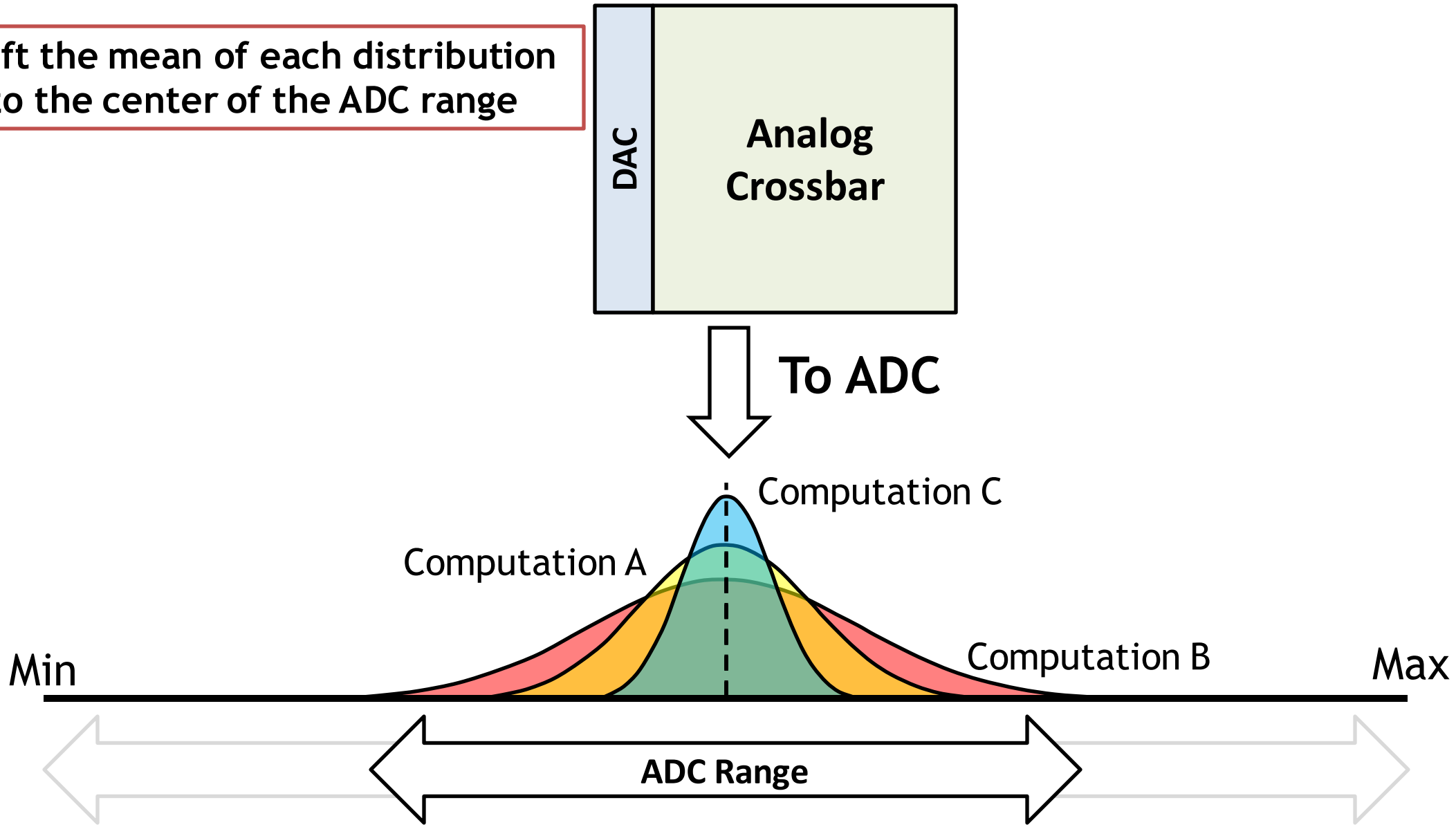


Shifting Distributions

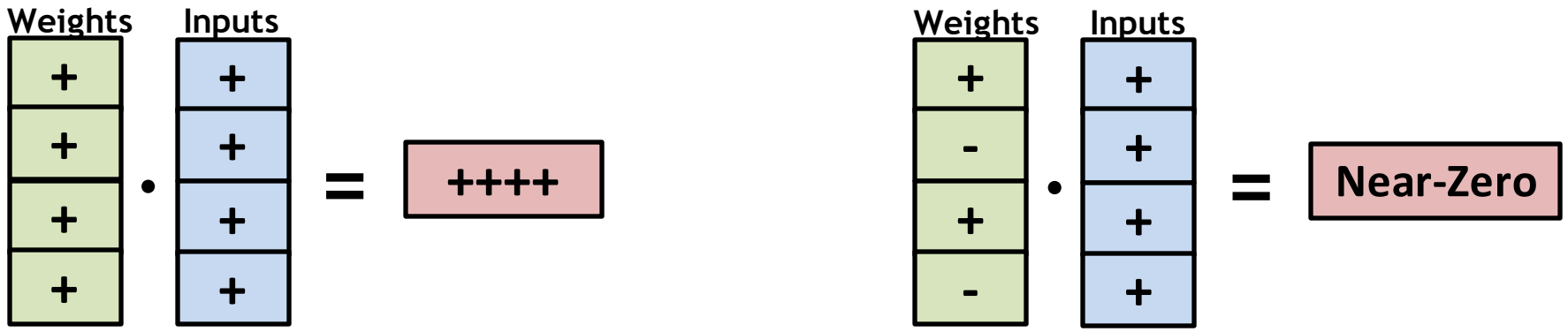


Shifting Distributions

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

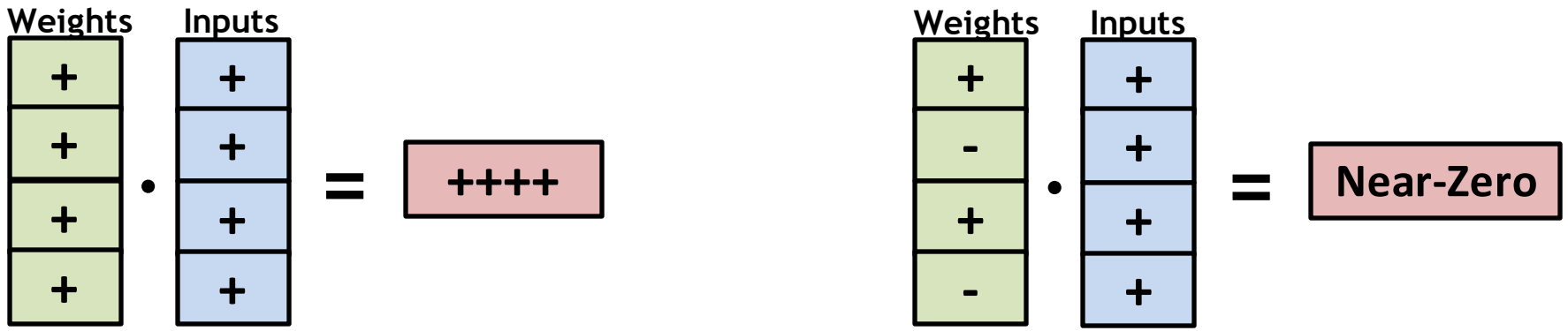


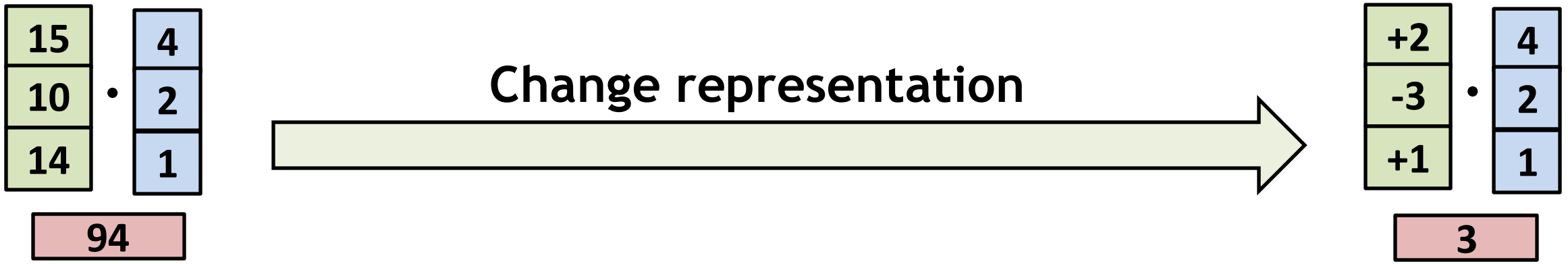
Center+Offset: Zero-Average Analog Results



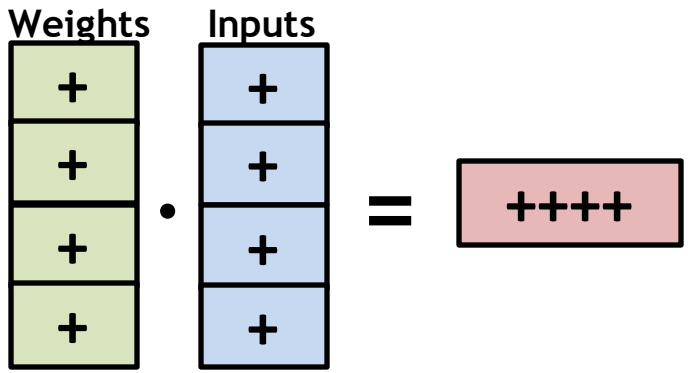
Computations with zero-average weights produce near-zero results

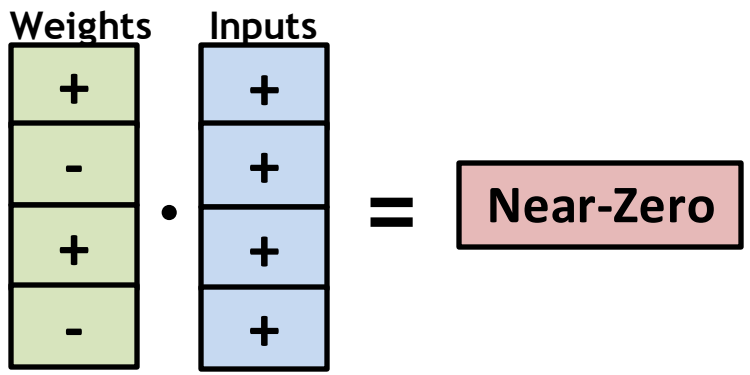
Center+Offset: Zero-Average Analog Results



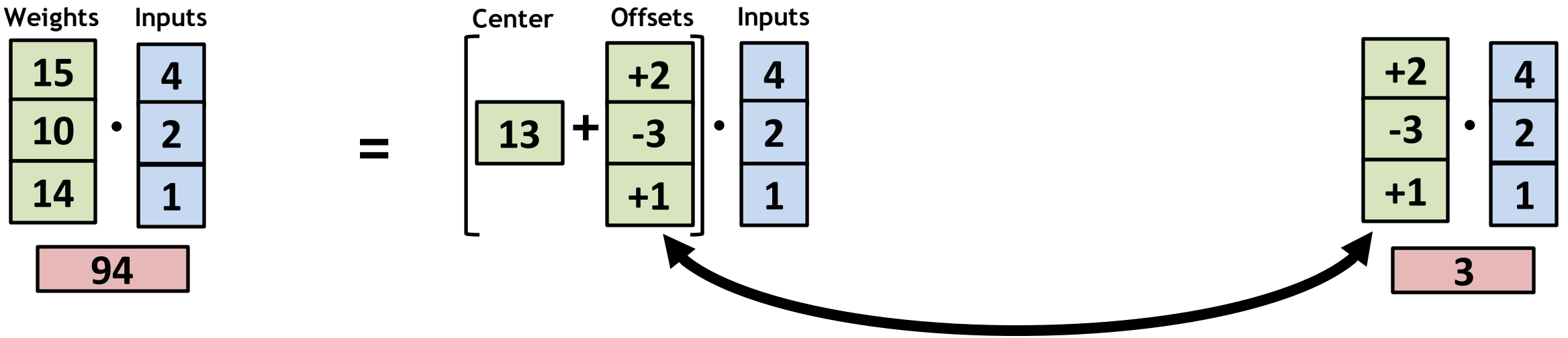


Center+Offset: Zero-Average Analog Results





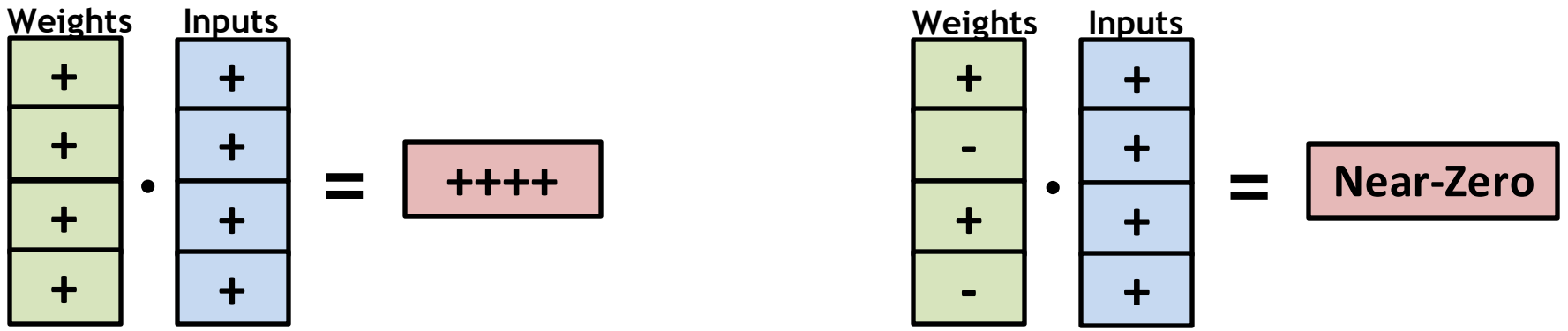
Pick a center to balance +/- offsets



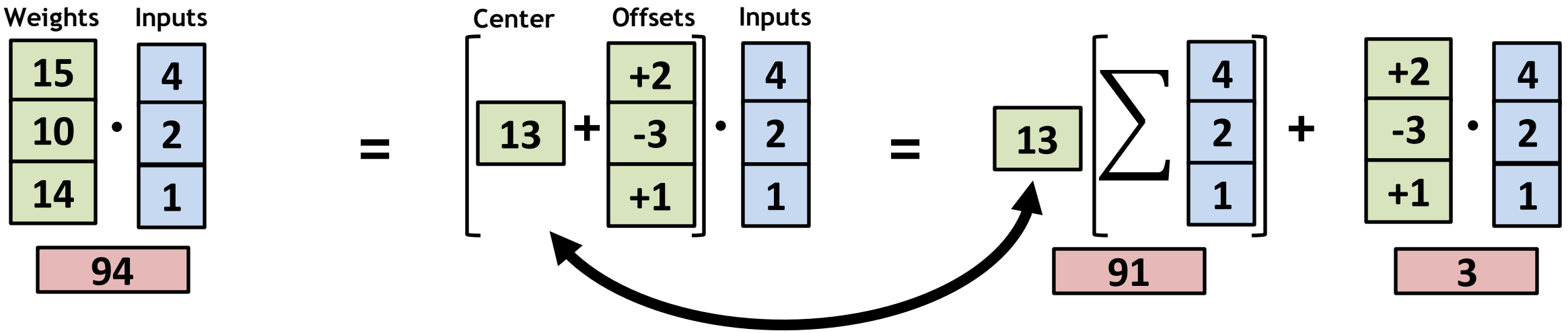
Extract a center value

Weights are center +/- zero-average offsets

Center+Offset: Zero-Average Analog Results

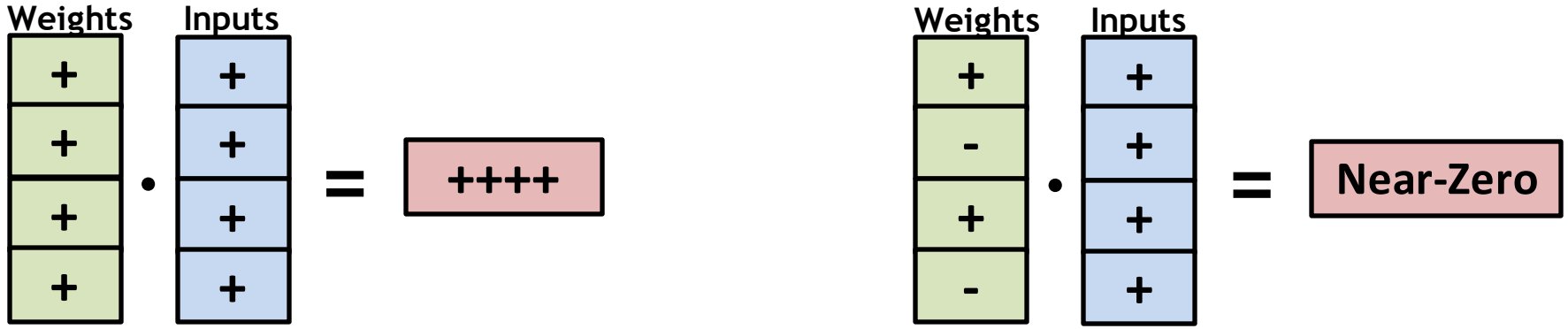


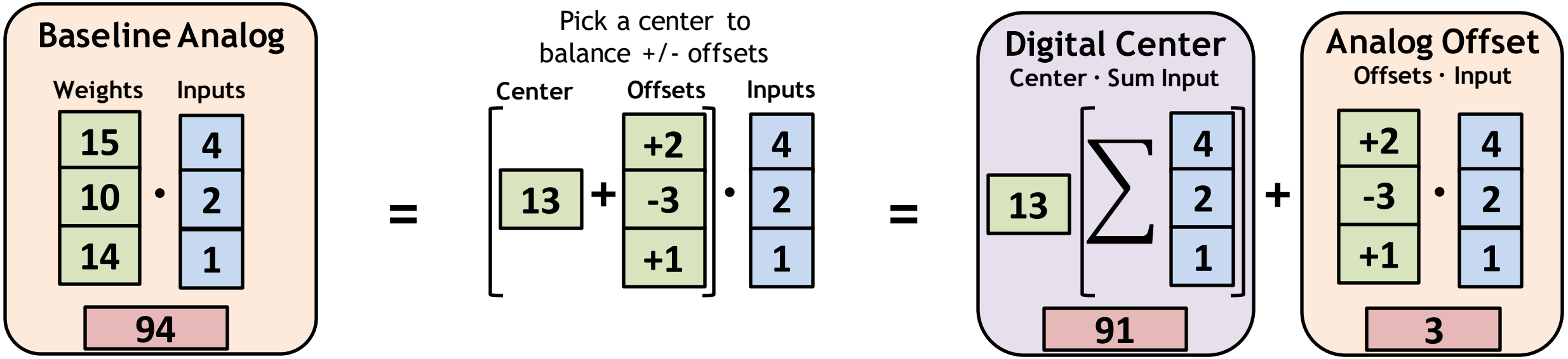
Pick a center to balance +/- offsets



Move center arithmetic to a different term

Center+Offset: Zero-Average Analog Results





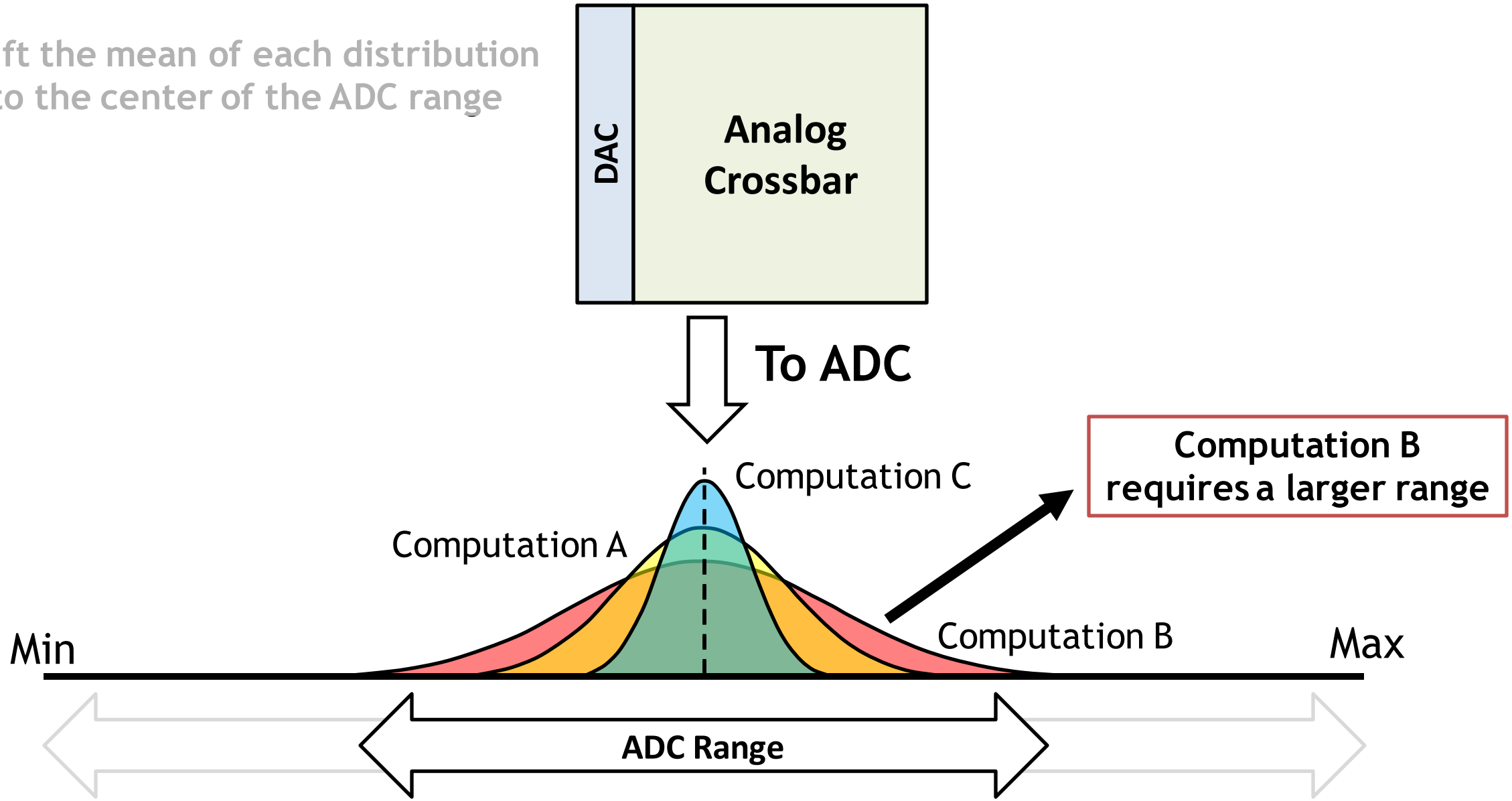
Key Takeaway: Partition computation

Digital calculates high-resolution center operations

Analog calculates parallel offset operations

Shifting Distributions

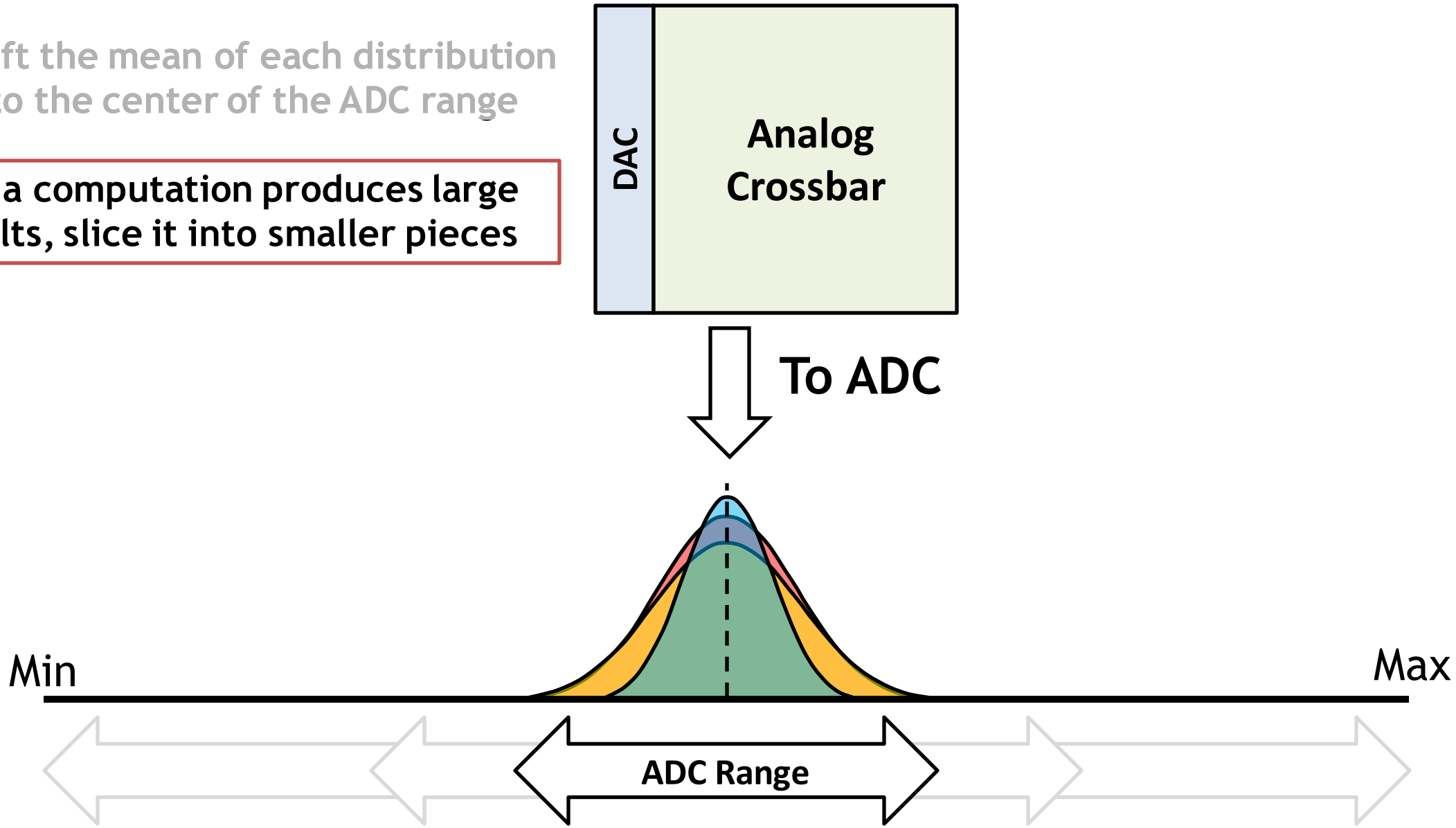
1. Shift the mean of each distribution to the center of the ADC range



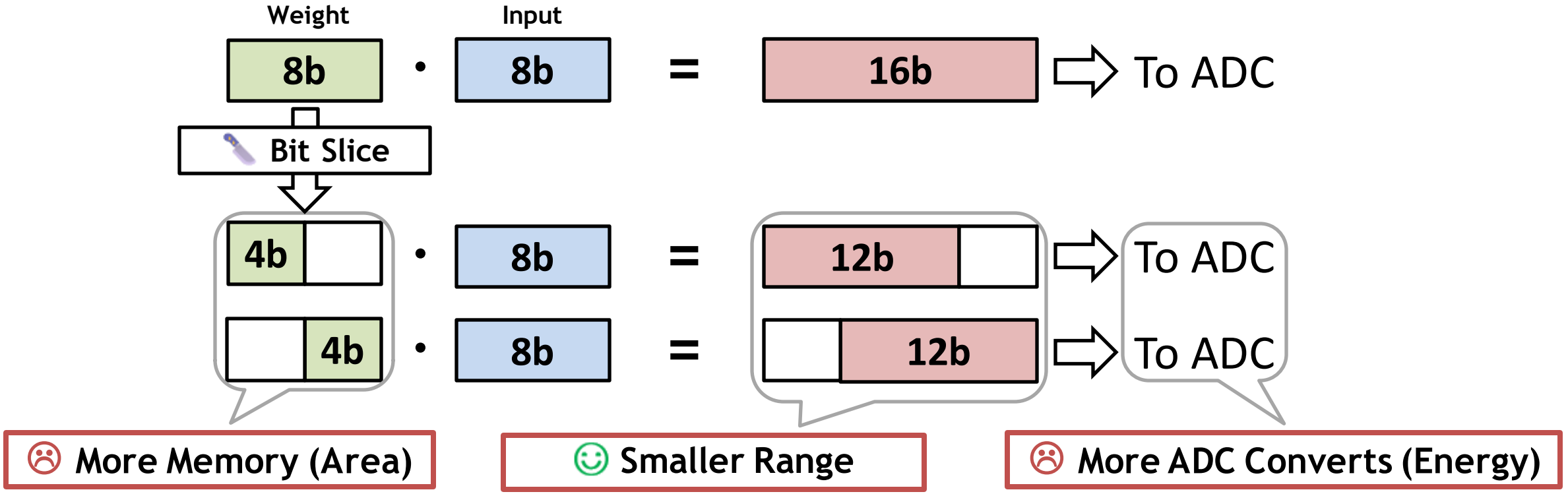
Shifting Distributions

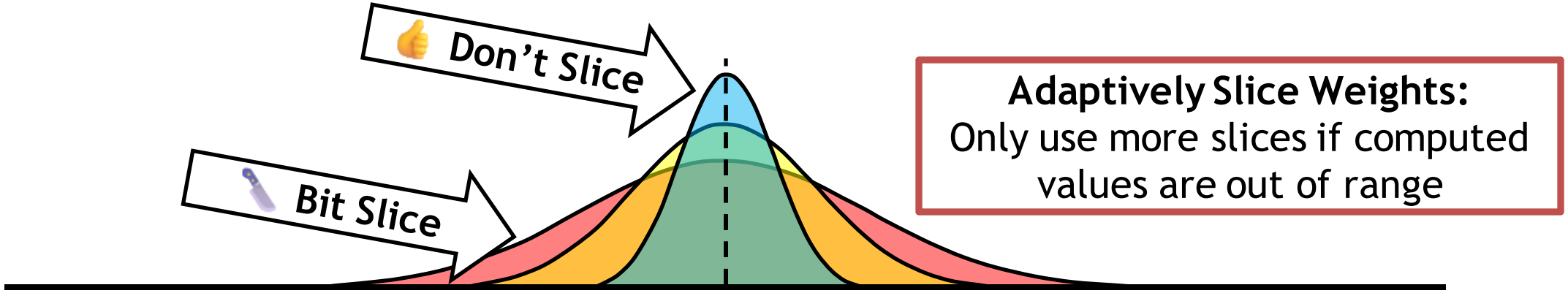
1. Shift the mean of each distribution to the center of the ADC range

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



Adaptive Weight Slicing: Slice Large-Result Computations



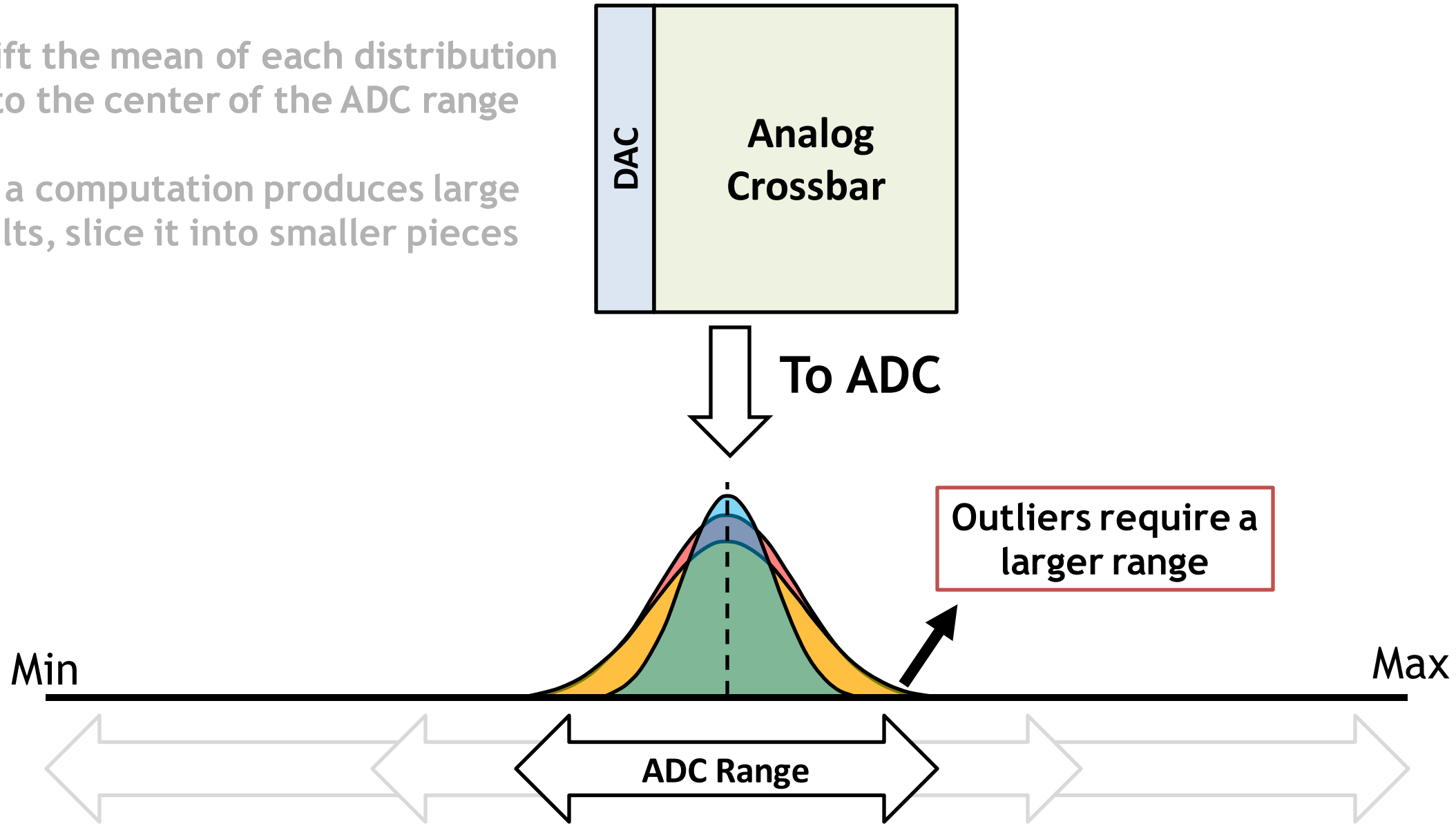


Adaptively Slice Weights:
Only use more slices if computed values are out of range

Shifting Distributions

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

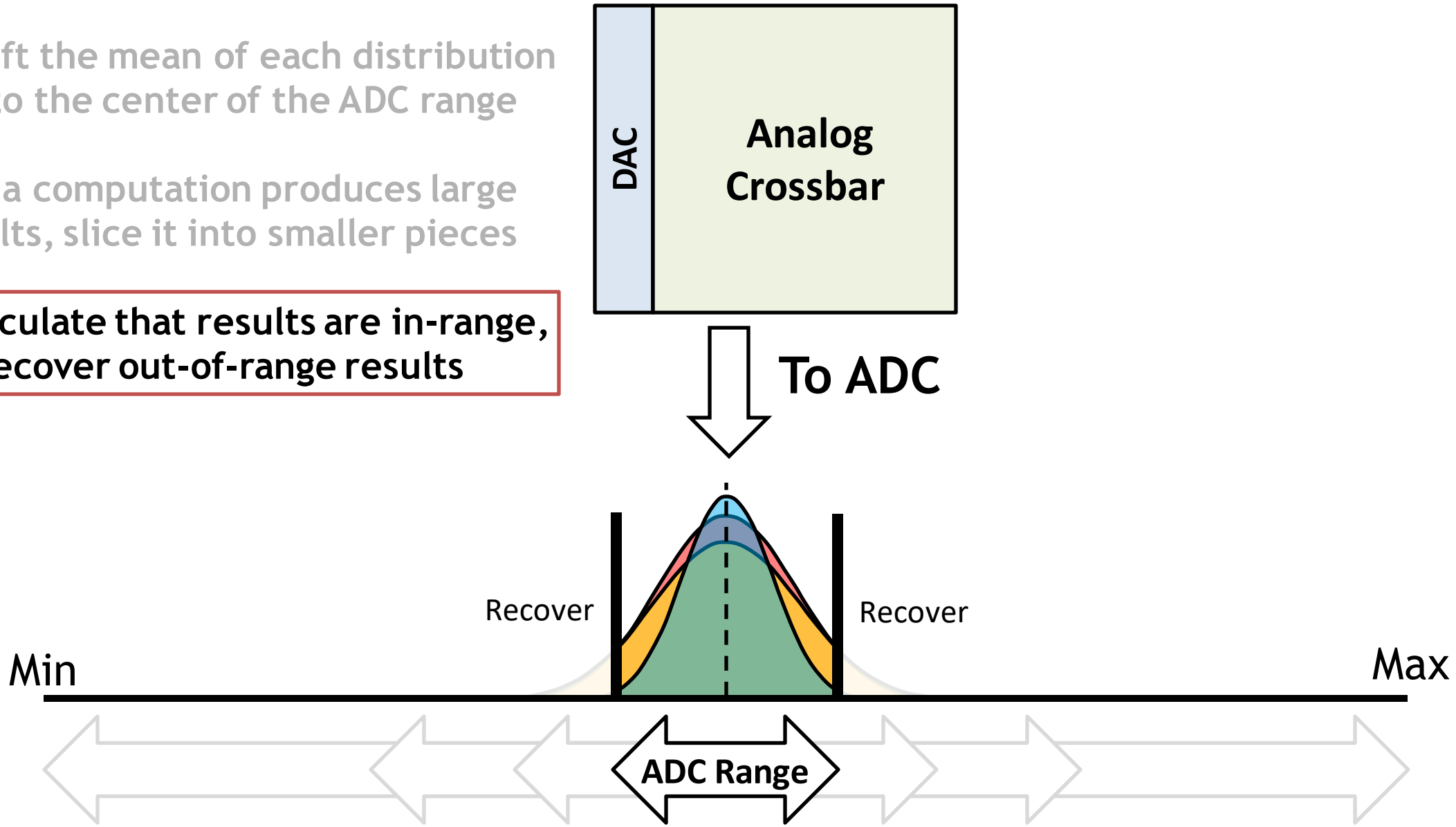


Shifting Distributions

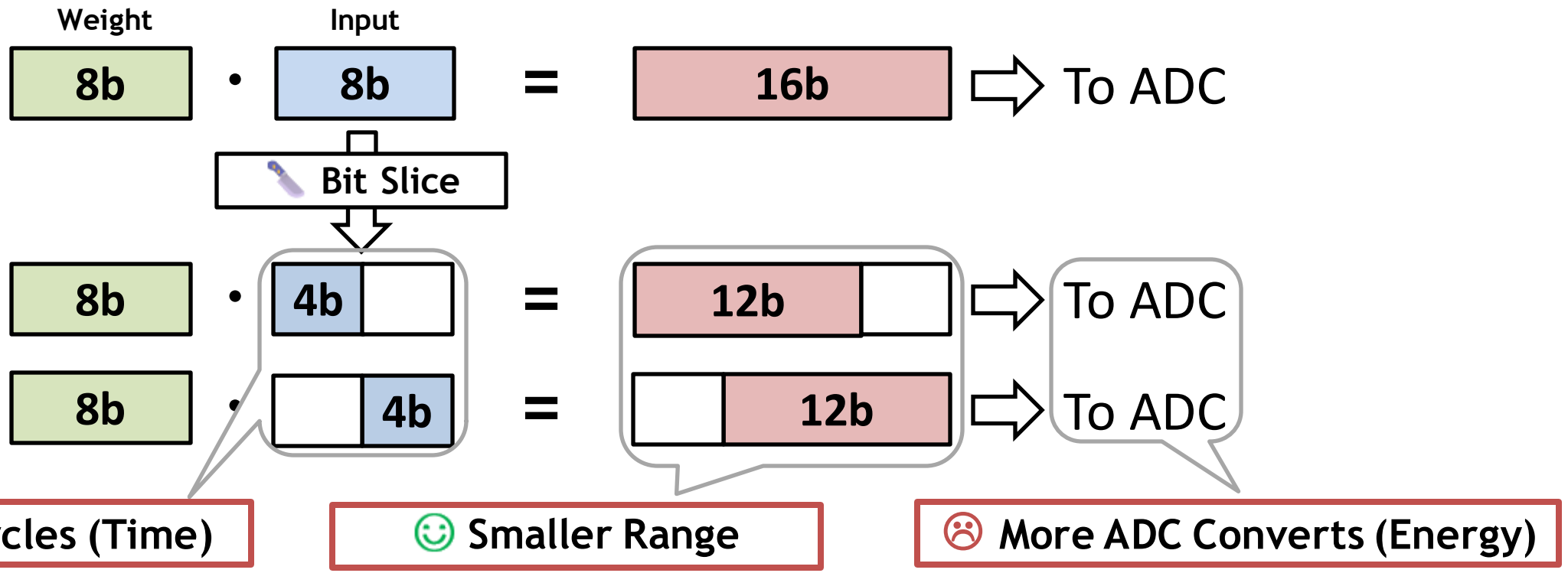
1. Shift the mean of each distribution to the center of the ADC range

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

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

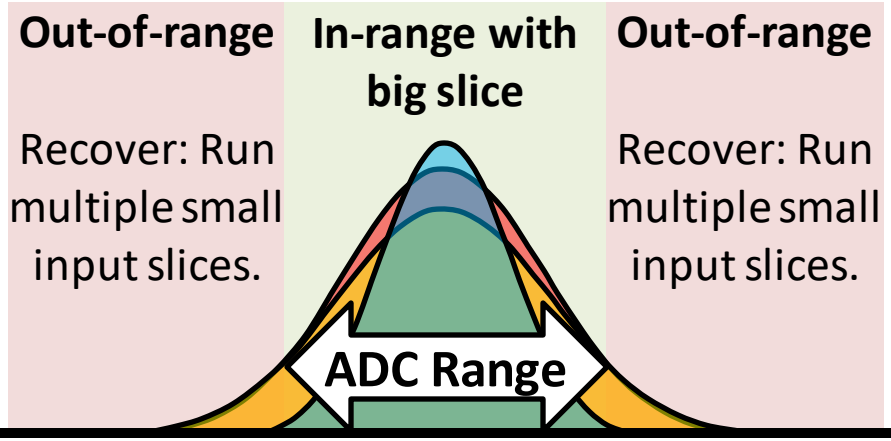


Dynamic Input Slicing: Try Again with Smaller Slices



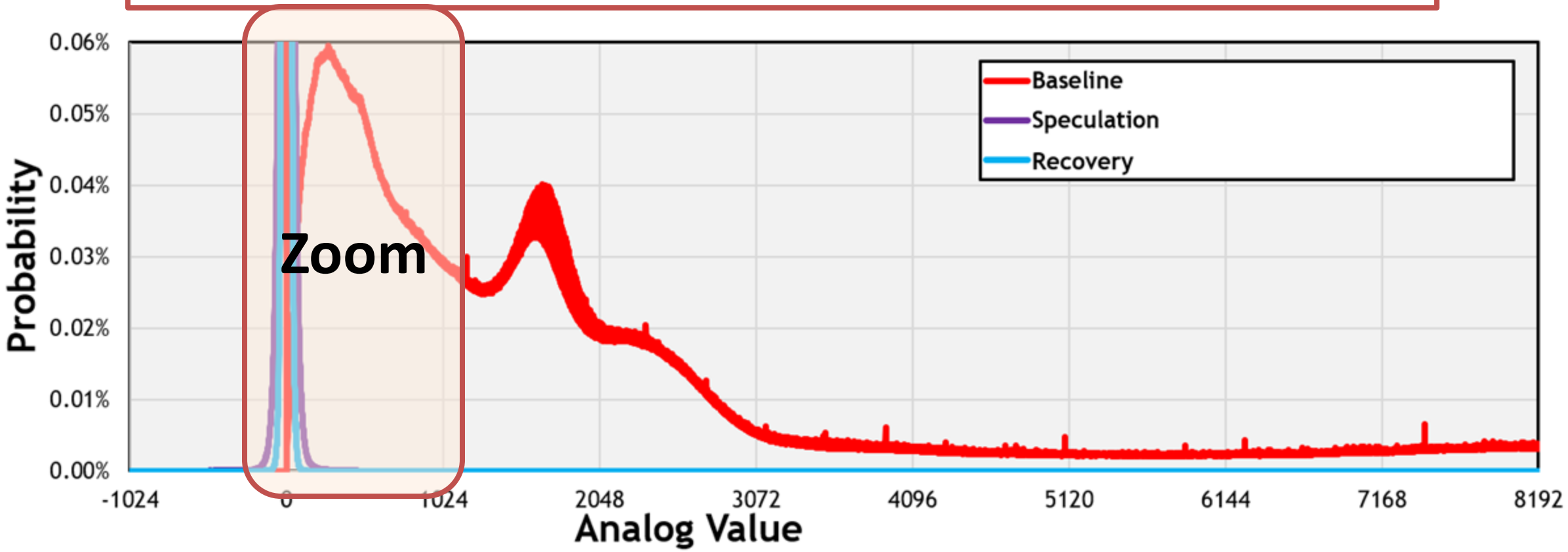
1. Speculate with big input slice

2. Recover out-of-range results with multiple smaller input slices



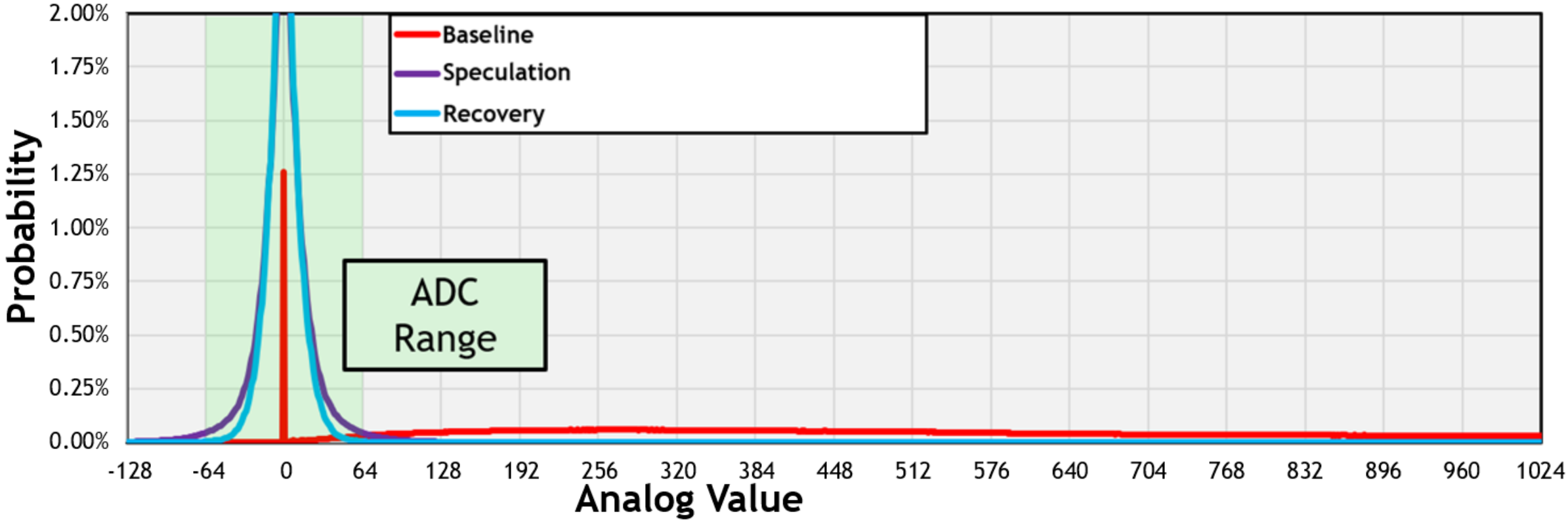
Reshaping the Distributions of Analog Values

1024x reduction in required ADC range



Reshaping the Distributions of Analog Values

1024x reduction in required ADC range



→ Lower Energy ADC (↑ **Efficiency**) and/or

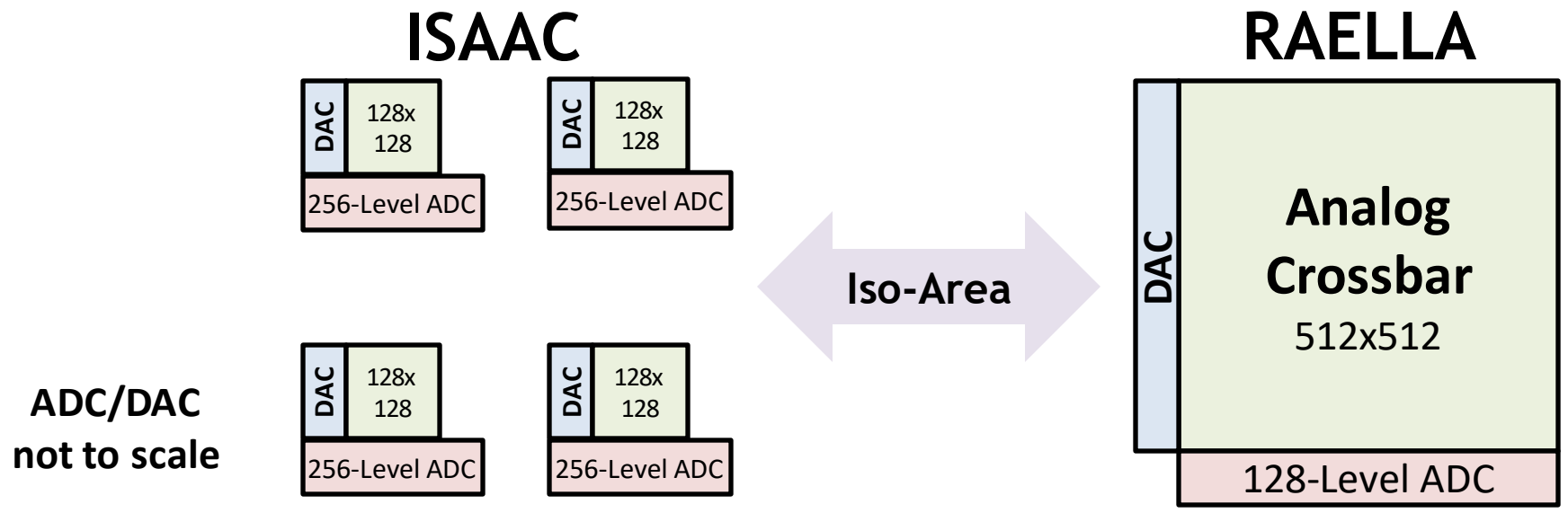
→ More computations per ADC convert (↑ **Efficiency, Throughput**)

Evaluation

- Full System Simulation comparing accelerators ISAAC and RAELLA
- Both low-accuracy-loss, run DNNs without modification/retraining

Green = Compute

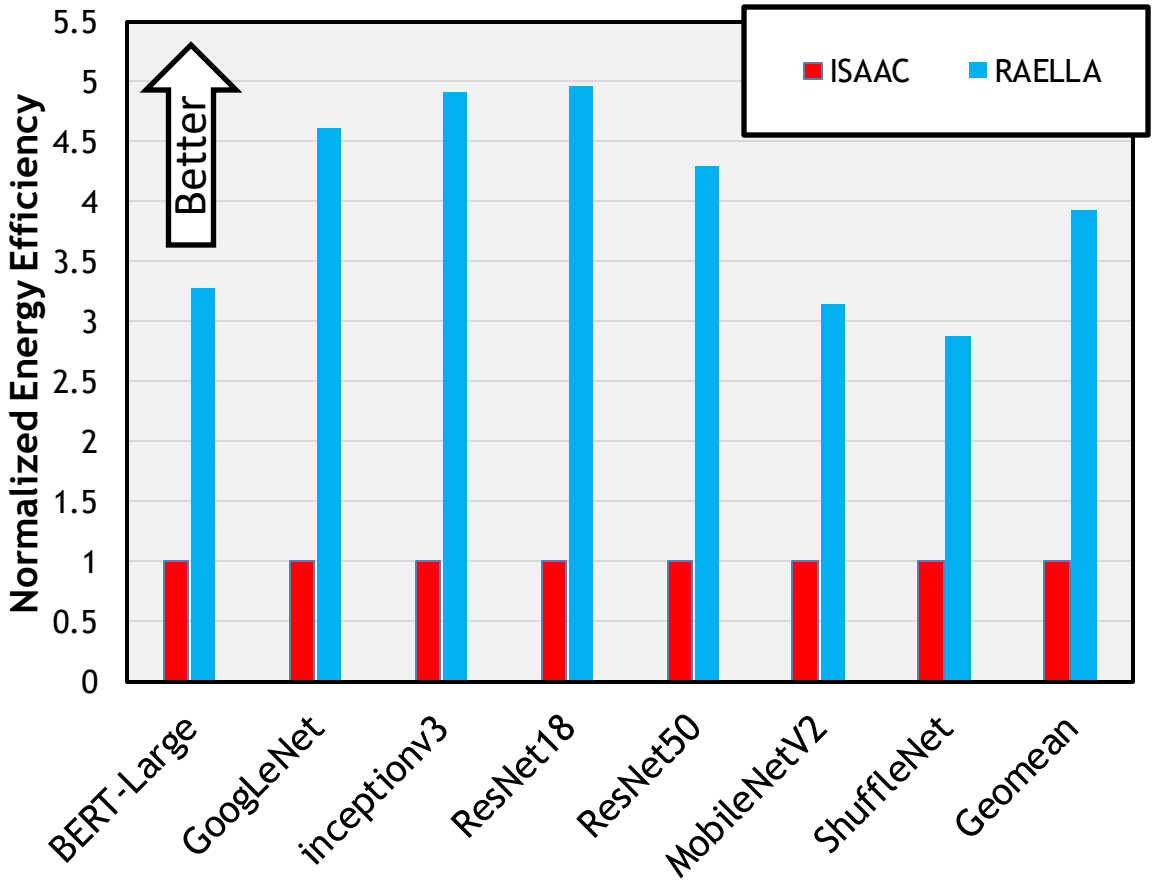
RAELLA gets more compute per unit area, more compute per ADC convert



Evaluation: ISAAC and RAELLA

RAELLA improves efficiency by 3.9x geomean

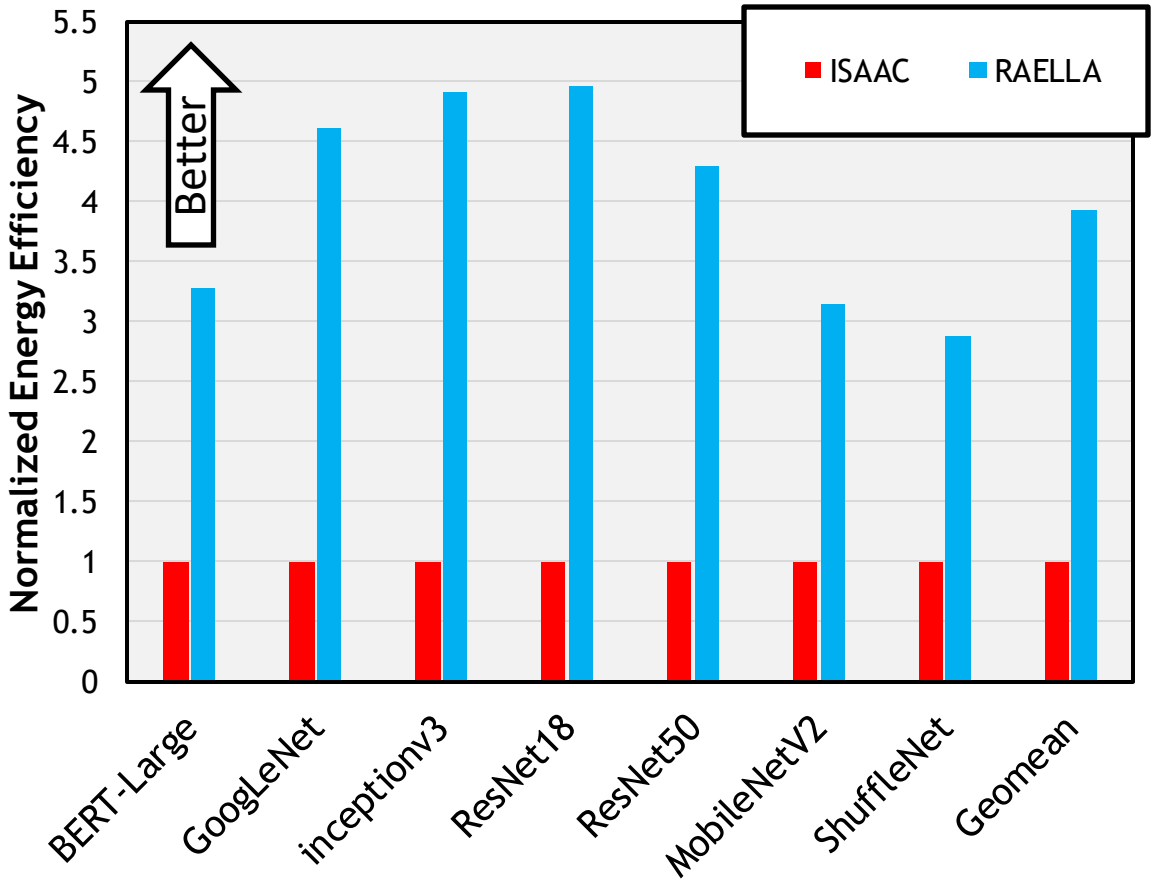
Efficiency



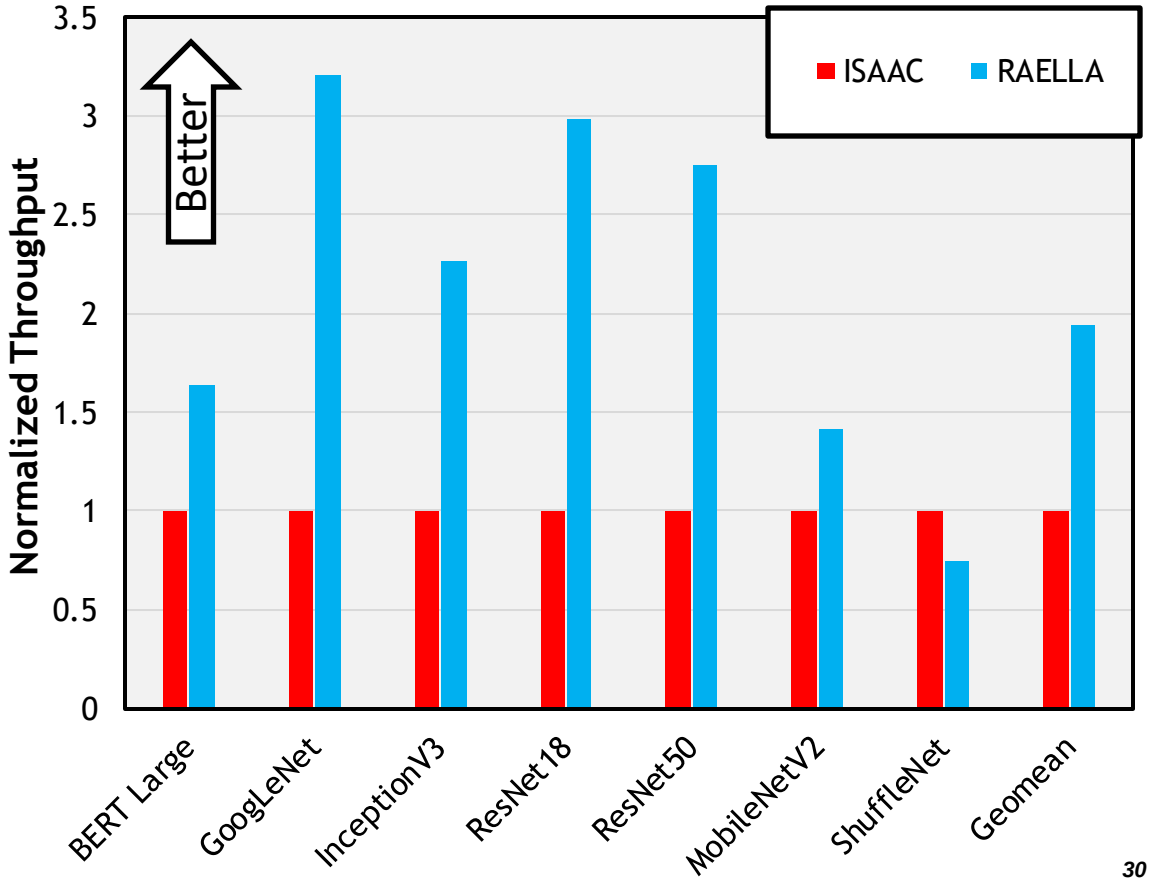
Evaluation: ISAAC and RAELLA

RAELLA improves efficiency by 3.9x geomean
RAELLA improves throughput by 1.8x geomean

Efficiency



Throughput



Key Takeaways

- High ADC energy is a challenge in PIM architectures:
  - **Titanium Law can be used to analyze ADC energy tradeoffs**

- Reduce ADC energy; make analog computations produce small results:
  - **Center+Offset:** Shift the mean of each computed distribution to the center of the ADC range
  - **Adaptive Weight Slicing:** If a computation produces large results, slice it into smaller pieces
  - **Dynamic Input Slicing:** Speculate that results are in-range, recover out-of-range results

- Small-result analog computation enables:
  - **Lower-energy** ADC and/or **more analog compute** with the same ADC range
  - Up to **5x higher efficiency** and **3x higher throughput**
  - **Without modifying or retraining DNNs!**

This work was funded in part by Ericsson, the MIT AI Hardware Program, and MIT Quest

Key Takeaways

- High ADC energy is a challenge in PIM architectures:
  - **Titanium Law can be used to analyze ADC energy tradeoffs**

- Reduce ADC energy; make analog computations produce small results:
  - **Center+Offset:** Shift the mean of each computed distribution to the center of the ADC range
  - **Adaptive Weight Slicing:** If a computation produces large results, slice it into smaller pieces
  - **Dynamic Input Slicing:** Speculate that results are in-range, recover out-of-range results

- Small-result analog computation enables:
  - **Lower-energy** ADC and/or **more analog compute** with the same ADC range
  - Up to **5x higher efficiency** and **3x higher throughput**
  - **Without modifying or retraining DNNs!**

This work was funded in part by Ericsson, the MIT AI Hardware Program, and MIT Quest