Accuracy-Aware Computing: Opportunities and Pitfalls

Marilyn Wolf
School of ECE

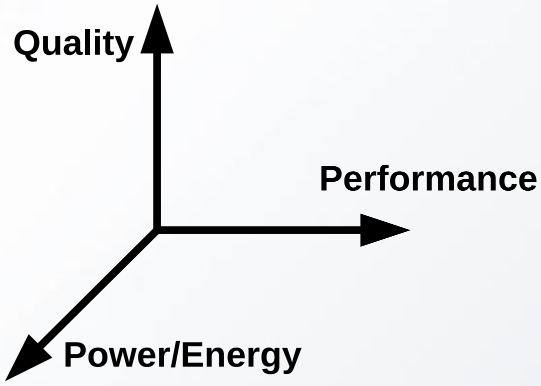
Georgia Institute of Technology

Outline

- Se Hun Kim, Marilyn Wolf, Saibal Mukophdhyay: Error-aware image compression.
  - Perceptual models of image quality.
  - Error models for subthreshold logic.
  - The system-level challenge.
  - Error-aware architectures.

Problem Definition

- Error Tolerance of DSP applications
  - Limit of human perceptual system
  - Relax 100% correctness
  - Acceptable quality
  - Accuracy (Quality)-energy tradeoff



- Research Goal
  - Low power/energy consumption without significant quality degradation

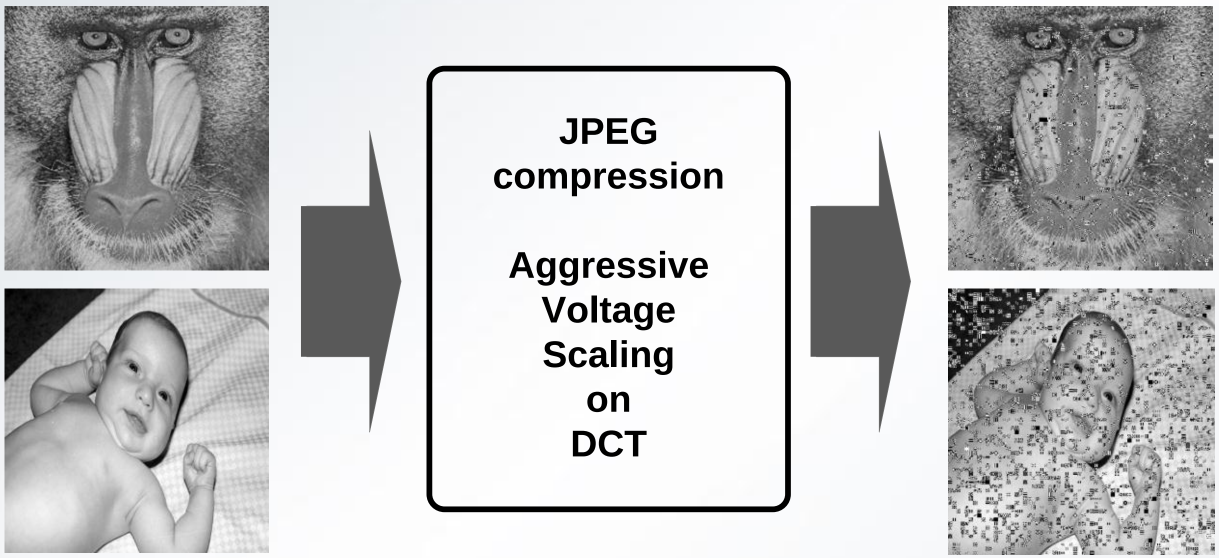
Georgia Institute of Technology

Input dependence on Energy Savings

- Motivation

Img 1

Img 2



MSSIM = 0.5577

MSSIM = 0.3402

- Two reasons
  1. The natural disparity in error tolerance due to the characteristics of the human visual system
  2. Error rate depends on image types under aggressive voltage scaling

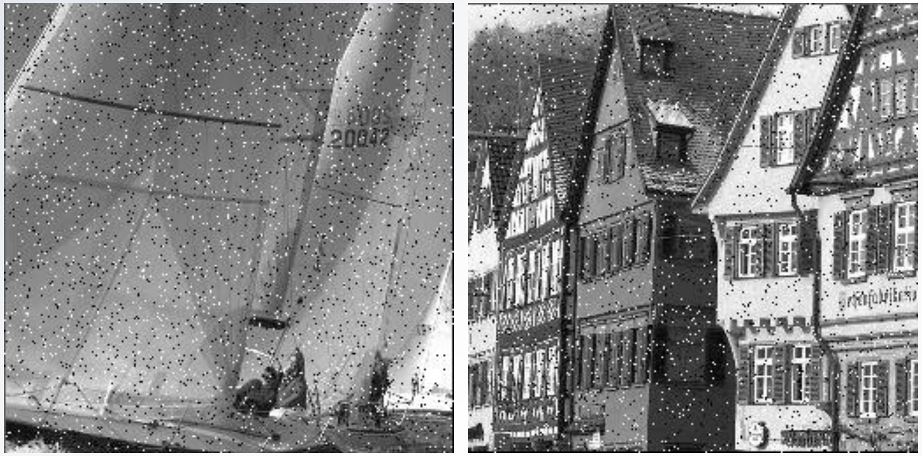
* MSSIM : Mean Structural SIMilarity – a perceptual image quality assessment

Georgia Institute of Technology

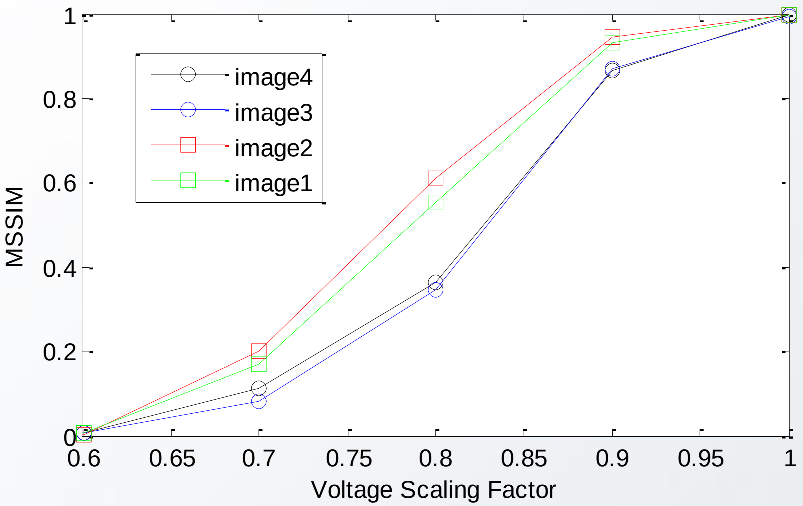
Natural disparity in error tolerance

- Several aspects of human visual system
  - Contrast masking, Texture masking, Frequency masking
  - Discrimination threshold depends on these three

➔ Images with higher contrast, texture, and frequency tolerate more error signals



Comparison of two images with same salt & pepper error



SRAM simulation result under aggressive voltage scaling, images 1 and 2 have higher contrast, texture, and frequency than image 3 and 4

Structural Similarity Index Measurement (SSIM)

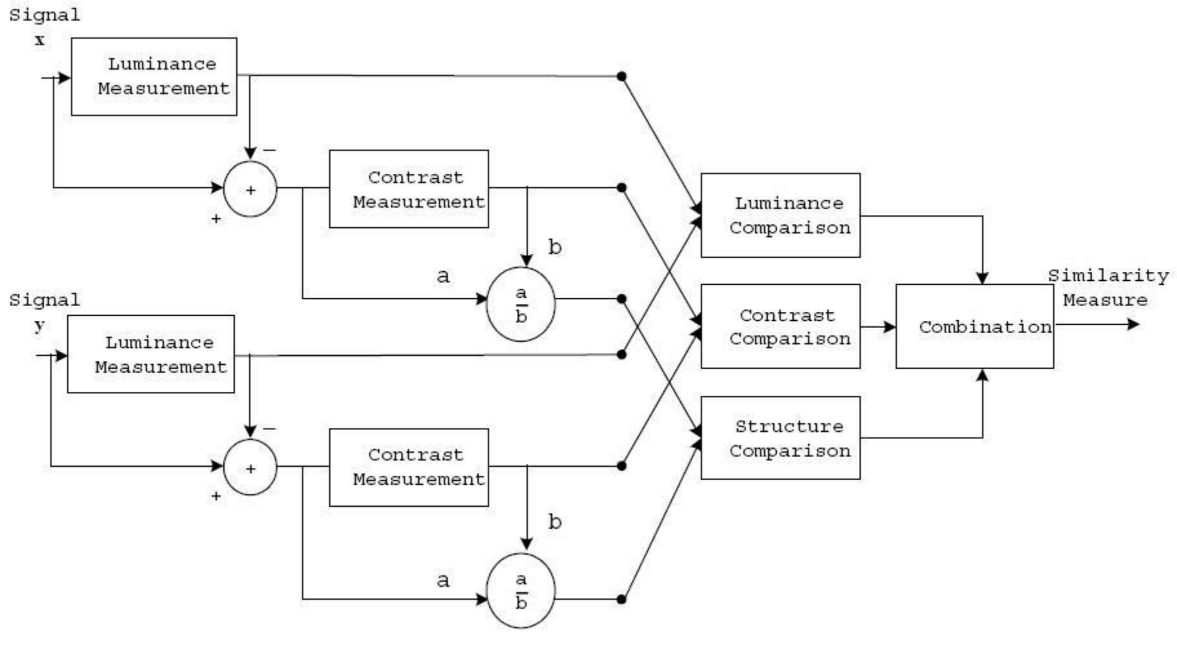
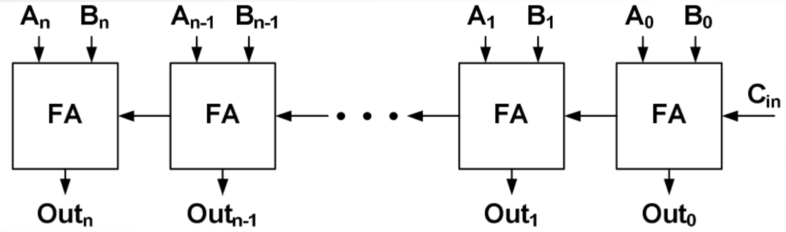


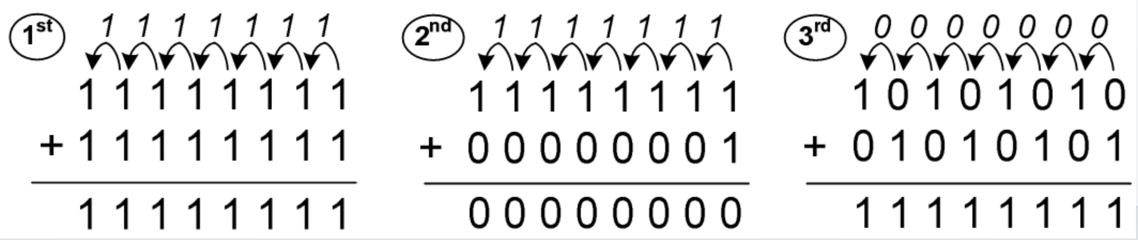
Fig. 2. Structural Similarity (SSIM) Measurement System [6]

Delay Oriented Error Analysis

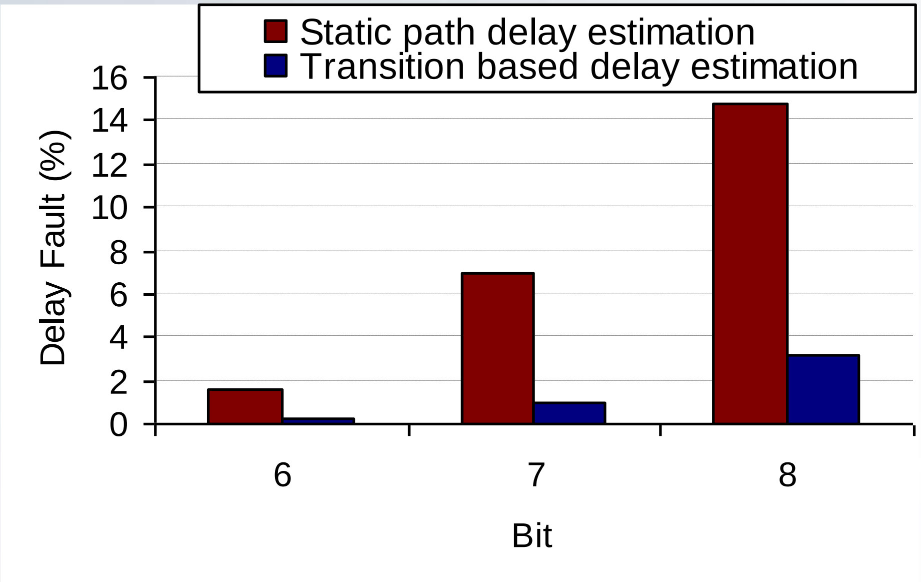
- Input dependent delay estimation is required for accurate delay estimation
- Static path delay based estimation
  - Based on only current input and carry propagation
- Transition delay based estimation
  - Based on previous and current input and transition

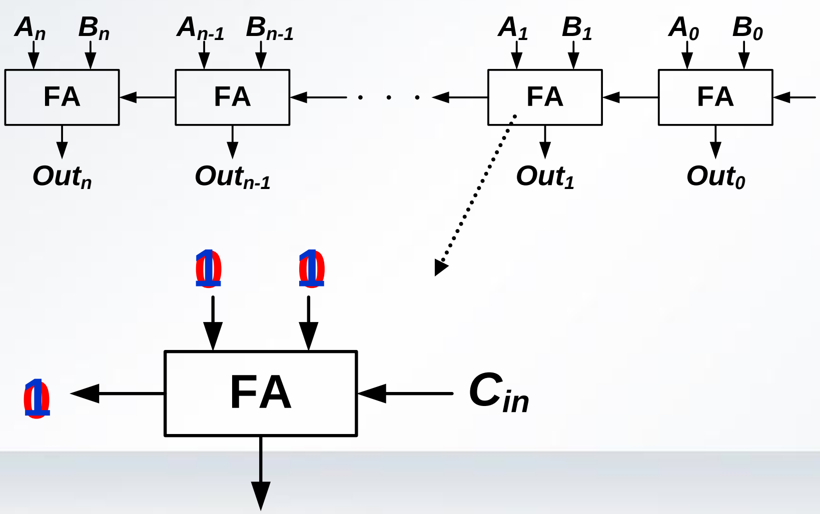


Ripple carry adder



Delay Oriented Error Analysis





- Comparison between two delay estimation methods

- The behavior of a full adder
  - consecutive *10* or *01* → long delay propagation path

Input analysis

- Characterizing the relationship between input and error
  - Three cases to cause a long delay propagation path

Case 1

$$
\begin{array}{r}
00000001\ (+1) \\
+\ 11111110\ (-2) \\
\hline
11111111\ (-1)
\end{array}
$$

λ

Case 2

$$
\begin{array}{r}
10111111\ (-65) \\
+\ 01000011\ (+67) \\
\hline
00000010\ (+2)
\end{array}
$$

λ

Case 3

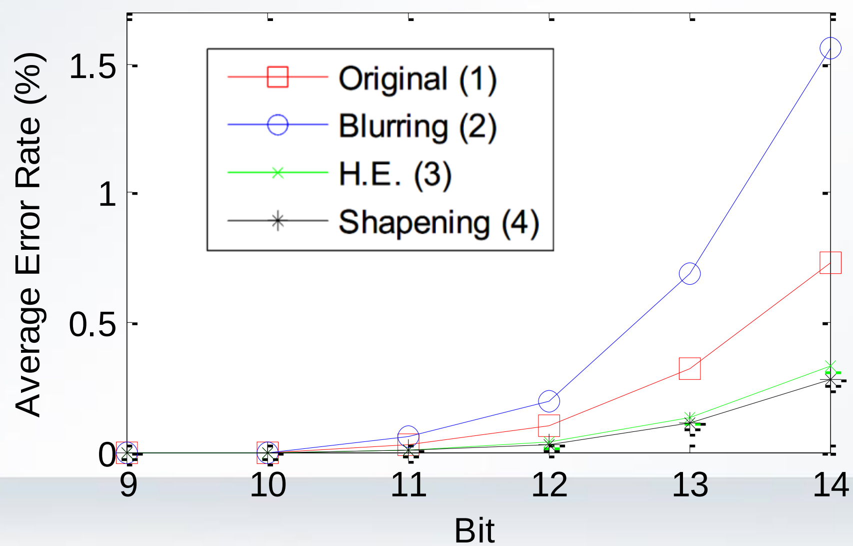
$$
\begin{array}{r}
00000001\ (+1) \\
+\ 01111111\ (+127) \\
\hline
10000000\ (+128)
\end{array}
$$

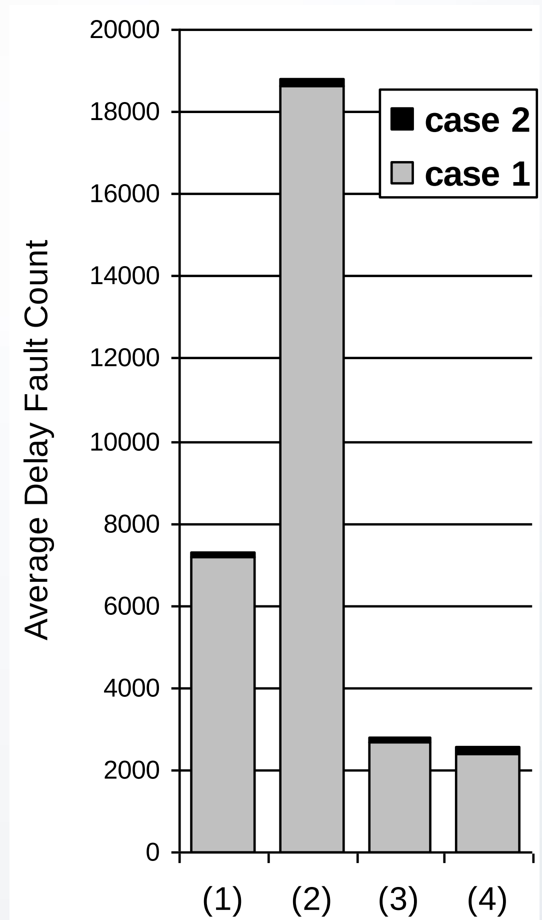
λ

1. sign(*A*) ≠ sign(*B*) & mag. of *A* and *B* is small
2. sign(A) ≠ sign(B) & mag. of A and B is large and similar
3. sign(A) = sign(B) & mag. diff. of A & B is very large

Input analysis

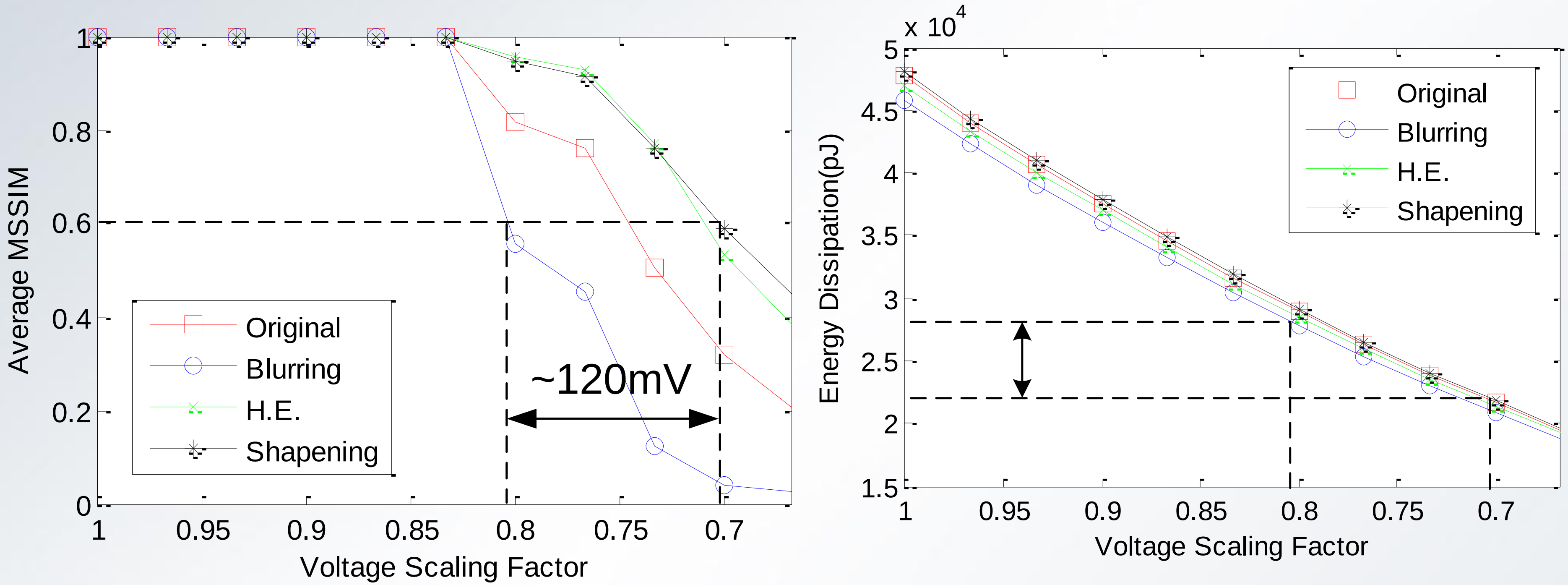
- Case 1 is dominant
- Input image analysis
  - Image type 1: blurred
  - Image type 2: sharpened, histogram equalized





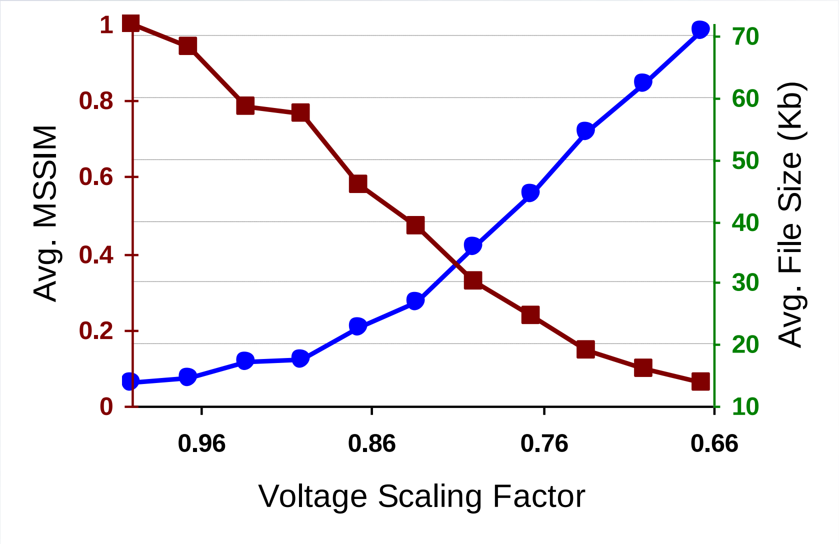
Georgia Institute of Technology

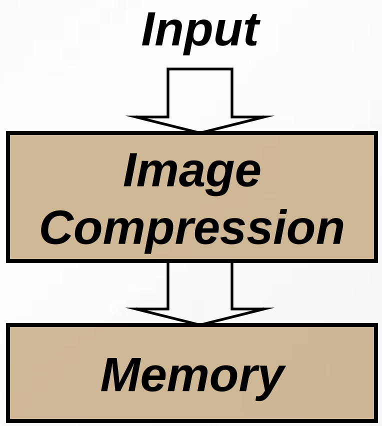
Experimental Results and Discussion



- Difference in voltage scalability (energy savings) among different image types for a given quality requirement

System-level Analysis



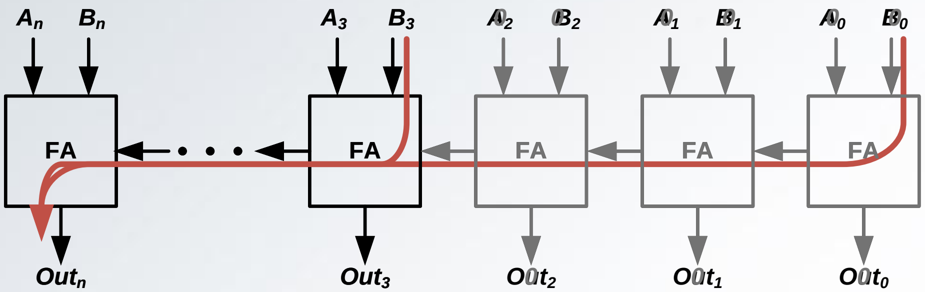


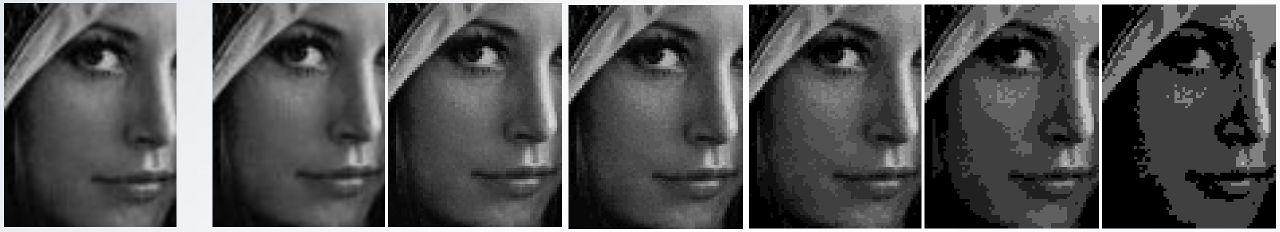
- Impact of aggressive voltage scaling on image compression
  - Degradation in output quality
  - Reduction in compression ratio → Increase output file size
    - Run-length of zero
    - Average codeword length

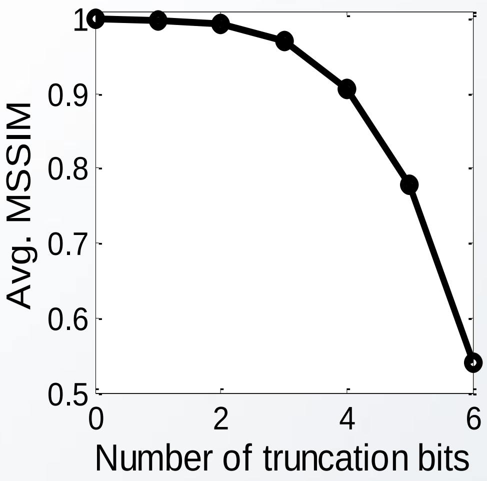
➔ Increase in energy consumption in other subsystems

e.g., Memory for storing the output images

Pixel and Coefficient Truncation

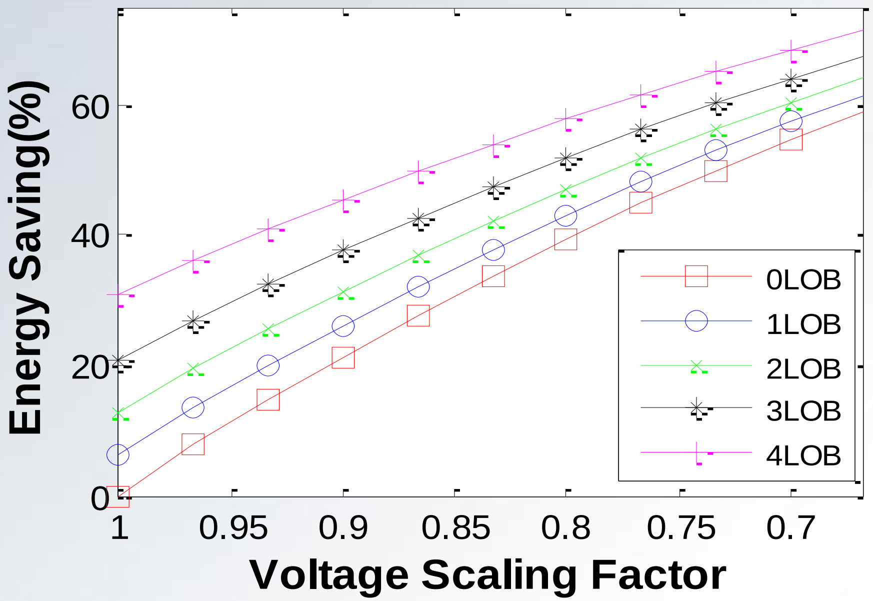




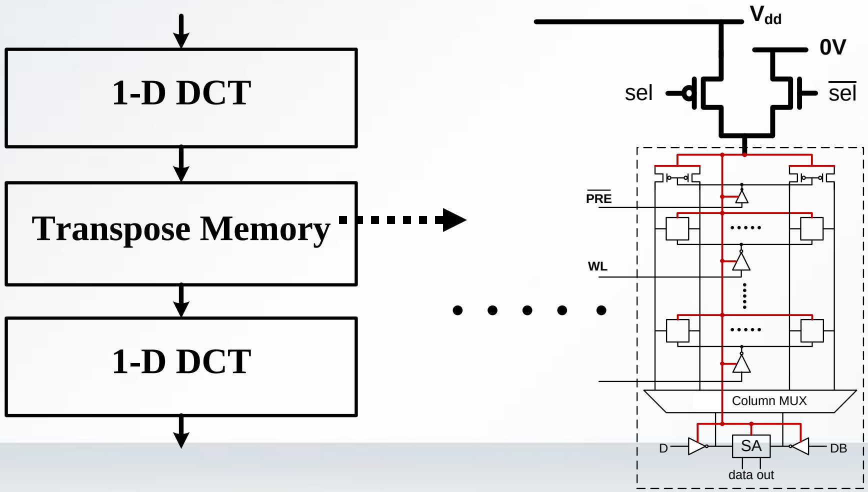


- Directly reduces delay → Error rate reduction
- Low order bit truncation→ Minimize quality degradation

Pixel and Coefficient Truncation

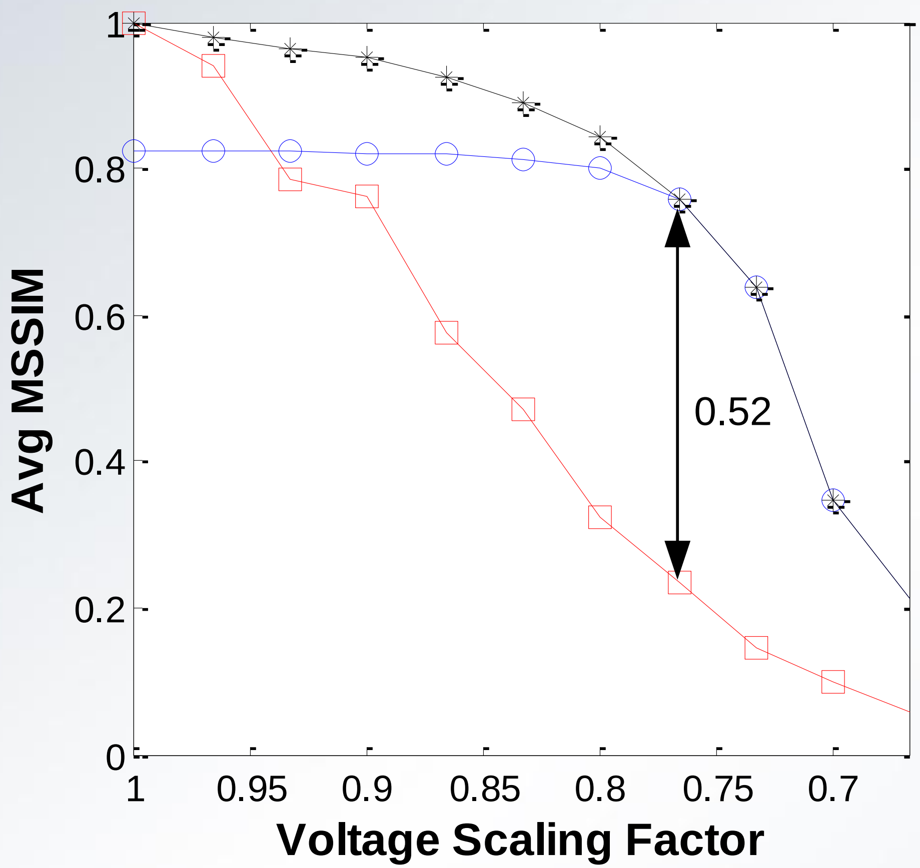


- Energy savings
  - Switching activity reduction
  - Supply voltage reduction

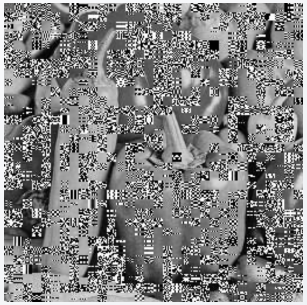


- Disabling the memory cells
  - → Additional energy savings
  - → Trivial implementation cost (6% area increase)

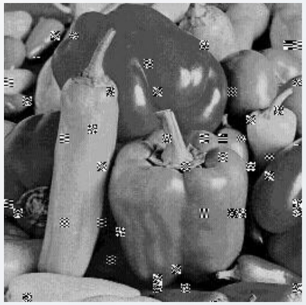
Results



Original

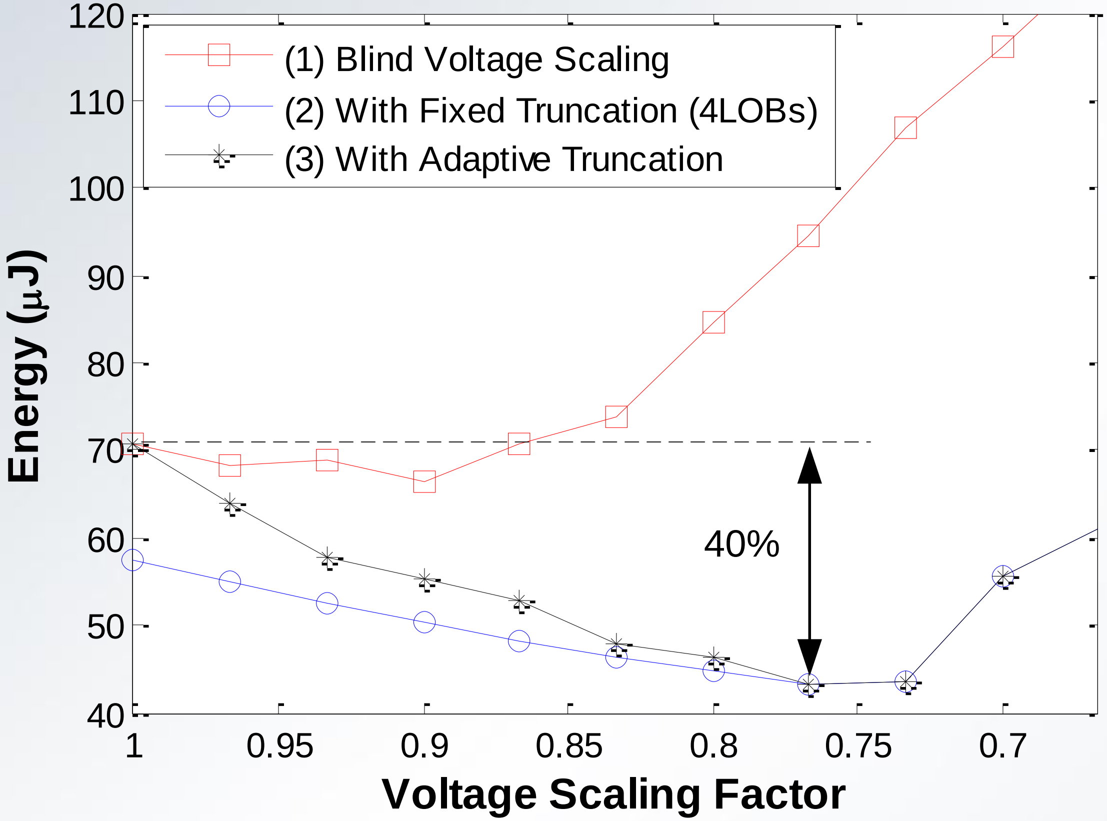


(1)

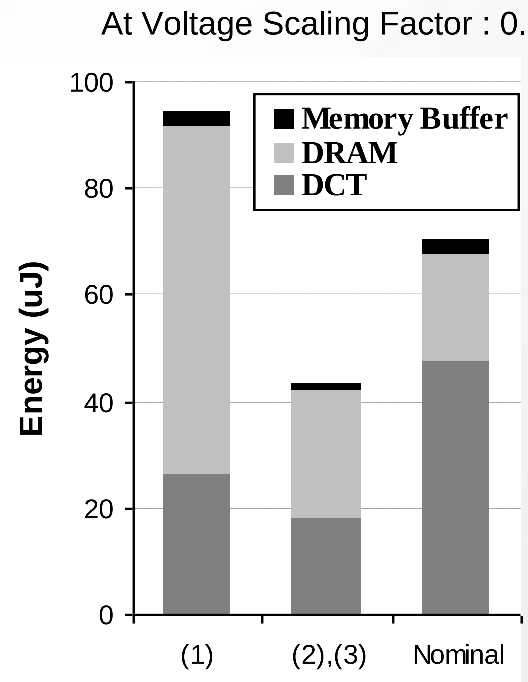


(2), (3)

Results



120
110
100
90
80
70
60
50
40
Energy (μJ)
(1) Blind Voltage Scaling
(2) With Fixed Truncation (4LOBs)
(3) With Adaptive Truncation
40%
1
0.95
0.9
0.85
0.8
0.75
0.7
Voltage Scaling Factor



At Voltage Scaling Factor : 0.77
100
80
60
40
20
0
Energy (uJ)
Memory Buffer
DRAM
DCT
(1)
(2),(3)
Nominal

Conclusions

- Trade-offs need to be evaluated at the system level.
- Subthreshold error modeling is a sequential problem, not a combinational one.
- Algorithms and platforms can take advantage of perceptual properties to optimize other design criteria.

Georgia Institute of Technology