Time-Encoded Values for Highly Efficient Stochastic Circuits

IEEE Transaction on VLSI 2017

M. Hassan Najafi, Shiva Jamali-Zavareh, David J. Lilja, Marc D. Riedel, Kia Bazargan, and Ramesh Harjani

Najaf011@umn.edu

ISCAS 2017, BALTIMORE, MD, USA
Overview

• Introduction
  – Stochastic Computing, advantages, weaknesses
  – Representation of stochastic numbers

• Stochastic Number Generation
  – Conventional approach: LFSR-based
  – Proposed approach: PWM-based

• Stochastic Operations with PWM Signals
  – Multiplication, Scaled Addition, Absolute value subtraction
  – Multi-level circuits with PWM signals

• Experimental results
  – Hardware cost, operation time, energy, performance comparison
    • Conventional binary, Prior Stochastic, PWM-based Stochastic

• Sources of computational error

• Conclusions
Introduction

• Stochastic Computation
  – A re-emerging computing paradigm: introduced in 1969
  – Logical computation on random bit streams

  – Value: probability of obtaining a one versus a zero
    • Unipolar \([0, 1]\) positive
      – Each bit has probability \(X\) of being 1
    • Bipolar \([-1, 1]\) positive, negative
      – Each bit has probability \((X+1)/2\) of being one

    • 000111, 1010, 110010 = 0.5 (unipolar), 0.0 (bipolar)

  – Variable length bit streams
**Key Advantages**

- Simple hardware for complex operations
  - Multiplication: **AND** (unipolar), **XNOR** (Bipolar)
  - Scaled Addition: **MUX**

- Gracefully tolerates noise
  - Redundant representation provides error tolerance
  - Stochastic: 0010000011000000 (3/16) → 4/16 = 0.25
  - Binary: 0.0011 = 0.1875 → 0.1011 = 0.68

- Skew tolerance
  - Polysynchronous stochastic circuits [Najafi et al, ASP-DAC, 2016]
Introduction

• Weaknesses
  – Stochastic Number Generators (SNGs) are costly
    • 80% or more of the overall hardware cost
    • and power consumption
  – High accuracy ~ Long stochastic streams

  – Energy = Power x Time

  – Accurate SC -> High energy consumption
    • Much slower
    • More energy consumption
      – than conventional binary design
Introduction

• Representation of Stochastic Numbers
  – Digital
    • Probability of obtaining a one versus a zero

  ![AND gate diagram with time-encoded inputs and output]

  \[0,1,1,0,1,0,1,1,1 \text{ (6/10)}\]
  \[1,0,1,1,0,0,1,0,1,1 \text{ (5/10)}\]
  \[0,0,1,0,0,1,0,1,1 \text{ (3/10)}\]

  ![Time-encoded signals]

  \[\text{In1} \quad \text{2ns} \quad \text{3ns} \quad \text{1ns} \quad 6/10 = 0.6\]
  \[\text{In2} \quad \text{2ns} \quad \text{3ns} \quad 5/10 = 0.5\]
  \[\text{AND\_OUT} \quad \text{1ns} \quad \text{2ns} \quad 3/10 = 0.3\]

  – Analog
    • Encoding the value as the fraction of time the signal is high
Stochastic Number Generation

- **Conventional approach**
  - Using random or pseudo-random constructs
    - e.g. LFSR

  ![Diagram of Random Number Generator and Comparator]

- **Proposed approach**
  - Pulse Width Modulation
  - Analog periodic pulses signals as the stochastic number
Stochastic Number Generation

- **PWM signals as the stochastic number**
  - Defined by a **frequency** and a **duty cycle**.
  - **Duty cycle** describe the amount of high time

PWM signals with different duty cycles

(a) 20% duty cycle

(b) 50% duty cycle

(c) 80% duty cycle

PWM signals with different duty cycles
Stochastic Operations

- Stochastic Operations based on their inputs
  - **Independent** or uncorrelated. 110101, 101100
    - **AND**: Multiplication
  - **Correlated**: 111100, 110000
    - **XOR**: Absolute value subtraction $|X_1 - X_2| : 001100$
    - **AND**: Minimum: 110000
    - **OR**: Maximum: 111100
  - **Insensitive** to correlation
    - **MUX**: Scaled addition/subtraction
Stochastic Operations with PWM signals

- **Multiplication**
  - Connecting two PWM signals with same duty cycle and same frequency to an AND gate will not work!
    - Output signal = Input signals
  - Choosing same frequency leads to poor results

- **Key:**
  - **Choosing different frequencies.**
    - Results in a new phase between the signals in each repetition
Stochastic Operations with PWM signals

- **Multiplication**
  - Assume two input values
    - \(X = \frac{3}{5}\) (duty cycle of 60%): 11100 (Length=5)
    - \(Y = \frac{1}{2}\) (duty cycle of 50%): 1100 (Length=4)
  - We multiply \(X\) and \(Y\) with AND gate for 20 cycles: 6/20

  \[
  \begin{align*}
  X &= 11100111001110011100 \\
  Y &= 11001100110011001100 \\
  X \cdot Y &= 11000100000010001100
  \end{align*}
  \]

  - The multiplication result is always correct if
    - one chooses stream lengths that are relatively prime and let them run up to the common multiple.
Stochastic Operations with PWM signals

- **Multiplication**
  - Relatively prime length rule for digital bit streams
  - Inharmonic PWM signals

- **Verifying the argument**
  - Simulated multiplication on a large set of random inputs
  - **Fixed** the period of the first PWM signal at 20ns
  - **Varying** the period of the second from 1 to 20ns
  - Run for 1000ns
Stochastic Operations with PWM signals

- An example of multiplying two PWM signals using an AND gate.
  - IN1 represents 0.5 (50% duty cycle) with a period of 20 ns
  - IN2 represents 0.6 (60% duty cycle) with a period of 13 ns

- The output after 260 ns represents 0.30 (78 ns/260 ns)
Stochastic Operations with PWM signals

The average error rate of multiplying 1000 pairs of random numbers when varying the operation time.

The period of the PWM signals: 20ns and 13ns

- Best running time: LCM or multiples of the LCM of period of the inputs

The average error rate for multiplying 1000 pairs of random numbers

The period of the PWM signal: relatively prime integers in the interval [2, 20].

- The larger the LCM => the higher the accuracy
• Scaled Addition

Inharmonic input and select signal for MUX

The average error rate of performing scaled addition on 1000 pairs of random numbers

• Large LCMs are not necessarily required to produce accurate results.

• Optimal choice: an “even” value for the period of the select input and an “odd” value for the period of the main inputs.

• The operation should run for the LCM of the periods.
Stochastic Operations with PWM signals

- **Absolute Value Subtraction**
  - requires highly correlated inputs

- **High correlation in PWM signals**
  1) choosing the **same frequency** for the input signals
  2) having **maximum overlap** between the high parts

For these operations the period of the output signal, and, thus, the operation time, equals the period of the input signals.

- **Advantage:** accurate result after running for only one period
- **Limitation:** difficult to provide correlation for higher levels
Multi-level circuits with PWM signals

- The output of each level can be used as the input of the next level even though the output is not a PWM signal.
  - Output is still a periodic signal
  - Output of (a) and (b) has a period of $P_1 \times P_2 \times P_3 \times P_4$

- The important trade-off is to select small or large periods
  - Shorter operation time vs. higher accuracy
Experimental Results

- Validated the idea using:
  - **Robert’s** cross edge detection, **Gamma** correction function

Inputs period = 0.51ns
Select period = 0.34ns
Operation time = 1.02ns

\[ X \text{ period} = 0.60\text{ns} \]
\[ b \text{ period} = 0.90\text{ns} \]
Operation time = 1.8ns

**Area, delay, power and energy** comparison of the implemented circuits

<table>
<thead>
<tr>
<th>Circuit</th>
<th>Approach</th>
<th>Area ((\mu m^2))</th>
<th>Delay (ns)</th>
<th>Power ((\mu W)) @max freq.</th>
<th>Energy (pJ)</th>
<th>Area×Delay ((\mu m^2 \times \mu s))</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Core</td>
<td>SNG</td>
<td>Output Circt.</td>
<td>Total</td>
<td>Delay</td>
</tr>
<tr>
<td>Robert</td>
<td>Conventional binary</td>
<td>1626</td>
<td>-</td>
<td>-</td>
<td>1626</td>
<td>0.78</td>
</tr>
<tr>
<td></td>
<td>Stochastic-Prior</td>
<td>22</td>
<td>739</td>
<td>199</td>
<td>960</td>
<td>87.04</td>
</tr>
<tr>
<td></td>
<td>Stochastic-PWM</td>
<td>22</td>
<td>464</td>
<td>110</td>
<td>596</td>
<td>1.02</td>
</tr>
<tr>
<td>Gamma</td>
<td>Conventional binary</td>
<td>1980</td>
<td>-</td>
<td>-</td>
<td>1980</td>
<td>1.03</td>
</tr>
<tr>
<td></td>
<td>Stochastic-Prior</td>
<td>76</td>
<td>982</td>
<td>199</td>
<td>1257</td>
<td>153.6</td>
</tr>
<tr>
<td></td>
<td>Stochastic-PWM</td>
<td>76</td>
<td>678</td>
<td>110</td>
<td>864</td>
<td>1.8</td>
</tr>
</tbody>
</table>
Experimental Results

- PWM signal generator

The area cost of the PWM generator (when period=2ns) is roughly as expensive as the cost of the conventional SNG with 8-bit LFSR.

Decreasing PWM period = Increases the error rate but Lowers implementation cost

- The area cost of the PWM generator (when period=2ns) is roughly as expensive as the cost of the conventional SNG with 8-bit LFSR.
Experimental Results

- Performance comparison

Average Error Rate of processing sample images with the implemented circuits.
Sources of Computational Errors

- **Five primary sources of errors** in performing stochastic operations on PWM signals

  - \( E_G = \) Error in generating PWM signals
    - The difference of the **expected** and the **generated** duty cycle
    - Achieving the desired frequency is also not always feasible
      - Limitations in controlling the period of the PWM generator
    - In our simulations, \( E_G < 0.4\% \).

  - \( E_S = \) Error due to skew between input signals
    - Some operations need perfectly synchronized PWM signals
    - On-chip variations, other noise sources affecting clock generators result in deviation from
      - expected period, phase shift, slew rate
Sources of Computational Errors

- $E_M = \text{Error in measuring the output signals}$
  - Analog integrator
    - Measuring the fraction of time the output signal is high
  - Longer rise and fall times
  - Imperfect measurement of the high and low voltages
    - Result in inaccuracies in measuring the correct output
  - In our simulations, average error rate of measurements: 0.16\% for Robert 0.12\% for Gamma

- $E_T = \text{Error due to truncation}$
  - Running the operation for any time less or more than the required operation time
Sources of Computational Errors

- \( E_A \) = Error due to function approximation
  - Functions often must be approximated to implement in SC
  - i.e. Bernstein approximation for Gamma correction

\[ E_{Total} = E_G + E_S + E_M + E_T + E_A \]

- Some of these sources of errors can **offset or compensate for each other**, resulting in an acceptable total error.

- In an actual chip fabrication, the effect of **thermal noise** and the **influence of process and temperature variations** might introduce more inaccuracy in the generated signals
Sources of Computational Errors

- **Average error rate** of the produced images (20 trials) for different rates of inaccuracy in the duty cycle and period of PWM signals.

![Graphs showing output average error rate vs. duty cycle inaccuracy and period inaccuracy](image)

- Even with 20% relative error in the duty cycle and period, still stochastic circuits can produce acceptable outputs.
Conclusions

- **Conventional approach** of generating stochastic signals is costly!
  - Expensive SNGs
  - High latency, High energy consumption
- Proposed a **low-cost energy-efficient approach** for generating stochastic signals based on PWM signals
  - Area, latency, energy consumption are all greatly reduced
  - 99% performance speedup, 98% saving in energy dissipation, and 40% area reduction compared to prior stochastic approaches
- With time-encoding data, the proposed mixed-signal approach
  - Inherits the fault-tolerant advantage of stochastic design
  - While working as fast and energy efficiently as conventional binary designs
Further reading

• PWM signals in independent stochastic design

• PWM signals in correlated stochastic design
  • “High-Speed Stochastic Circuits Using Synchronous Analog Pulses”, M. Hassan Najafi and David J. Lilja, ASP-DAC 2017
Thank you

Questions?

M. Hassan Najafi
Najaf011@umn.edu