UNIVERSITY OF MINNESOTA
Driven to Discover℠

Low Latency, Parallel Computing Scheme

Technology No. 20180024-20180086

Parallel Computing Harnessing Unary Encoding

A new frontier in computing, opened in the last decade, uses unary encoding: the data is not packed, but complex computations such as Gamma correction and edge detection in an image can be performed using very simple logic. This new computing method provides a deterministic parallel bit shuffling network that uses a simple, deterministic thermometer encoding of data to achieve near optimal sub-sampling. The approach results in zero random fluctuation and high accuracy, yet keeps the output bit stream length constant, hence overcoming the impractical large code size in a zero-fluctuation stochastic code. It uses core "stochastic" logic circuits that do not employ constant coefficients, making them significantly smaller than traditional stochastic logic that potentially spend a significant amount of resources to generate such constant coefficients. The low-latency, parallel computation scheme in the context of unary stochastic and unary deterministic computing that enables computing a parallel unary stream (i.e., all bits are generated in a single digital clock cycle, using parallel copies of a stochastic logic, generating the output stream in one clock cycle). It uses a hard-wired, deterministic shuffling network that de-correlates inputs and allows for low errors in general (even zero errors in some cases). Rather than relying on randomness as an essential requirement for stochastic logic input, deterministic shuffling and sub-sampling techniques are described for generating inputs to the core stochastic logic.

Smaller Area than Stochastic Methods

Conventional binary has been the dominant encoding of data in digital systems due to its compact representation. However, it requires the data to be "unpacked" before computation (e.g., multiplication has to be broken into partial product and accumulation operations). Stochastic computing suffers from random fluctuations and unpredictability of the output. Previous solutions to this problem involve significantly increasing the code size, which is not practical. Stochastic computing and deterministic computing on unary streams use simple "stochastic" logic to perform complex computation. However, one of the major limitations with this method is its long latency and circuit depth. This technology offers a parallel implementation that uses both a thermometer code and a hard-wired deterministic shuffling method, which significantly decreases latency with a moderate increase in area. When compared to previous stochastic computing methods, results on feed-forward and feedback

circuits show, on average, an (area × delay) value 7x smaller than of conventional binary and 8x smaller than previous stochastic work at a10-bit binary resolution.

Comparison vs. Conventional Binary				
No. of Bits	**10**	**11**	**12**	**13**
Area-Delay Product	8x smaller	4x smaller	2x smaller	1x

BENEFITS AND FEATURES:

- Simple, deterministic thermometer encoding of data
- Achieves very good to optimal sub-sampling
- Zero random fluctuation and high accuracy
- Keeps output bit stream length constant
- Core “stochastic” logic circuits
- Low-latency, parallel computation scheme
- Enables computing a parallel unary stream
- Hard-wired, deterministic shuffling method de-correlates inputs and allows for low errors
- Superior to binary computing for medium resolutions (8-12 bits)

APPLICATIONS:

- FPGA architectures: FPGAs provide two useful features this method uses: abundant buffered routing resources and flipflops, and the ability to handle large fanouts
- ASIC
- Full-custom digital chips operating in domains (e.g., image processing and signal processing)
- Low-power signal/image processing, e.g., in Internet of Things
- Machine learning/deep learning systems
- Approximate computing applications and applications that tolerate some degree of uncertainty (e.g., video processing, image tagging)

Phase of Development - Conceptual

Researchers

David Lilja, PhD
Professor, Electrical and Computer Engineering
External Link (ece.umn.edu)
Kia Bazargan, PhD
Associate Professor, Electrical and Computer Engineering
External Link (ece.umn.edu)
Marc Riedel, PhD
Associate Professor, Electrical and Computer Engineering
External Link (ece.umn.edu)

Soheil Mohajer, PhD
Assistant Professor, Electrical and Computer Engineering
External Link (ece.umn.edu)

Interested in Licensing?

The University relies on industry partners to scale up technologies to large enough production capacity for commercial purposes. The license is available for this technology and would be for the sale, manufacture or use of products claimed by the issued patents. Please contact Doug Franz to share your business needs and technical interest in this technology and if you are interested in licensing the technology for further research and development.

https://license.umn.edu/product/low-latency-parallel-computing-scheme