Stochastic Computing: A Design Sciences Approach to Moore’s Law

Naresh Shanbhag

Department of Electrical and Computer Engineering
Coordinated Science Laboratory
University of Illinois at Urbana-Champaign
Computing and Moore’s Law

- Computing (Deterministic Foundations)
  - Deterministic Computing
    - Turing Machine
    - Worst-case design
    - Computing
      - Boole
        - Boolean logic
      - Von Neumann
        - Programming models
          - Architecture
          - ideal switch
          - Devices
          - Applications
          - Stochastic
          - Deterministic
          - Physical Sciences
            - based
            - Moore's Law
              - 'avoiding statistics'
              - Current
              - Bardeen
              - Kilby
Stochastic Computing and Moore’s Law

- Design Sciences based on Moore’s Law ('embracing statistics')
- Communications (Statistical Foundations)
- Stochastic Computing
- nominal-case design
- Device Physics (Statistical Foundations)
- scaled and post-si devices

Proposed

Applications

Programming models

statistical inference

probabilistic switch

Devices
Von Neumann (1956)


....treatment of error is *unsatisfactory and ad hoc*

....*error should be treated .....as information has been*, by the work of L. Szilard and C. E. Shannon.

The present treatment *falls short of achieving this*..........

communications-inspired (stochastic) computation
Alternative Computational Models

Research theme in the Gigascale Systems Research Center (GSRC)

Sponsors: DOD & SRC
Stochastic Computing Panorama

Theoretical Foundations

- Kernels: FIR, FFT, Viterbi, video, comm., ML

Design Methodologies

- R. Kumar (UIUC), Roychowdhury (UCB), Malik (Princeton)

Application-specific Kernels

- Jones (UIUC)
- Singer (UIUC)

Custom ICs

- SSNOC
- ECG
- FIR

Stochastic Processors

- FPGA Prototypes

FPGA Prototypes

- BEE2- ERSA

Stochastic SOC

- ERSA [Mitra]
**Deterministic Computing**

- Input space: $2^N$
- Output space: $2^M$
- One-to-one (relabeling)

**Stochastic Computing**

- Many-to-one (clustering)
- Many-to-many (don't cares)
- Many-to-many (probabilistic)

Logic minimization

Estimation and Detection

Error statistics: $P_{\varepsilon,\eta}(\varepsilon,\eta)$

Corrected output: $\hat{y}$
Statistical Estimation & Detection

\[ \hat{y} = \arg \max_{H_i} P(y_1, y_2, ..., y_N \mid H_i) \]

\[ \tilde{y} = \arg \max \prod_{k=1}^{N} P(y_k \mid \tilde{y}) \]

Metrics: maximum a posteriori probability (MAP), maximum likelihood (ML), minimum mean-squared error (MMSE), minmax, minimum absolute error
Error Statistics

VOS induced timing violation

Path delay distribution (8b RCA)

$V_{dd} = K_{VOS}V_{dd,crit}$

effective error rate vs. energy trade-off
engineer circuit error statistics
prefer long-tailed PDDs

FIR filter in 180nm

Measured error PMF at Vdd = 0.76V

Error Probability Mass Function (PMF)

16-bit ripple-carry adder
Stochastic Computing Techniques

Algorithmic noise-tolerance (ANT)

Soft NMR

Stochastic sensor NOC (SSNOC)

Stochastic Computing Framework

Likelihood Processing

IEEE Spectrum, Nov. 2010

The Era of Error-Tolerant Computing

Errors will abound in future processors... and that’s ok!

By DAVID LAMMERS / NOVEMBER 2010
Algorithmic Noise-Tolerance (ANT)

[Hegde, Wang, Shim, Varatkar, Abdallah]

\[ y_a = y_o + \eta \]
\[ y_e = y_o + e \]

\( \hat{y} \) corrected

\[ \text{Power} \]
\[ \text{Voltage} \]

\[ P_{\text{EC}} \]
\[ P_{\text{TOTAL}} \]
\[ P_{\text{main}} \]

\[ SNR_{uc} \ll SNR_e \ll SNR_{ANT} \cong SNR_o \]

high error-rates (up to 60%)
overhead (gate-count): 5%-22%
energy savings: 40%-70% (<1dB SNR loss)
ANT Techniques

**prediction-based**

[Diagram of prediction-based technique]

**reduced-precision replica**

[Diagram of reduced-precision replica]

**adaptive error-cancellation**

[Diagram of adaptive error-cancellation]

**input subsampled replica**

[Diagram of input subsampled replica]

**maximum a posteriori (MAP)**

[Diagram of maximum a posteriori (MAP)]
ANT-based Error-resilient FIR Filter

Prediction-based ANT (Hegde)

Wiener-Hopf predictor

3-tap predictor

0.5π BW
29-tap FIR

Simulation results

5X energy savings

Chip architecture

microphotograph

0.35μm
3.3V CMOS
32-tap FIR

FMAC: 88MHz;
ECMAC: 11MHz;
V_{dd-crit} = 3.55V; 2.25V
V_{dd-min} = 2.32V

measured results

3X energy savings
Error-resilient Motion Estimation

input sub-sampled replica (ISR-ANT) (Varatkar)

area overhead = 26%

2.5X energy-savings

130nm CMOS

Conventional vs. Proposed:
- PSNR increase: 1.5 dB
- PSNR variance reduction: 7X

ideal conventional proposed
Stochastic Sensor Network-on-a-Chip

Robust estimation

\[ Y_i = \theta + \eta_i \]

\( \theta \) is the estimate

\( \eta_i \) is the noise

\( Y_i \) is the observation

[Varatkar, Narayanan, Jones]

(sim) 130nm CMOS WID process variations

Error PMF at Vdd = 0.85V

5.8X energy reduction

86% error-rate handling, \( P_{det} > 90\% \)

500

400

300

200

100

Energy (pJ)

error-tolerance

2170X

energy reduction

5.8X

\( V_{dd} \) = 0.85V

prototype IC in 180nm CMOS
ECG Analysis IC @ MEOP

[Abdallah]

**ECG waveform**

*BIH-MIT ECG DB: 11bits, 200Hz*

**PAM-Tompkin Algorithm**

**45nm, IBM process**

![Diagram of the PAM-Tompkin Algorithm](image)
Soft NMR

- soft NMR architecture
- [Kim] MAP rule
- soft voter

**Error Statistics**

\[
P(y_0 | y_1, y_2, \ldots, y_N) = \frac{P(y_1, y_2, \ldots, y_N | y_0) \cdot P(y_0)}{P(y_1, y_2, \ldots, y_N)}
\]

**Soft DCT**

- 2X
- 10X
- 3.8X

**Soft DMR vs. TMR**

- 35% energy savings
- 2X more robust
Matching the statistical requirements of applications to the statistics of nanoscale fabrics

statistical application-level metrics

STOCHASTIC COMPUTING

statistics of nanoscale fabrics
Implications on Device Design

- In return for energy-efficiency, we can handle
  - non-deterministic device behavior
  - improved average case behavior for worse corner-case → long-tailed distributions
  - ‘low SNR switches’ → smaller gap between 1-variable and a 0-variable
  - ‘multi-state switches’ (instead of two)
Summary

• Device design combined with stochastic computing can drive Moore’s Law

• Device design for stochastic platforms
  – what are the device properties that result in favorable error statistics?
  – ideal switch model unnecessary
  – relaxed device specifications

• Design and programming methodologies

• New applications in biomedical, energy and security