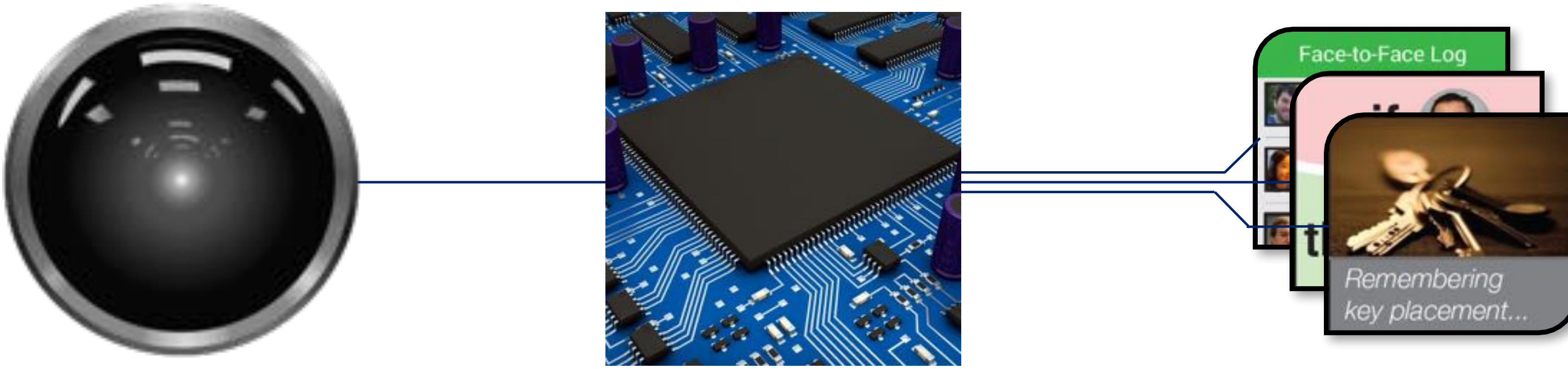
RedEye

Analog ConvNet Image Sensor Architecture for Continuous Mobile Vision

Robert LiKamWa ~~roblkw@rice.edu~~ likamwa@asu.edu
Yunhui Hou houyh@rice.edu
Yuan Gao ~~yg18@rice.edu~~ julianyg@stanford.edu
Mia Polansky mia.polansky@rice.edu
Lin Zhong lzhong@rice.edu

RICE

A vision of vision...



Sense Compute Interact

Energy efficiency goal: 10 mW

- Idle power consumption of smartphone
- Week-long use of small battery (2 Wh)
- Opens door to energy-harvesting solutions

... continuous mobile vision!

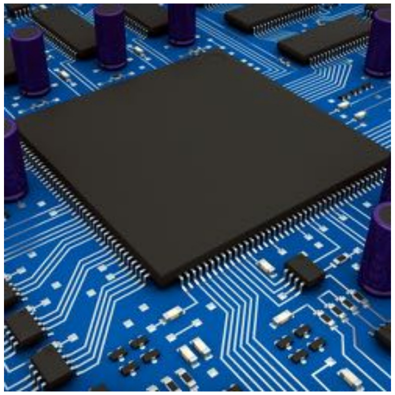
Vision demands energy

Sense

1 nJ per pixel

Ultra-low-power CMOS imager

(Himax 2016)



Compute

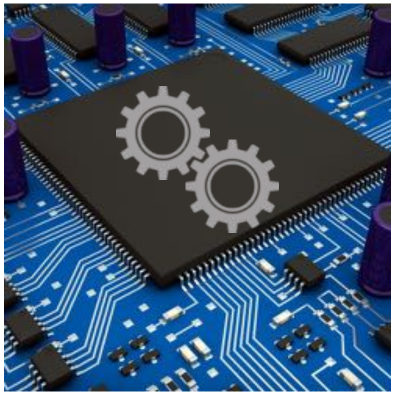
12 nJ per data movement

Quantifying Energy Cost of [Mobile] Data Movement

(Pandiyan, Wu IISWC 2014)

Key Idea:
Shift processing into the analog domain!

Process + **Sense**



Compute

Analog Challenges:

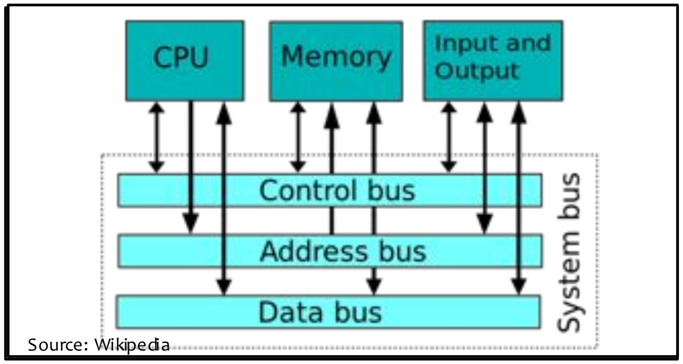
Design complexity

Noisy signal fidelity

Challenge #1: Design complexity

No bus for control/data

- Analog exchanges data on pre-routed interconnects
- Congestion and overlap cause parasitics



Complexity limits the extent of analog computing

Challenge #2: Noisy signal fidelity

Analog circuits suffer from

thermal noise

$$\overline{v_n^2} = k_B T / C$$

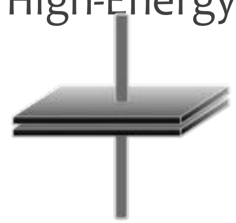
or

energy cost

$$E = CV^2/2$$

Low C
High-noise
Low-Energy

High C
Low-noise
High-Energy



Accumulating signal noise limits the extent of efficient analog computing

Complexity and noise limit the efficiency of prior analog architectures

Analog neural processing
(St. Amant et al @ UT-Austin, 2014)

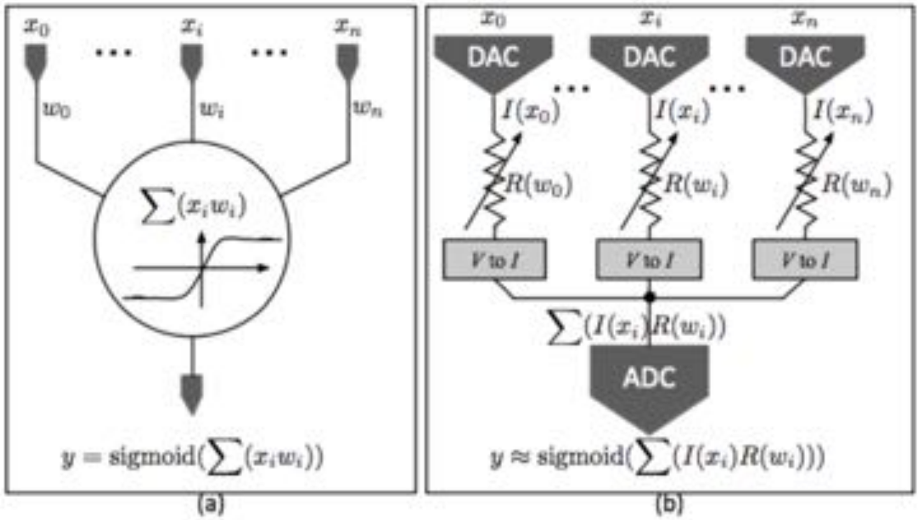
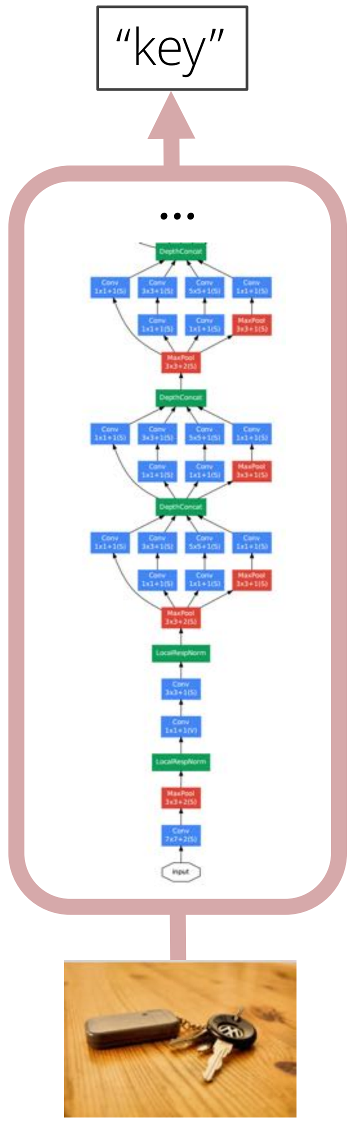


Figure 2: One neuron and its conceptual analog circuit.

ADC consumes >90% of energy consumption

Insight #1:
Vision is highly structured

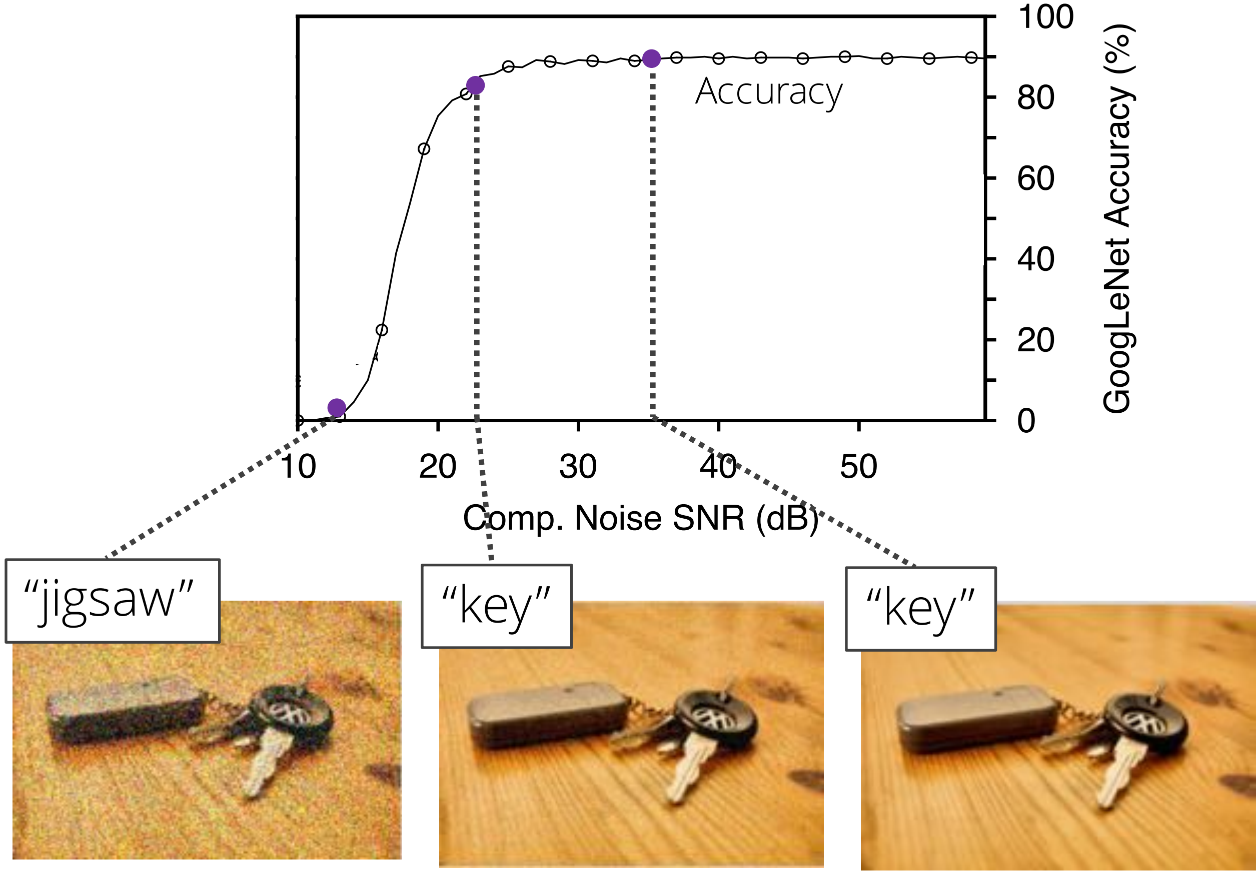
ConvNet blocks
Convolution
Max Pooling



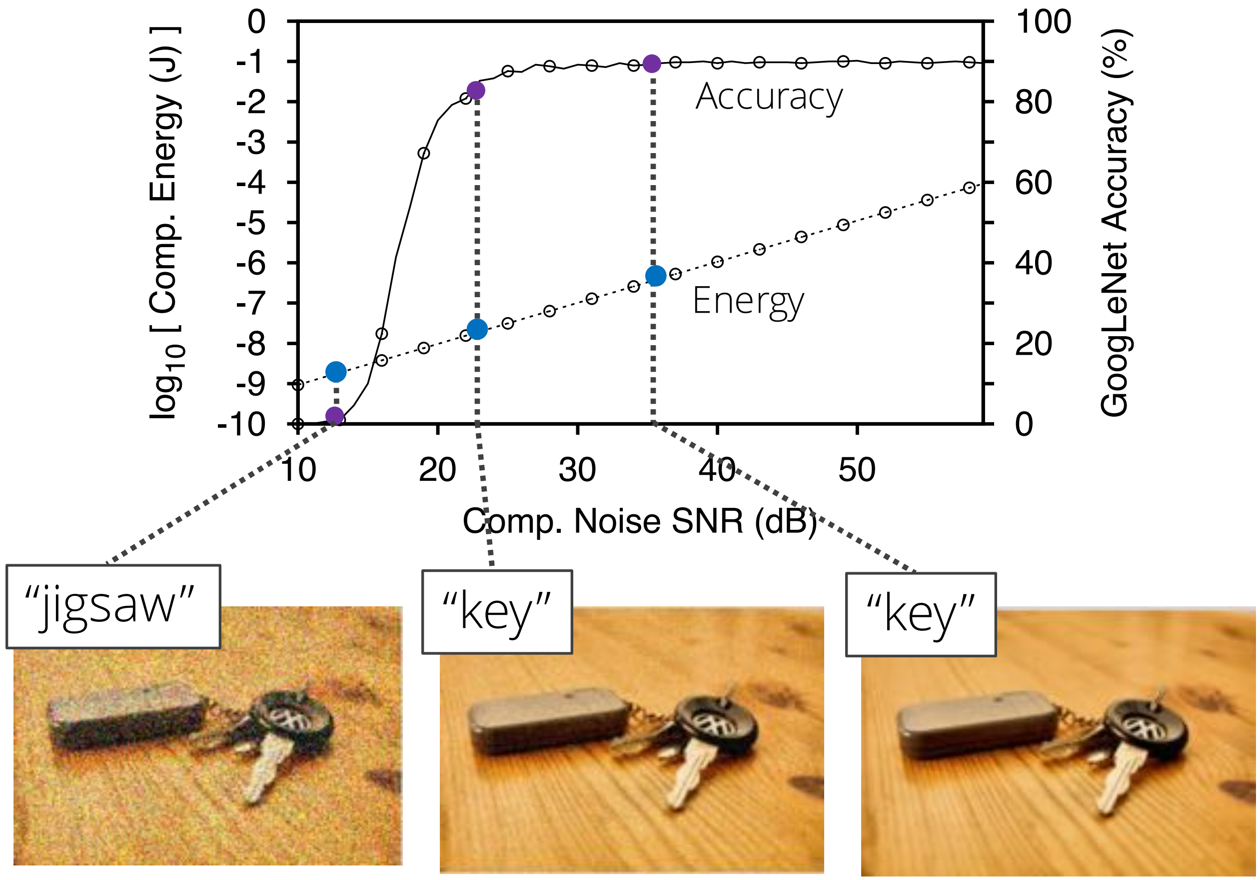
- Repetitive building blocks
  - Reusable structure
- Patch-based operations
  - Data locality
- Dataflow bandwidth reduces with processing
  - "Feed-forward"

What about noise?

Insight #2: Noisy images are okay for vision



Insight #2: Noisy images are okay for vision



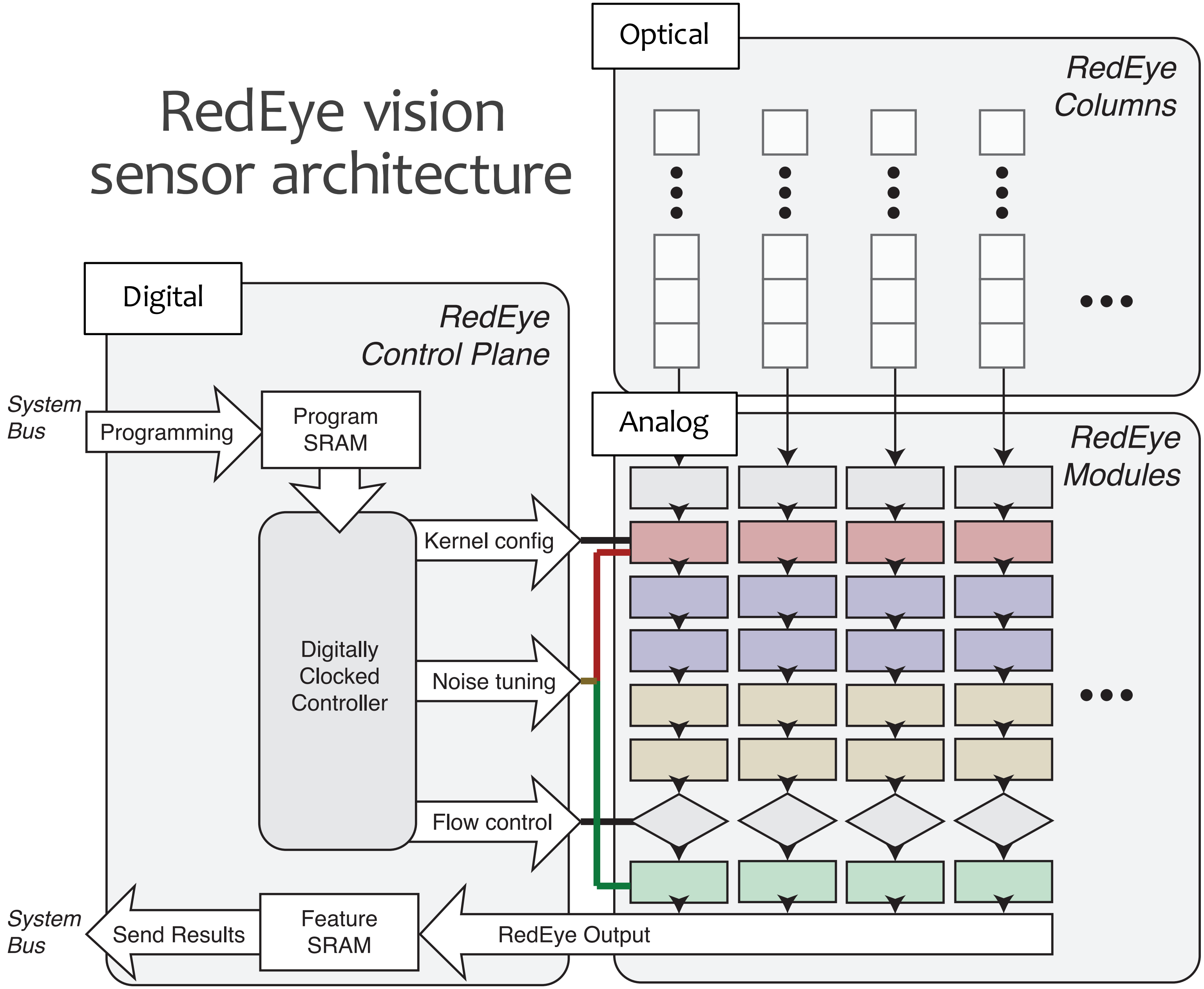
RedEye
vision sensor architecture

Programmable analog ConvNet execution

- Low-complexity modules for design scalability
- Noise mechanisms to trade accuracy/efficiency

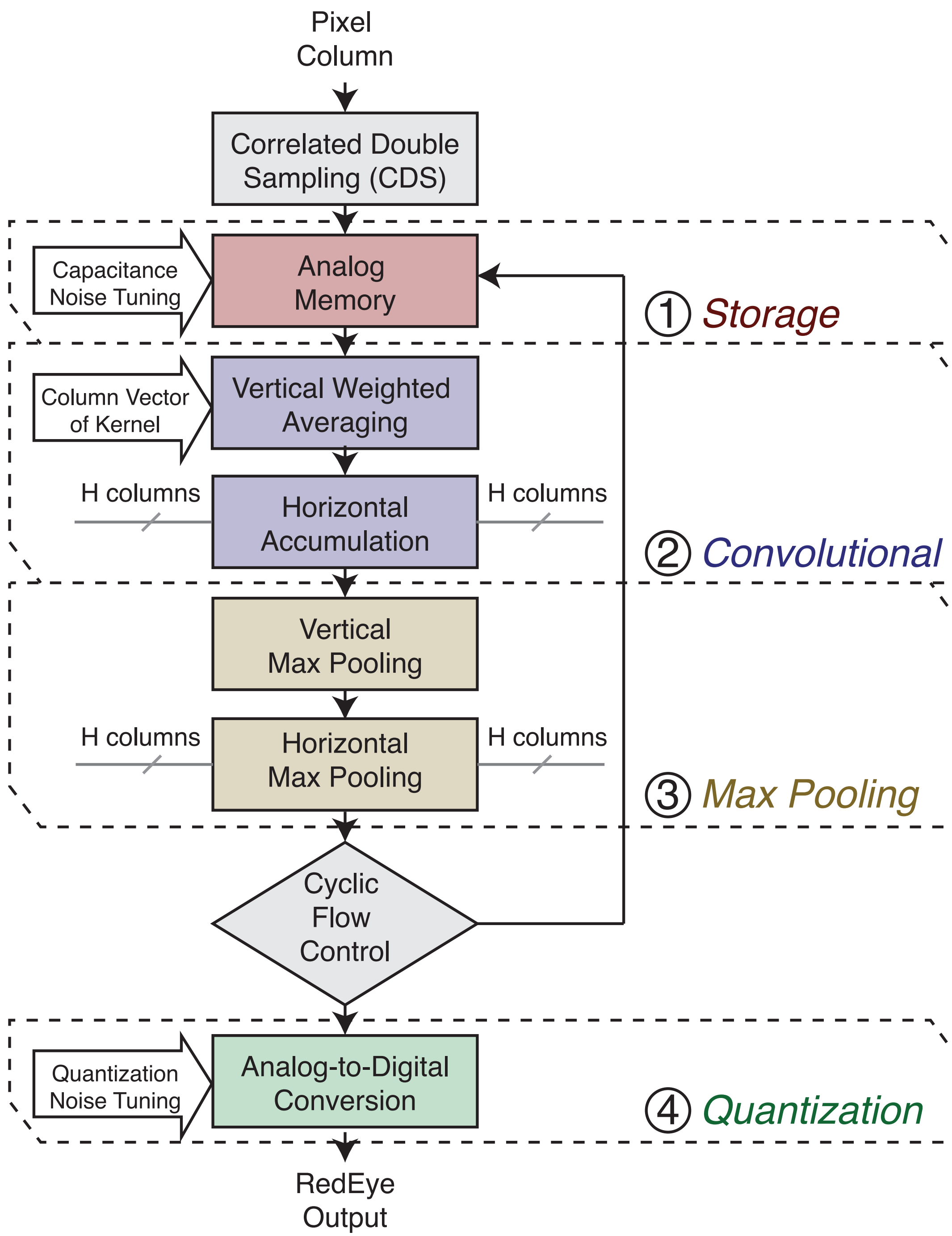
Reduce readout energy by 100x

RedEye vision sensor architecture



Reusable Modules

- Programmable kernel
- Cyclic flow for reuse

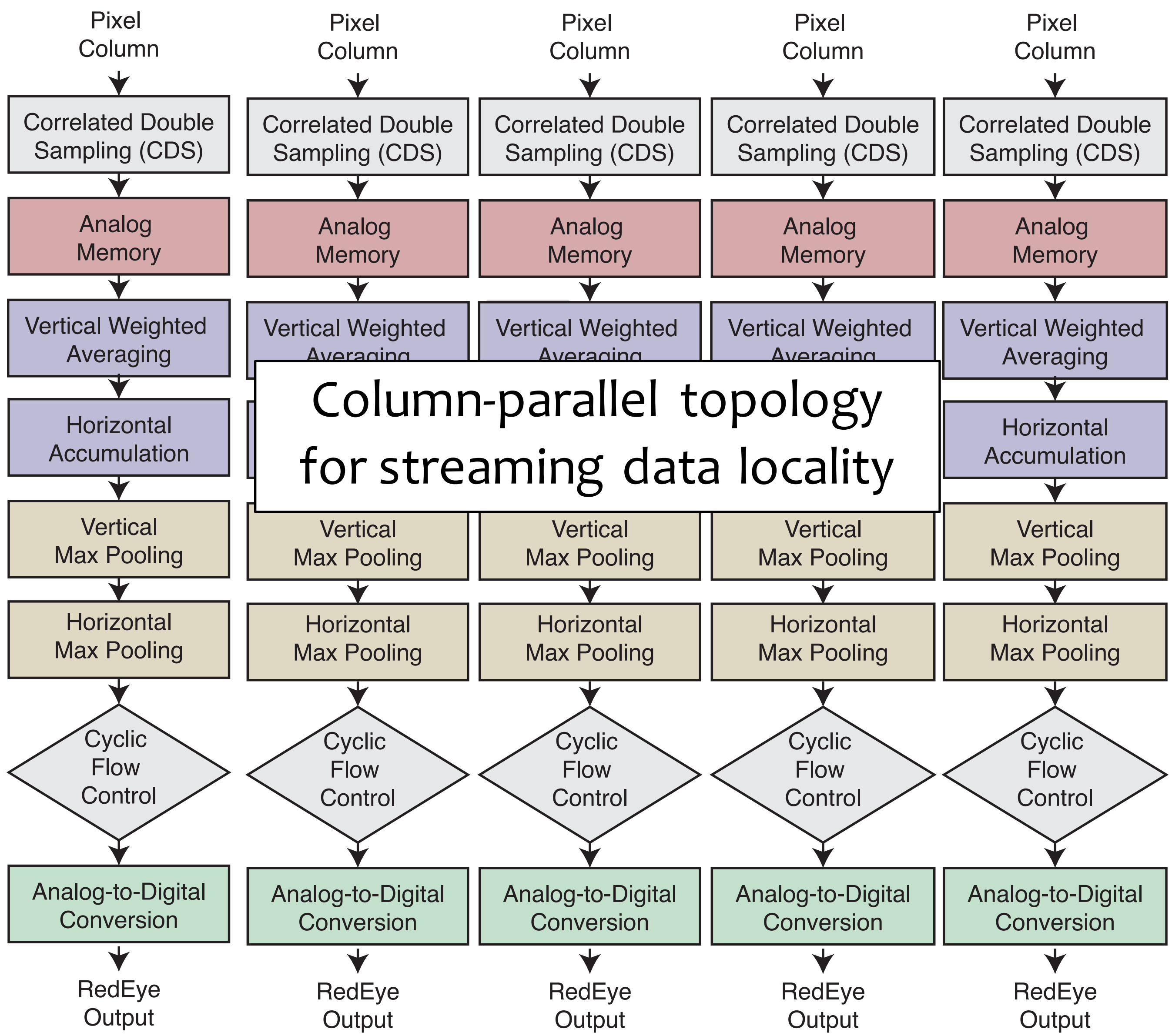


Reusable Modules

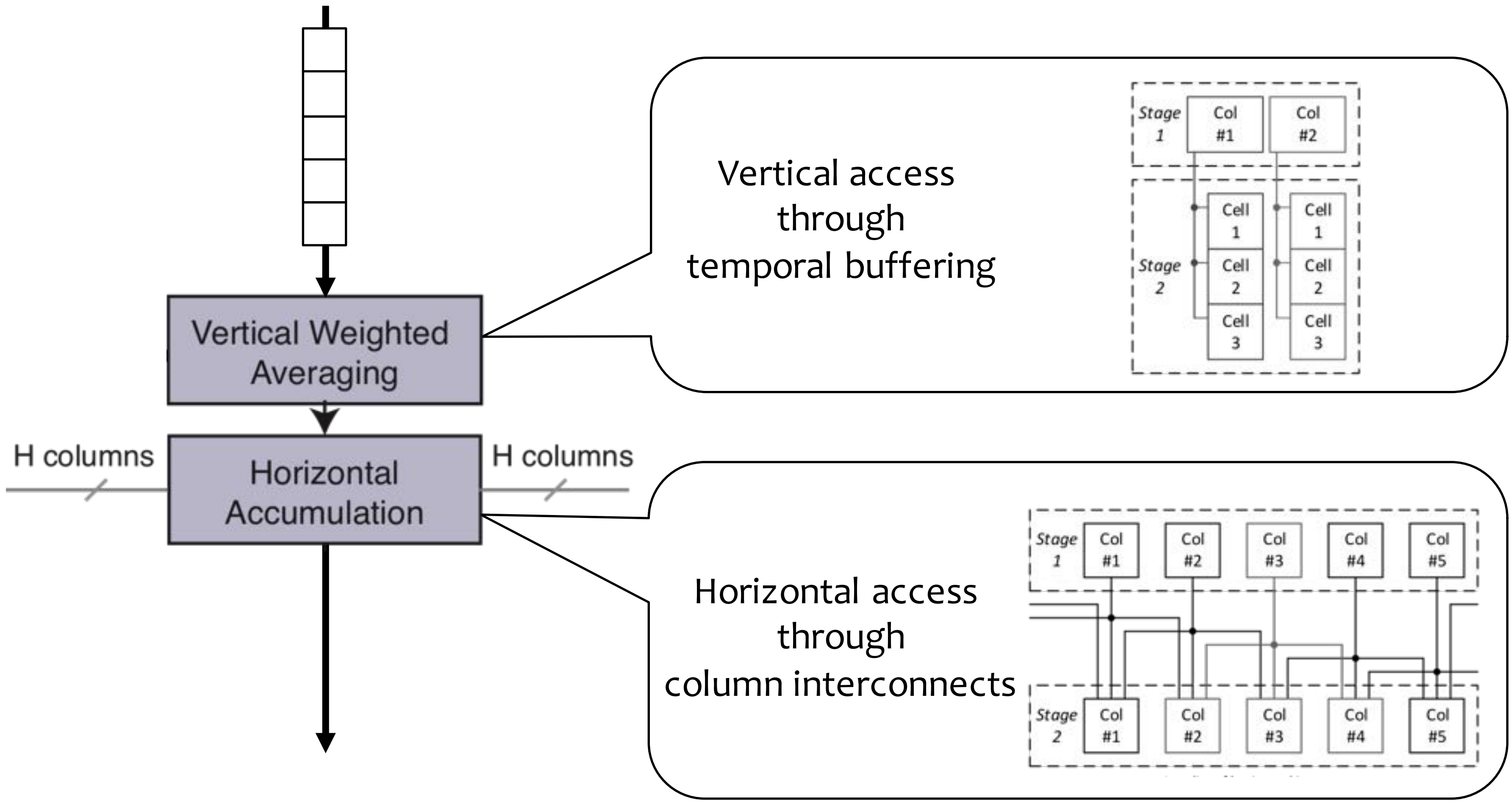
- Programmable kernel
- Cyclic flow for reuse

Data locality for patches

- Streaming processing
- Column topology



Streaming patch-based access

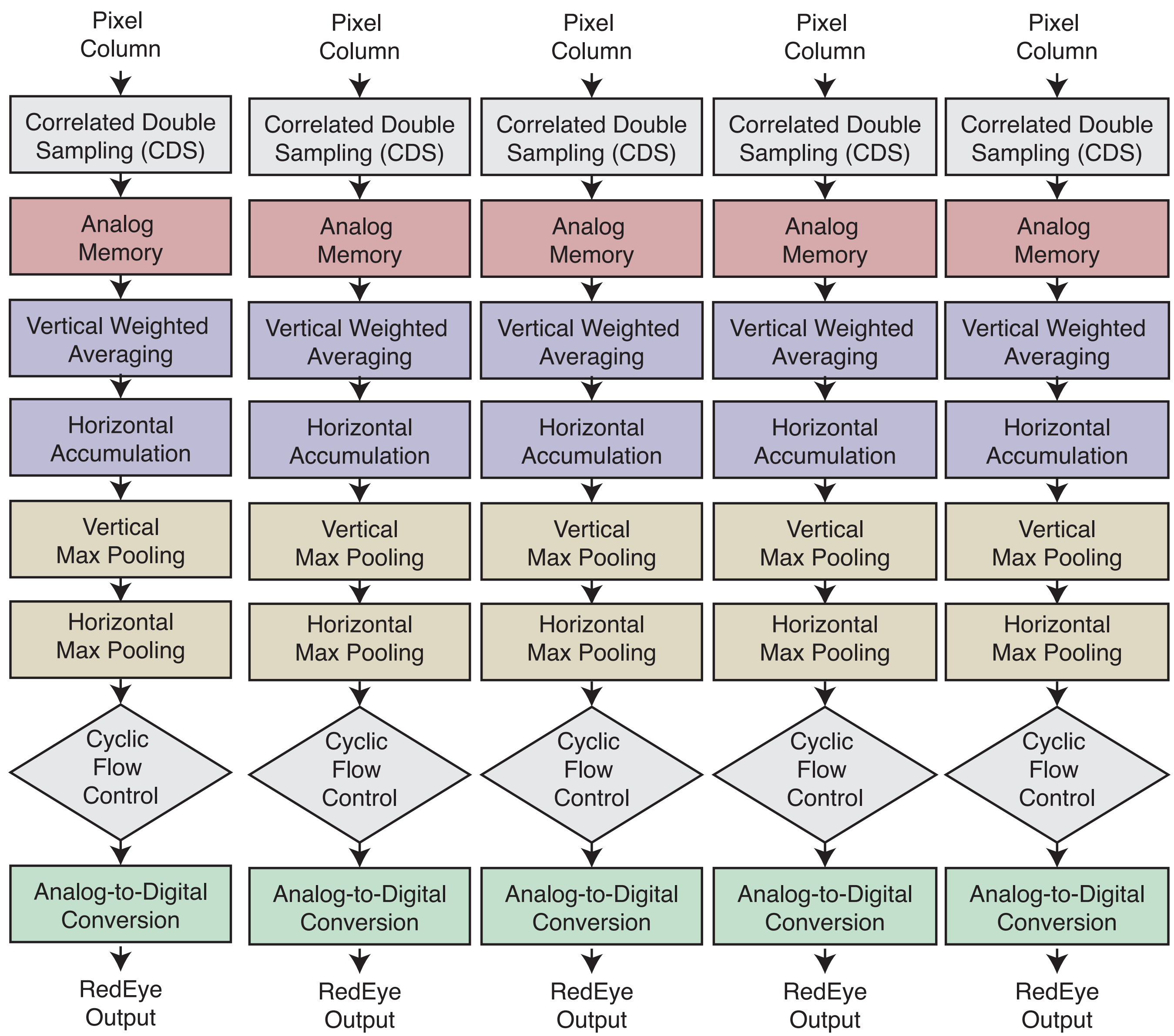


Reusable Modules

- Programmable kernel
- Cyclic flow for reuse

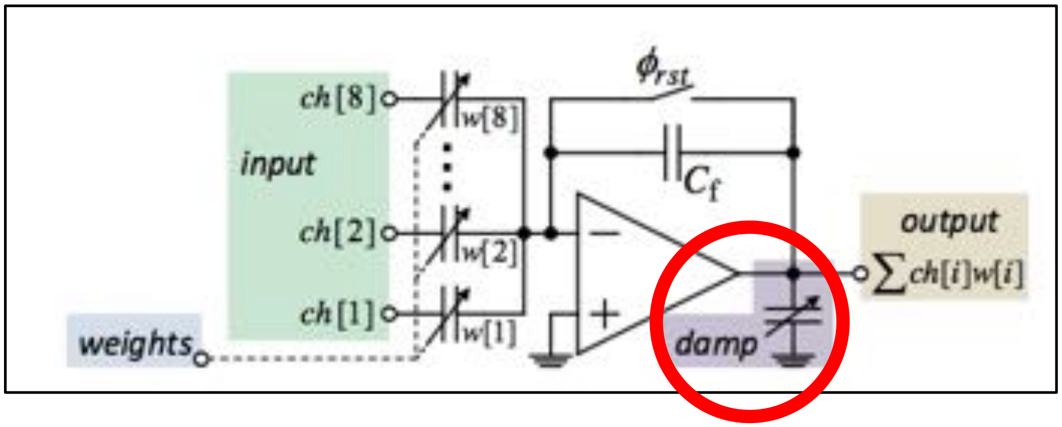
Data locality for patches

- Streaming processing
- Column topology

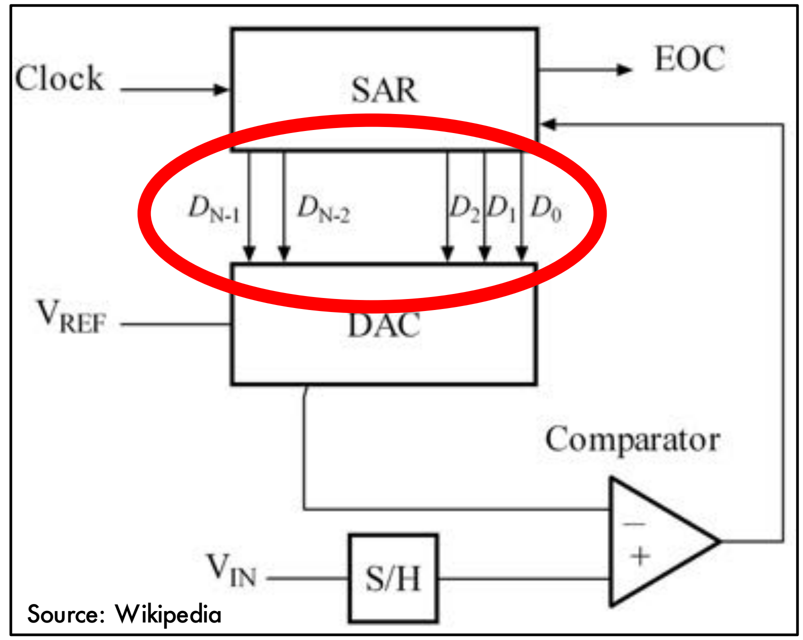


Noise-tuning mechanisms

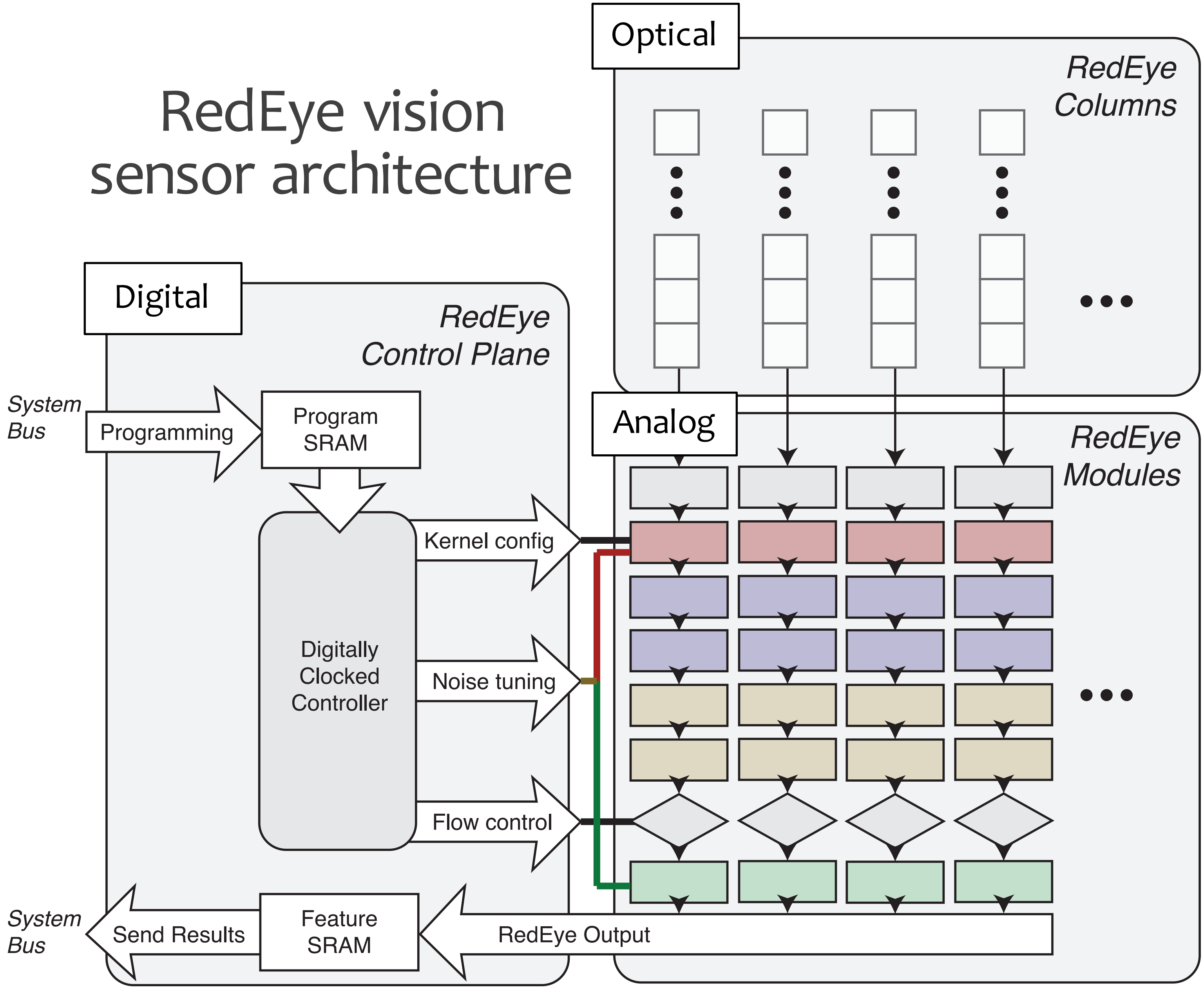
Mixed-signal Multiply-Accumulate
w/tunable fidelity vs. efficiency



SAR ADC
w/tunable-resolution vs. efficiency

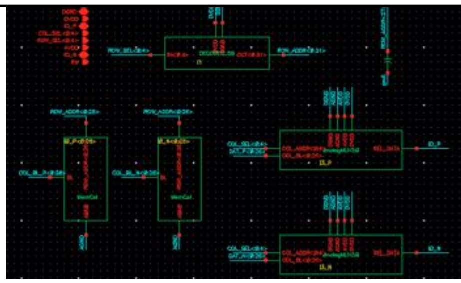


RedEye vision sensor architecture



Estimation and Evaluation

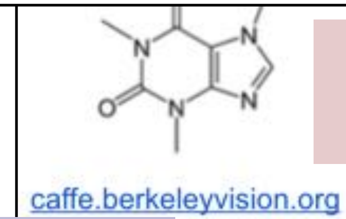
Cadence Spectre



Noise
Power
Timing

Parametrized Behavioral Model

RedEye-caffe Sim. Framework

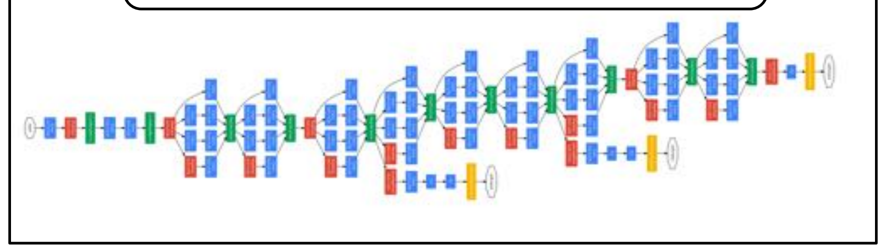


+ Quantized Weights

+ Processing Noise Layer

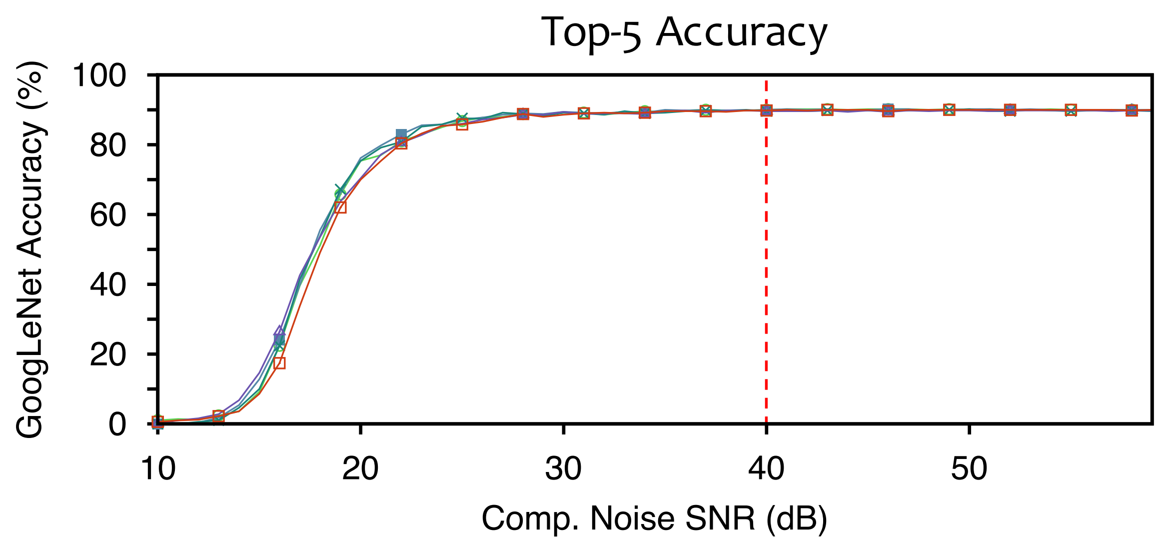
+ Quantization Noise Layer

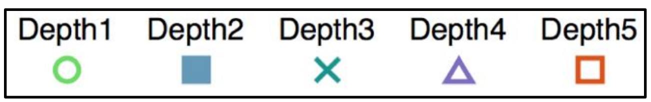
RedEye+GoogLeNet_v1

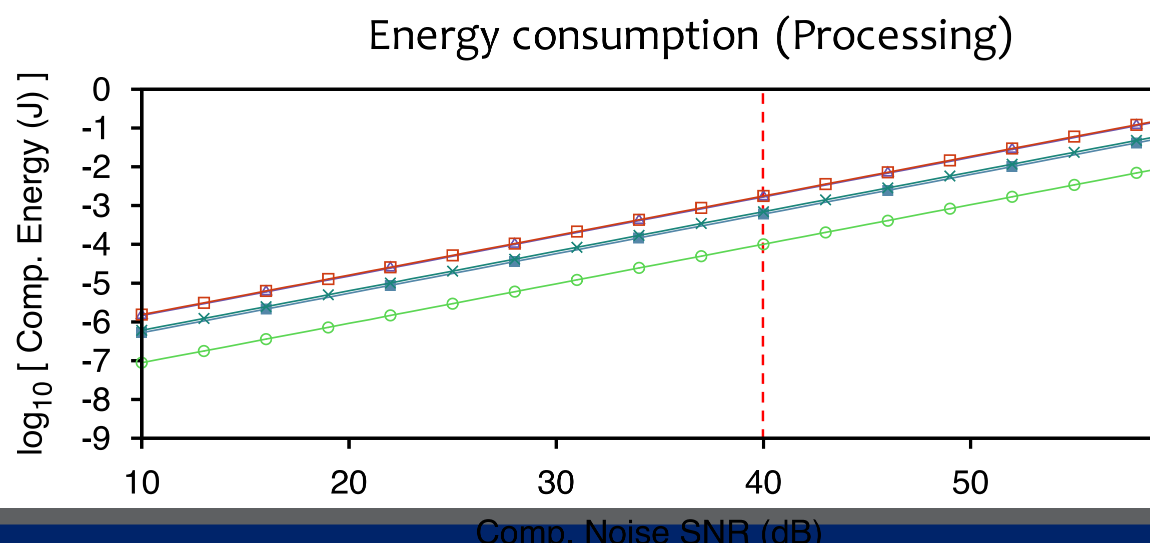


https://github.com/JulianYG/redeye_sim

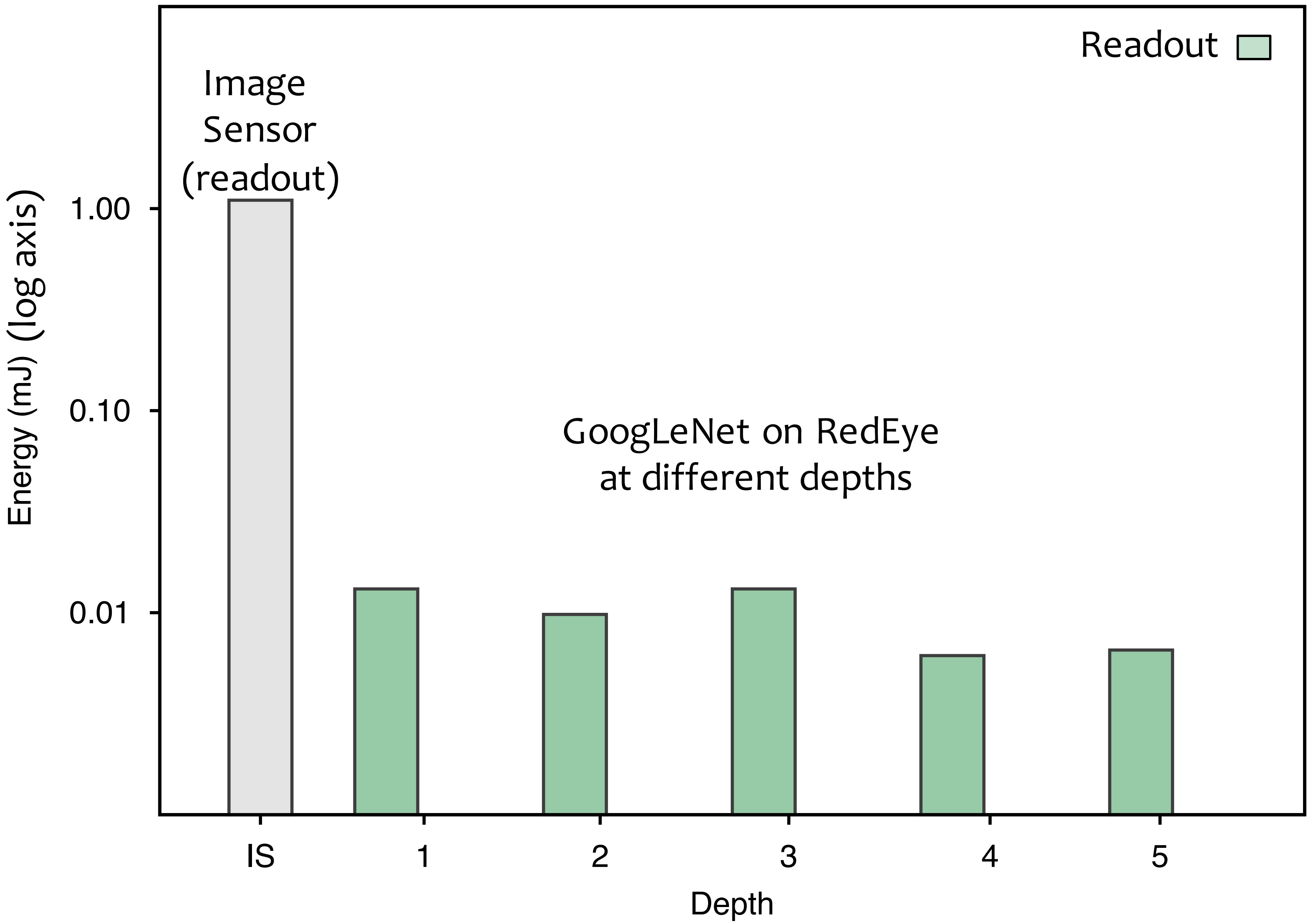
Admitting noise saves energy!
(but our current process limits us to 40 dB)



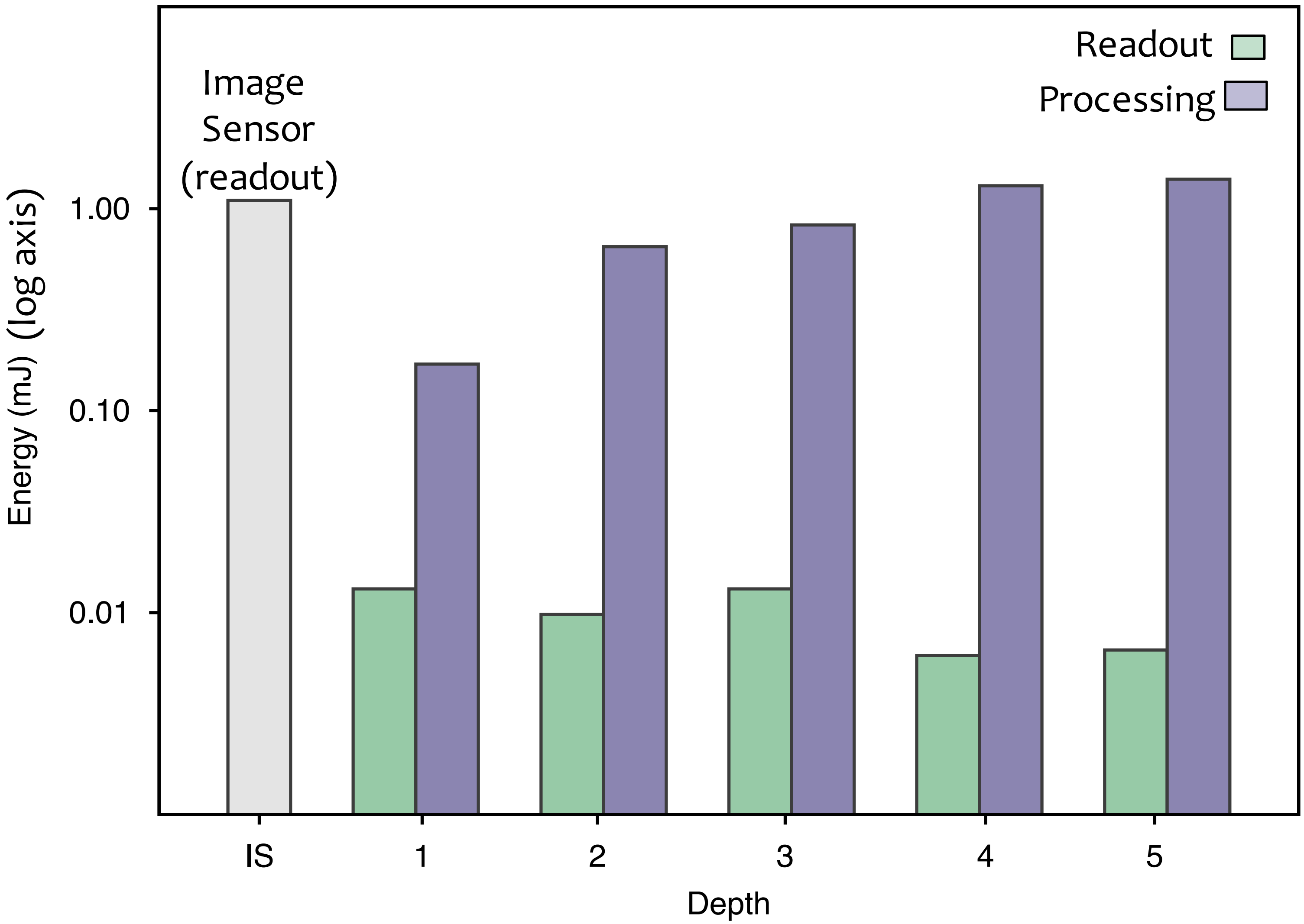




RedEye reduces **readout energy by >100x**

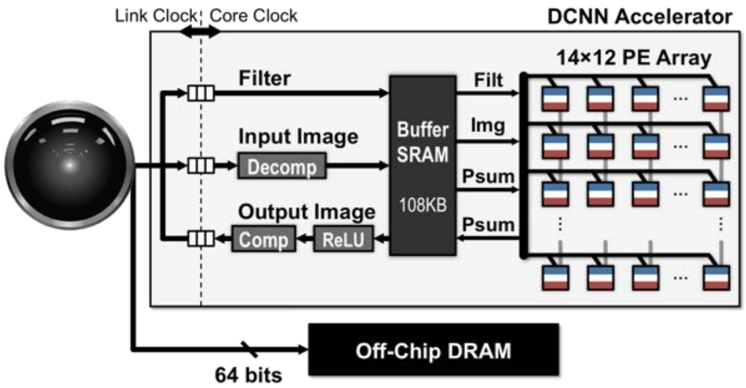


RedEye reduces readout energy by >100x at expense of processing energy

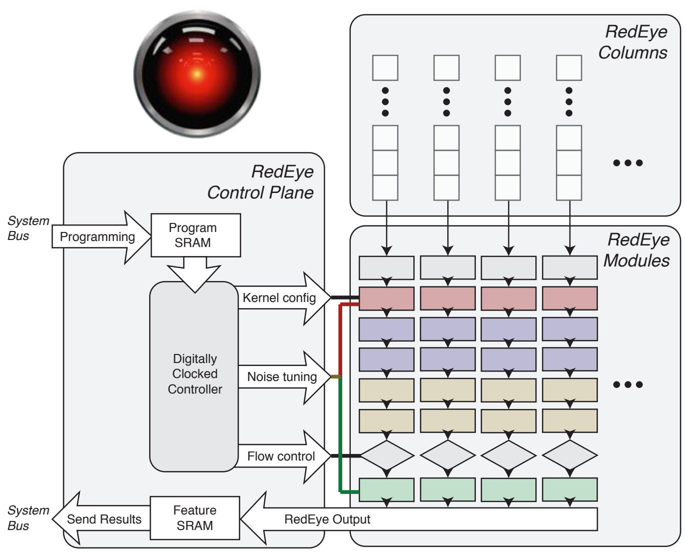


RedEye can help state of the art ConvNet processing efficiency by 2x

EyeRiss [ISCA '16, ISSCC '16] Chen et al



<u>Eyeriss+ Image Sensor:</u>
EyeRiss (Conv Layers): 5.9 mJ
Image Sensor: 1.0 mJ
EyeRiss (Full Layers): 2.1 mJ
Total: 9.0 mJ



<u>EyeRiss + RedEye:</u>
RedEye (Analog Conv): 2.5 mJ
RedEye Readout: 0.001 mJ
Eyeriss (Full Layers): 2.1 mJ
Total: 4.6 mJ

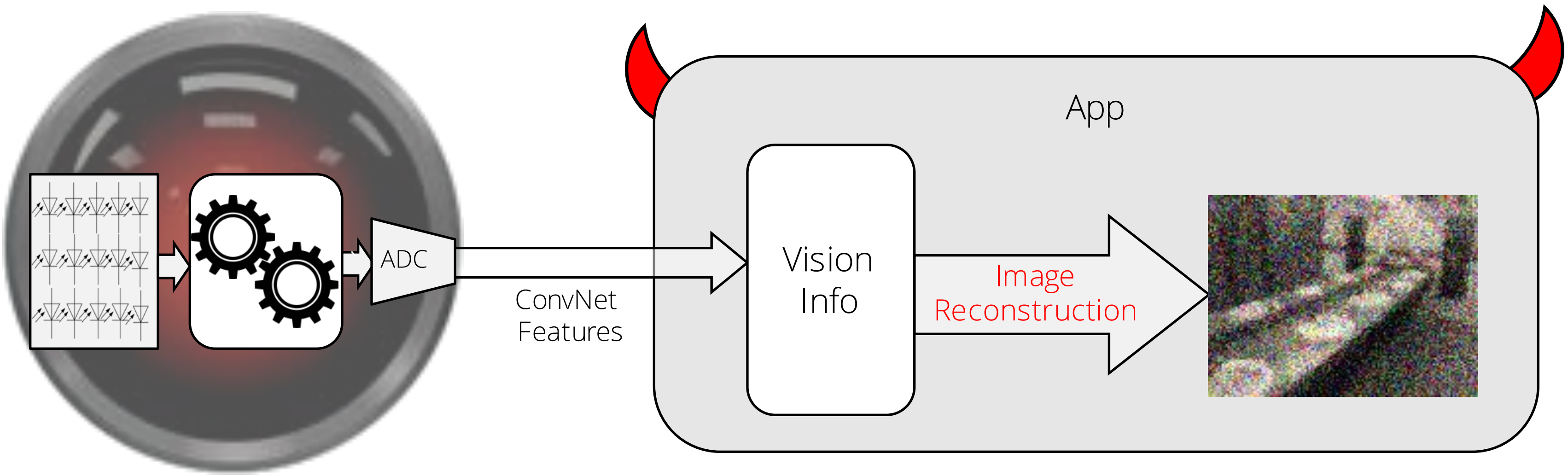
RedEye limitations (and opportunities!)

- RedEye is bounded to 40 dB (Limits energy savings)
  - Unit capacitance of process technology
- ConvNet not optimized for RedEye architecture
- RedEye is strictly feed-forward (no recurrence, e.g., LSTM nets)

Realizing RedEye chip

- Silicon validation in 65 nm TSMC
  - Non-idealities: noise, non-linearity, offset, process variation
  - Opportunities: voltage scaling, sub-threshold circuits

? Raw image privacy through noisy degradation ?



- Idea: App can have vision info, not image data.
- Degrade image and features (e.g., insert noise)
- Ensure vision usability, but image privacy

Depth 1 Reverse Depth 2 Reverse Depth 3 Reverse Depth 4 Reverse Depth 5 Reverse

"Understanding Deep Representations by Inverting Them", Mahendran et al.

Related Work

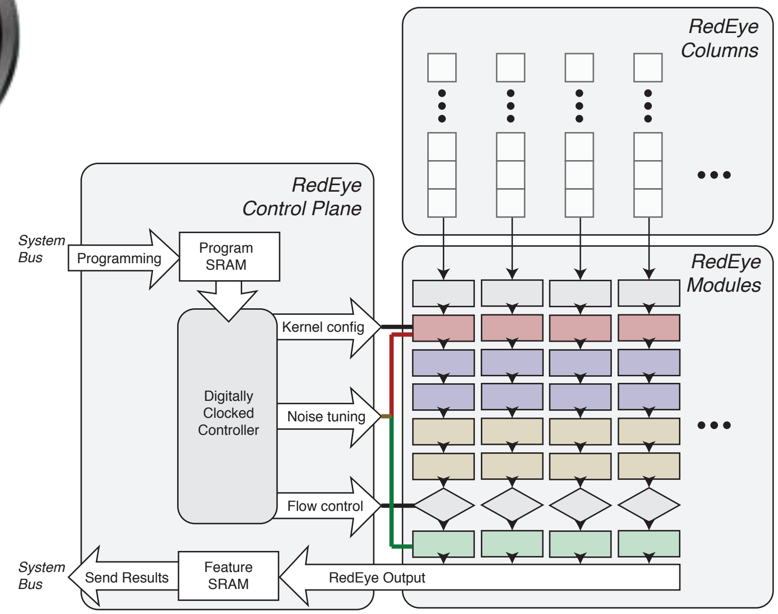
- **Hardware ConvNet acceleration**
  - Reconfigurable flexibility
    - NeuFlow: Dataflow vision processing system-on-a-chip (Pham et al, MSCS 2012)
    - Origami: A convolutional network accelerator (Cavigelli et al, GLSVLSI 2012)
    - A dynamically configurable coprocessor for convolutional neural networks (Chakradhar et al, SIGARCH News 2010)
  - Data Movement reduction
    - Convolution engine: balancing efficiency & flexibility in specialized computing (Qadeer et al, SIGARCH News, 2013)
    - Memory-centric accelerator design for convolutional neural networks (Peemen et al, ICCD 2013)
    - DianNao: A small-footprint high-throughput accelerator for ubiquitous machine-learning. (Chen et al, ASPLOS 2014)
    - **PRIME: A Novel Processing-in-memory Architecture for NN Computation in ReRAM-based Main Memory (Chi et al, ISCA 2016)**
    - **ISAAC: A Convolutional Neural Network Accelerator with In-Situ Analog Arithmetic in Crossbars (Shafiee et al, ISCA 2016)**
    - **EIE: Efficient Inference Engine on Compressed Deep Neural Network (Han et al, ISCA 2016)**
    - **Eyeriss: A Spatial Architecture for Energy-Efficient Dataflow for Convolutional Neural Networks (Chen et al, ISCA 2016)**

- **Limited-precision ConvNets**
    - General-purpose code acceleration with limited-precision analog computation (St. Amant et al, ISCA 2014)
    - Continuous real-world inputs can open up alternative accelerator designs (Belhadj et al, SIGARCH News 2013)
    - **Minerva: Enabling Low-Power, Highly-Accurate Deep Neural Network Accelerators (Reagen et al, ISCA 2016)**

RedEye

Analog ConvNet
Image Sensor Architecture
for
Continuous Mobile Vision

Robert LiKamWa likamwa@asu.edu
Yunhui Hou houyh@rice.edu
Yuan Gao julianyg@stanford.edu
Mia Polansky mia.polansky@rice.edu
Lin Zhong lzhong@rice.edu



Programmable analog ConvNet execution

- Modules for design scalability
- Tunable noise for accuracy and efficiency
- Programmability for flexibility

Open-Source simulation framework:

https://github.com/JulianYG/redeye_sim

RICE