4C-3: Streaming Accuracy: Characterizing Early Termination in Stochastic Computing

ASP-DAC 2022

Hsuan (Julie) Hsiao[1], Joshua San Miguel[2], Jason Anderson[1]

[1]University of Toronto, [2]University of Wisconsin-Madison

The Edward S. Rogers Sr. Department of Electrical & Computer Engineering
UNIVERSITY OF TORONTO

Executive Summary

Problem:

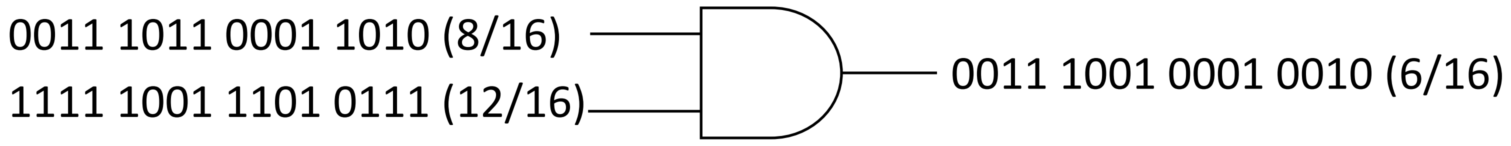
- Early termination is important for stochastic computing, but current methods for characterizing early termination have fundamental limitations

This work:

- We propose streaming accuracy, a metric for how good a bitstream is for early termination
- We introduce the concept of a streaming-accurate bitstream, which achieves the lowest possible error at all partial bitstream lengths
- We propose a design for a bitstream generator that generates streaming-accurate bitstreams

Stochastic Computing (SC)

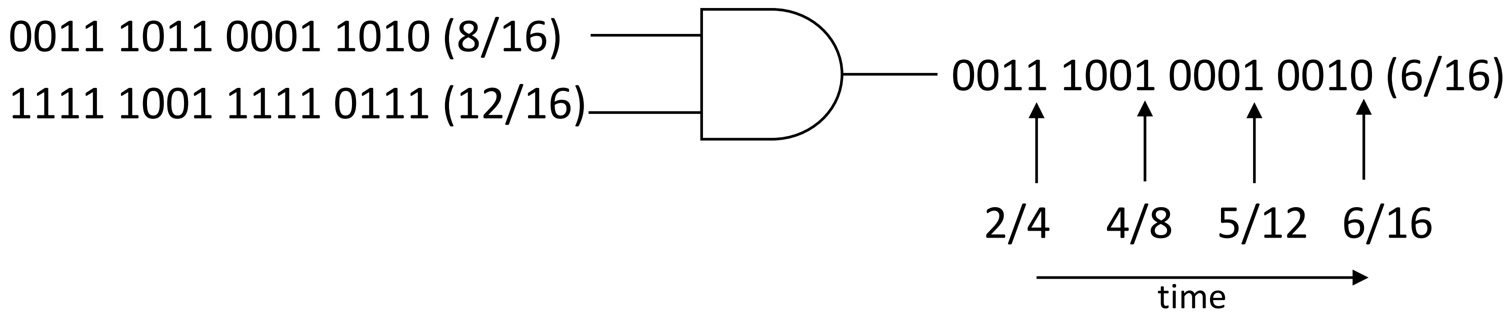
- Approximate computing paradigm
- Values are represented by the probability that a bit is set (e.g. fraction of 1s in a bitstream)
- Very small compute units
  - Example: multiplication → AND gate



- Potential for low power/area hardware
  - Example: image processing, LDPC decoding, neural networks

Early Termination in SC

- Value of a bitstream gets progressively more accurate/precise



- Terminate computation early if result is good enough
- Trade off latency to provide similar accuracy as binary encoding
  - Scales exponentially
- Early termination enhances the feasibility of SC circuits
  - Accuracy-energy tradeoff

Important to be able to measure early termination

Prior Work in Quantifying Early Termination

- Progressive Precision [1]
  - A bitstream is k-PP if the bit error of its initial partial bitstream of length 2^i is at most k for all i

[1] A. Alaghi and J. P. Hayes, "Fast and accurate computation using stochastic circuits," *2014 DATE*, pp. 1-4.

Prior Work in Quantifying Early Termination

- Progressive Precision [1]
  - A bitstream is *k*-PP if the bit error of its initial partial bitstream of length 2^i is at most *k* for all *i*

0111 1111 1111 0000 = $\frac{\mathbf{11}}{\mathbf{16}}$ = 0.6875 desired = $\frac{\mathbf{10}}{\mathbf{16}}$ = 0.625

[1] A. Alaghi and J. P. Hayes, "Fast and accurate computation using stochastic circuits," *2014 DATE*, pp. 1-4.

Prior Work in Quantifying Early Termination

- Progressive Precision [1]
  - A bitstream is k-PP if the bit error of its initial partial bitstream of length 2^i is at most k for all i

$$\boxed{\boxed{\boxed{\boxed{01}11}\ 1111}\ 1111\ 0000} = \frac{11}{16} = 0.6875 \qquad \text{desired} = \frac{10}{16} = 0.625$$

$$2\times\left|\frac{1}{2}-\frac{10}{16}\right| = 0.25 \quad 4\times\left|\frac{3}{4}-\frac{10}{16}\right| = 0.5 \quad 8\times\left|\frac{7}{8}-\frac{10}{16}\right| = 2 \quad 16\times\left|\frac{11}{16}-\frac{10}{16}\right| = 1$$

→ bitstream is 2-PP

[1] A. Alaghi and J. P. Hayes, "Fast and accurate computation using stochastic circuits," *2014 DATE*, pp. 1-4.

Prior Work in Quantifying Early Termination

- Progressive Precision [1]
  - A bitstream is k-PP if the bit error of its initial partial bitstream of length 2^i is at most k for all i

$\boxed{\boxed{\boxed{\boxed{01}11}\ 1111}\ 1111\ 0000} = \frac{11}{16} = 0.6875$ desired $= \frac{10}{16} = 0.625$

$2\times\left|\frac{1}{2}-\frac{10}{16}\right| = 0.25$ $4\times\left|\frac{3}{4}-\frac{10}{16}\right| = 0.5$ $8\times\left|\frac{7}{8}-\frac{10}{16}\right| = 2$ $16\times\left|\frac{11}{16}-\frac{10}{16}\right| = 1$ → bitstream is 2-PP

- only evaluate powers-of-2 lengths
- does not provide information on how much better k can get

[1] A. Alaghi and J. P. Hayes, "Fast and accurate computation using stochastic circuits," *2014 DATE*, pp. 1-4.

Prior Work in Quantifying Early Termination

- Normalized Stability [2]
  - A bitstream is *stable* if error never exceeds an error budget from now on

5% error budget 1010101010 1010101010101010101010 68.75% stable

stable (i.e., error always within budget)

[2] D. Wu, R. Yin and J. San Miguel, "Normalized stability: a cross-level design metric for early termination in stochastic computing," *2021 ASPDAC*.

Prior Work in Quantifying Early Termination

- Normalized Stability [2]
  - A bitstream is *stable* if error never exceeds an error budget from now on

5% error budget 1010101010 1010101010101010101010 68.75% stable

stable (i.e., error always within budget)

A 1010101010 1010101010101010101010 1.00

B 101010101010101010111 00010101010 0.50

C 101010101010101010101010111 00010 0.25

find maximum achievable stability
and normalize to it

[2] D. Wu, R. Yin and J. San Miguel, "Normalized stability: a cross-level design metric for early termination in stochastic computing," *2021 ASPDAC*.

Prior Work in Quantifying Early Termination

- Normalized Stability [2]
  - A bitstream is *stable* if error never exceeds an error budget from now on

5% error budget 1010101010 1010101010101010101010 68.75% stable

stable (i.e., error always within budget)

requires a user-specified error threshold

A 1010101010 1010101010101010101010 1.00

B 101010101010101010111 00010101010 0.50

C 101010101010101010101010111 00010 0.25

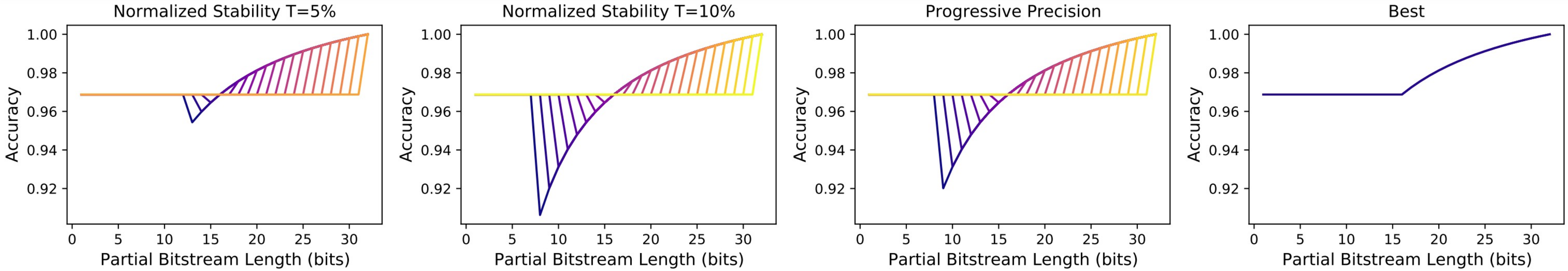
find maximum achievable stability
and normalize to it

uses an estimate

[2] D. Wu, R. Yin and J. San Miguel, "Normalized stability: a cross-level design metric for early termination in stochastic computing," *2021 ASPDAC*.

Limitations of Prior Work

- Select the best bitstream for early termination from all possible versions of $\frac{1}{32}$
  - Apply normalized stability and progressive precision to all versions of $\frac{1}{32}$
  - Plot the accuracy vs time for bitstream(s) that scored the highest on the metric



20 bitstreams

25 bitstreams

24 bitstreams

Streaming Accuracy

- Reflect the accuracy of a bitstream at any arbitrary early termination point

Streaming Accuracy

- Reflect the accuracy of a bitstream at any arbitrary early termination point

partial bitstream value

desired bitstream value

- Raw streaming accuracy $\Phi = 1 - \frac{\sum_{i=1}^{L} |P_{X_i} - P_X|}{L}$

bitstream length

- Best bitstream for early termination will have highest Φ value

Streaming Accuracy

- Reflect the accuracy of a bitstream at any arbitrary early termination point

partial bitstream value

desired bitstream value

- Raw streaming accuracy $\Phi = 1 - \frac{\sum_{i=1}^{L} |P_{X_i} - P_X|}{L}$

bitstream length

- Best bitstream for early termination will have highest Φ value

$$\text{Streaming accuracy}_{value-dependent} = \frac{\Phi}{\Phi_{best}}$$

for comparing bitstreams of the same value

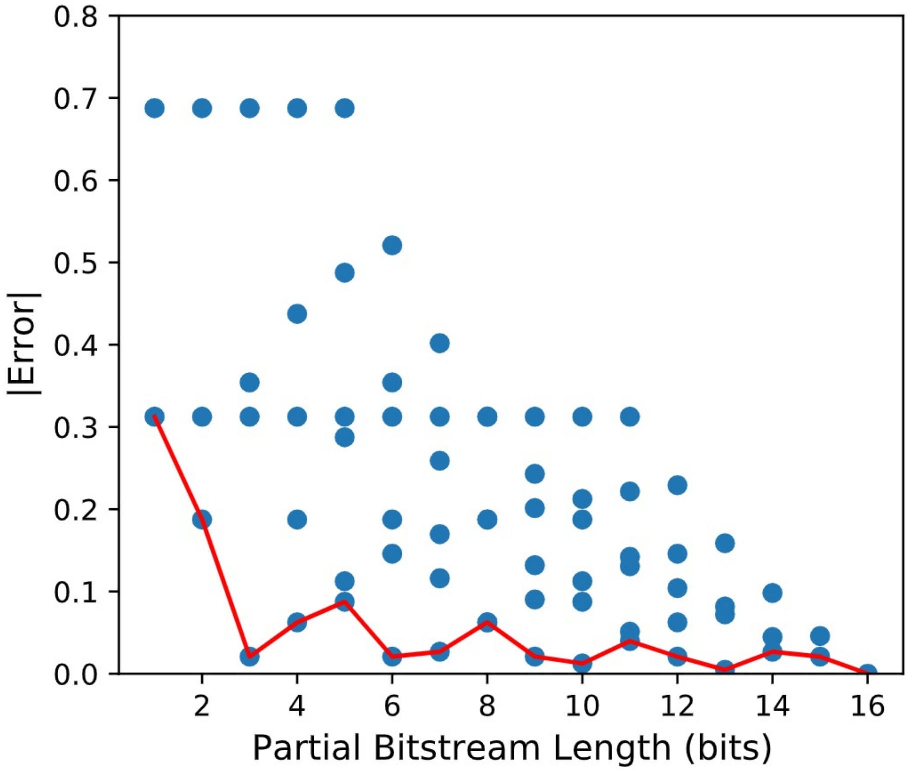
$$\text{Streaming accuracy}_{value-independent} = \frac{\Phi - \Phi_{worst}}{\Phi_{best} - \Phi_{worst}}$$

for comparing bitstreams of different values

Streaming Accuracy

- How do we find Φ_{best}
  - Observe that each successive partial bitstream lengths require monotonically increasing number of 1s by at most 1

All possible errors for $\frac{5}{16} = 0.3215$



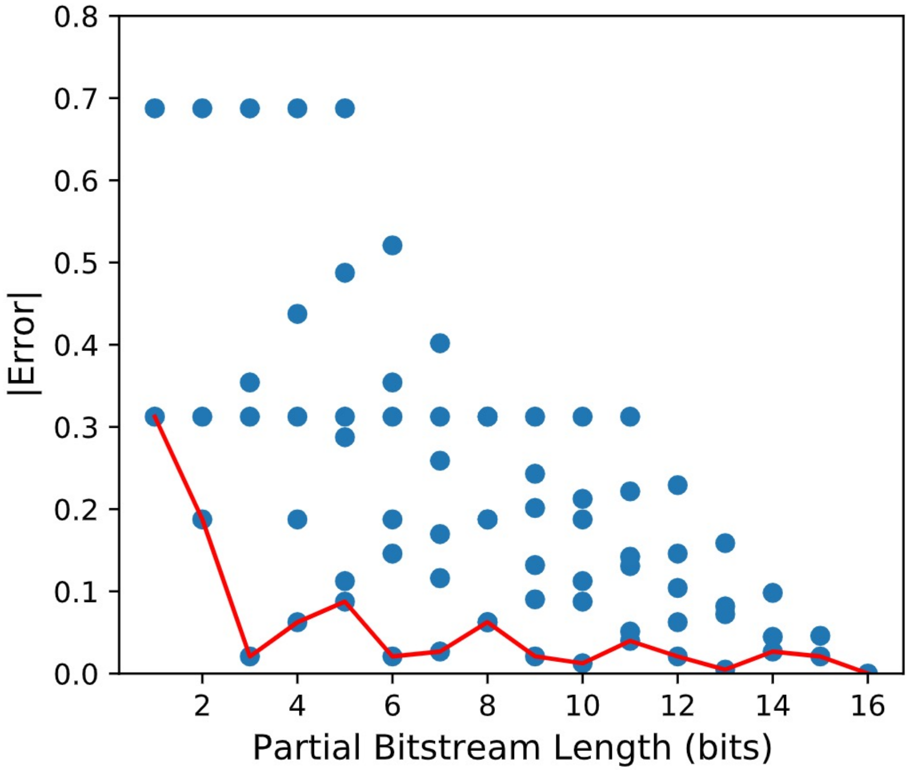
partial values = $\frac{0}{1}\ \frac{1}{2}\ \frac{1}{3}\ \frac{1}{4}\ \frac{2}{5}\ \frac{2}{6}\ \frac{2}{7}\ \frac{3}{8}\ \frac{3}{9}\ \frac{3}{10}\ \frac{3}{11}\ \frac{4}{12}\ \frac{4}{13}\ \frac{4}{14}\ \frac{5}{15}\ \frac{5}{16}$

Streaming Accuracy

- How do we find Φ_{best}
  - Observe that each successive partial bitstream lengths require monotonically increasing number of 1s by at most 1

    → There always exists a bitstream that is optimal for early termination at arbitrary points since a partial bitstream of length i+1 can be constructed from a partial bitstream of length i

All possible errors for $\frac{5}{16} = 0.3215$



partial values = $\frac{0}{1}\ \frac{1}{2}\ \frac{1}{3}\ \frac{1}{4}\ \frac{2}{5}\ \frac{2}{6}\ \frac{2}{7}\ \frac{3}{8}\ \frac{3}{9}\ \frac{3}{10}\ \frac{3}{11}\ \frac{4}{12}\ \frac{4}{13}\ \frac{4}{14}\ \frac{5}{15}\ \frac{5}{16}$

→ best bitstream = 0100 1001 0001 0010

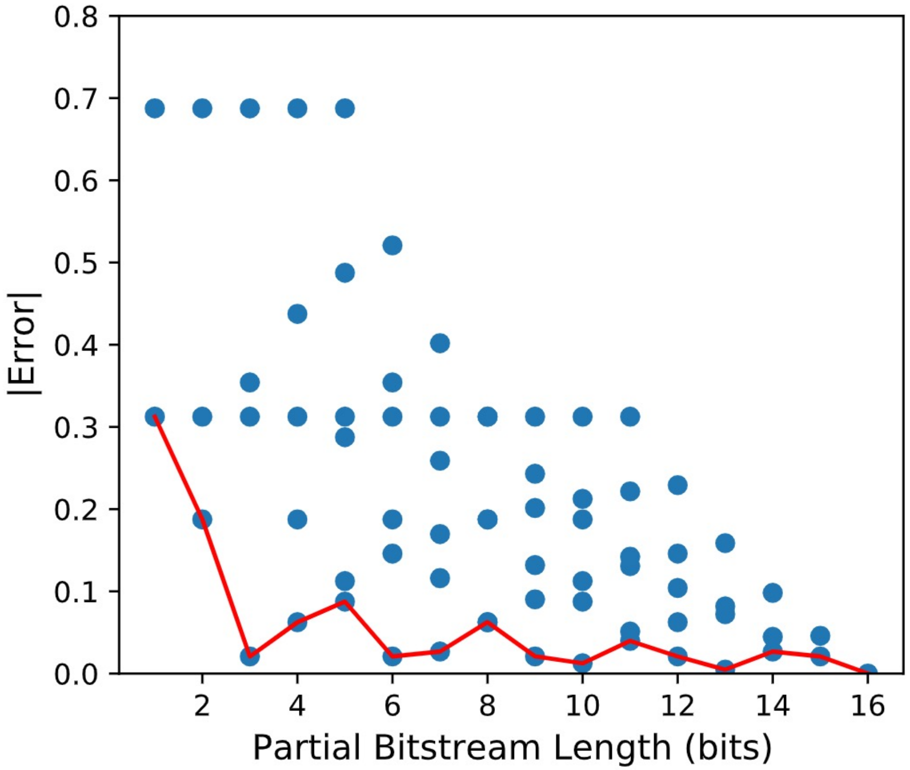
Streaming Accuracy

- How do we find Φ_{best}
  - Observe that each successive partial bitstream lengths require monotonically increasing number of 1s by at most 1

    → There always exists a bitstream that is optimal for early termination at arbitrary points since a partial bitstream of length i+1 can be constructed from a partial bitstream of length i

- Greedy approach:
  - Evaluate whether introducing 0 or 1 will minimize error at each additional bit
  - Call this a **streaming-accurate bitstream**

All possible errors for $\frac{5}{16} = 0.3215$



partial values = $\frac{0}{1}\ \frac{1}{2}\ \frac{1}{3}\ \frac{1}{4}\ \frac{2}{5}\ \frac{2}{6}\ \frac{2}{7}\ \frac{3}{8}\ \frac{3}{9}\ \frac{3}{10}\ \frac{3}{11}\ \frac{4}{12}\ \frac{4}{13}\ \frac{4}{14}\ \frac{5}{15}\ \frac{5}{16}$

→ best bitstream = 0100 1001 0001 0010

Streaming Accuracy

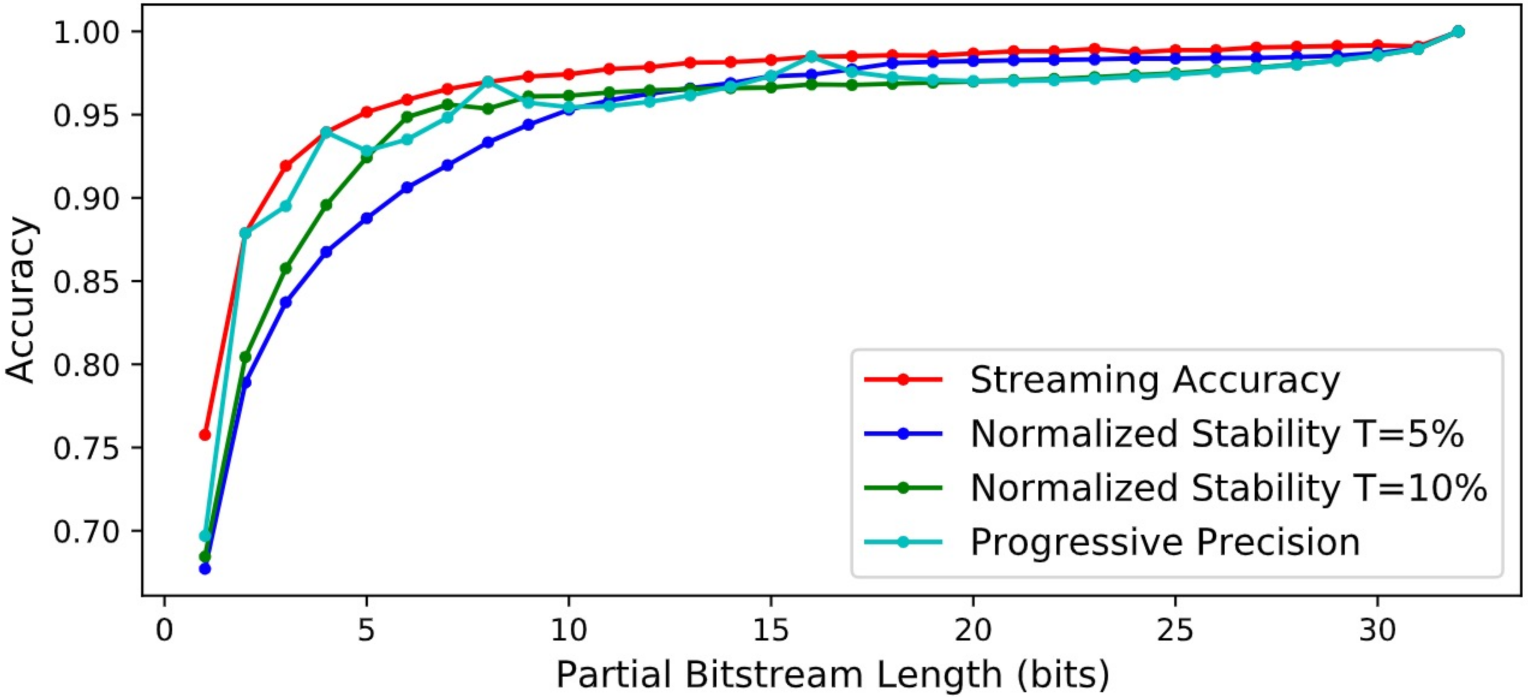
- How do we find Φ_{worst}
  - Bitstreams with value < 0.5 ← worst when all 1s clumped at the beginning
  - Bitstreams with value > 0.5 ← worst when all 1s clumped at the end

$$\text{Streaming accuracy}_{value-dependent} = \frac{\Phi}{\Phi_{best}}$$

$$\text{Streaming accuracy}_{value-independent} = \frac{\Phi - \Phi_{worst}}{\Phi_{best} - \Phi_{worst}}$$

Comparison Against Prior Metrics

- For each representable value, use each metric to select the best bitstream for early-termination
  - Average across all values
  - If metric selects multiple, average across all selections



Characterization

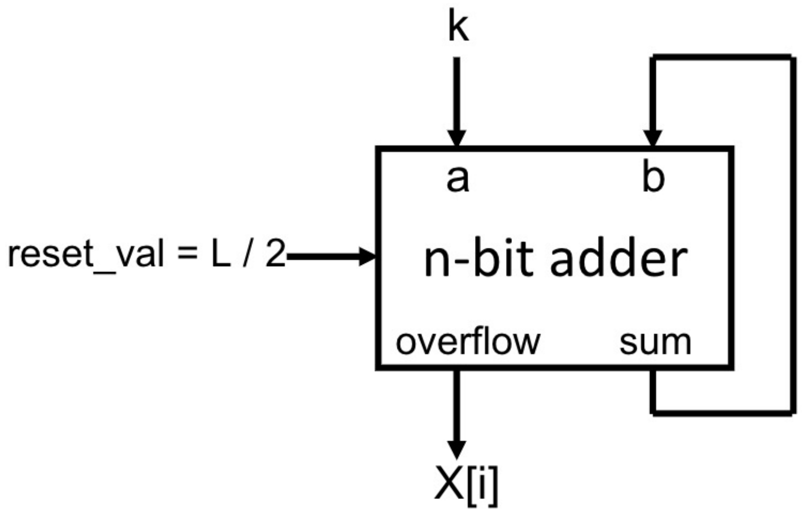
- Streaming accuracy allows designers to perform valuable characterizations

- Characterize different SC components (and different designs of a component)
  - See if a unit enhances or hinders the overall circuit's ability to early terminate
  - Examples of some arithmetic units in paper

- Explore relation between other bitstream properties
  - See if properties can be jointly optimized
  - Example of bitstream independence vs. early termination in paper

Streaming-Accurate (SA) Bitstream Generator

- We propose a design for a bitstream generator that generates streaming-accurate bitstreams
- Area and power estimates @ 500 MHz:
  - Implemented Verilog RTL for L=64
  - Synopsys design compiler with NanGate FreePDK45 standard-cell library



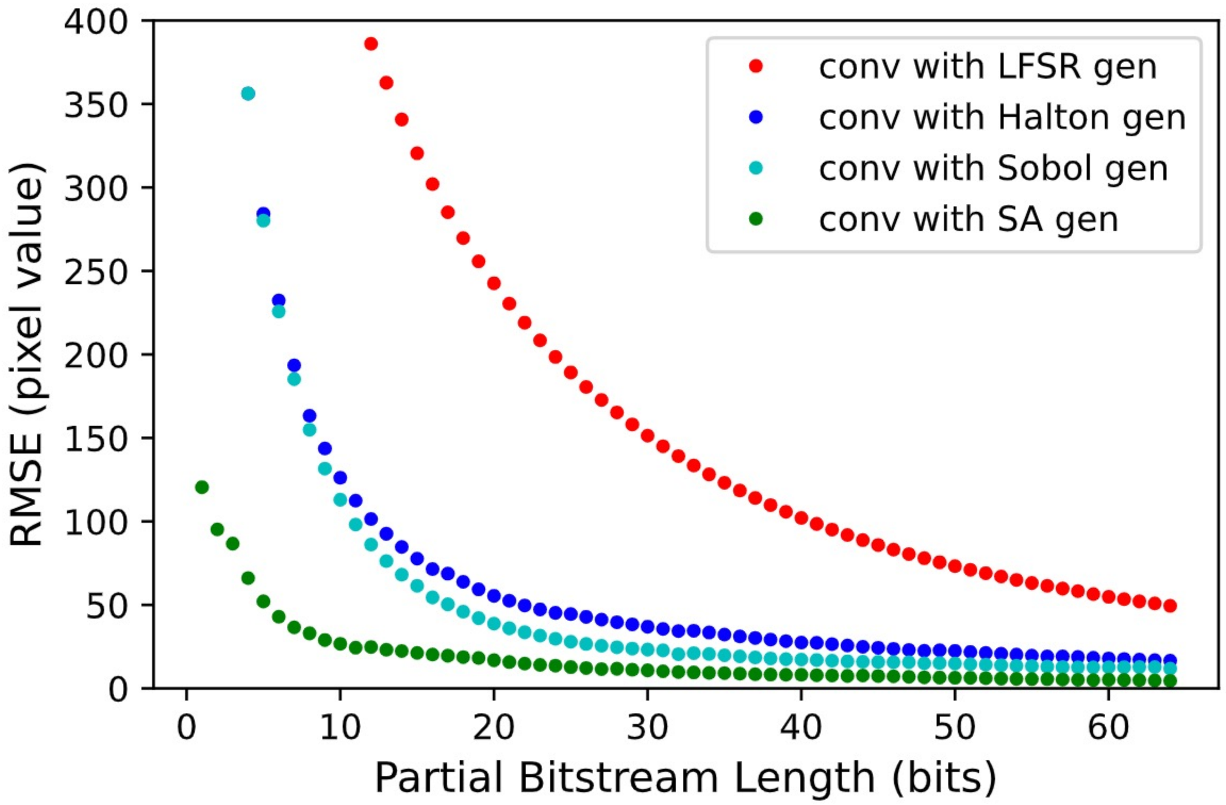
Generator	Area (μm^2)	Power (uW)
LFSR	75.278	26.851
Halton (base 2)	69.16	39.73
Halton (base 3)	245.252	140.796
Sobol	208.81	75.546
SA	69.16	31.306

Accuracy Evaluation

- Evaluate the impact of using streaming-accurate bitstreams in computation

Gaussian blur 2D convolution:
256 x 256 grayscale image, 3 x 3 weight filter,
6 images

- For each input pixel, use the bitstream produced by each generator
- Error (RMSE) relative to the floating-point version



Conclusion

- Proposed streaming accuracy, a metric that measures how close a bitstream resembles its streaming-accurate form
  - Showed how it fills the gaps left by existing metrics
- Proposed a new bitstream generator that is guaranteed to produce bitstreams in their streaming-accurate form
  - Showed it achieves higher accuracy compared to other popular bitstream generators

Thank you!

Questions: live Q&A at session 4C or email julie.hsiao@mail.utoronto.ca