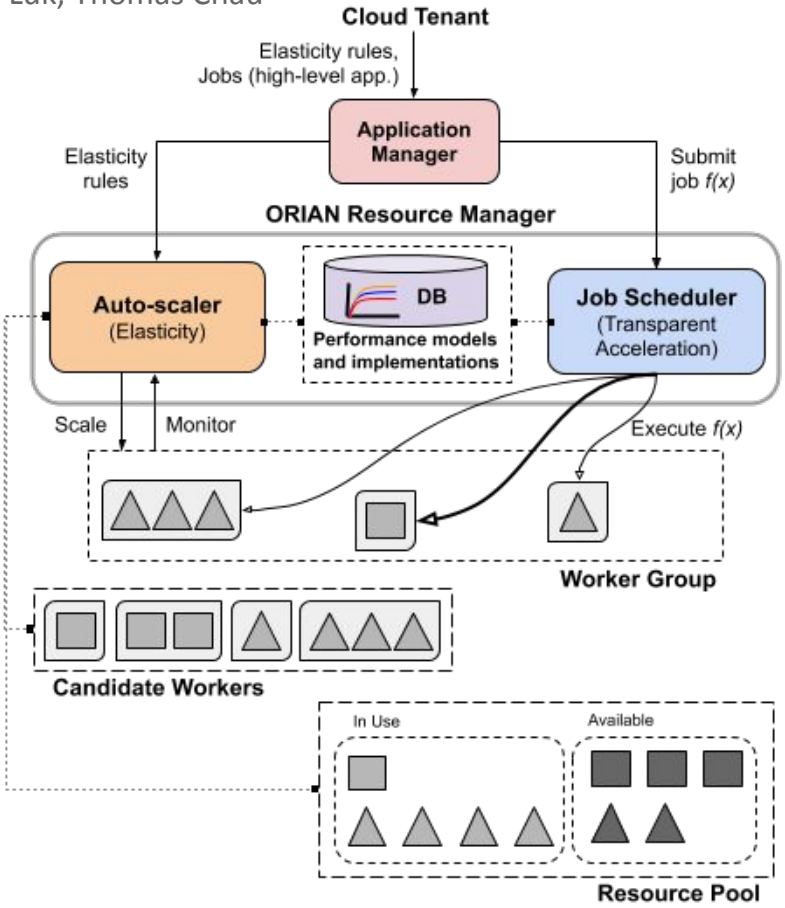
Transparent Heterogeneous Cloud Acceleration

Jessica Vandebon, José G. F. Coutinho, Wayne Luk, Thomas Chau

➔ **Goal:** make heterogeneous compute resources accessible to resource-oblivious cloud applications

➔ **ORIAN** extends homogeneous PaaS execution model:

1. **transparent acceleration:** automatic runtime mapping of jobs to the best worker
2. **heterogeneous elasticity:** automatic vertical (type) and horizontal (quantity) scaling to support QoS while minimising cost



Imperial College London

intel

CRbS: A Code Reordering Based Speeding-up Method of Irregular Loops On CMP

Authors：Li Yuancheng , Shi Jiaqi

- **CMP(chip multiprocessor) used to improve throughput and speed up multithreaded applications is becoming more and more commonplace.**
- **It is difficult for compilers to automatically parallelize irregular single-threaded applications which have complex data dependence caused by non-linear subscripts, pointers, or function calls within code sections.**

The Fig.1 shows the execution model. From the Fig.1(d) , we can see that RAW (Read-After-Write, can be detected by hardware mechanisms) occurs, the speculative thread will be restarted. So we adopt the code reordering method to reduce the RAW.

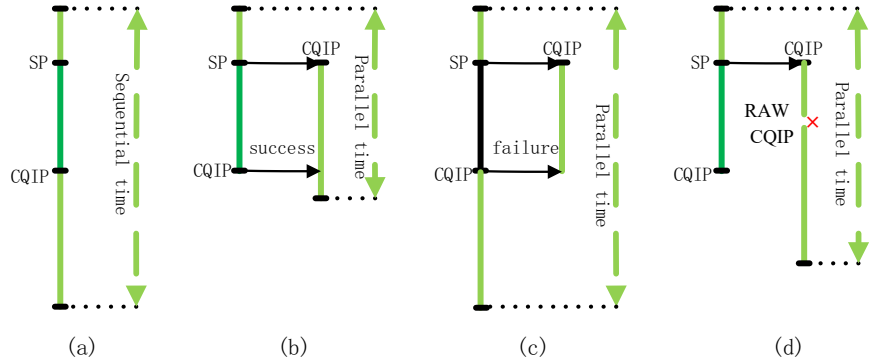


Fig.1 Sequential VS SpMT parallel execution

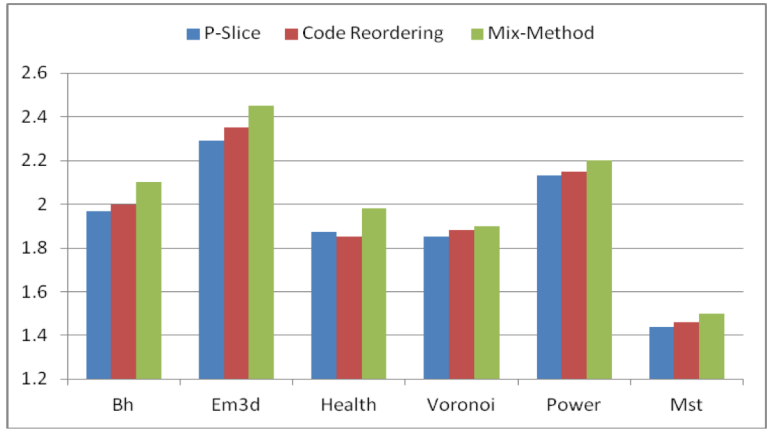


Fig.2 The Speedup based the code reordering method

By means of code reordering , we can cut down the data-dependence by reordering the key codes that may cause data-dependence. And Fig.2 shows the experimental results . The experimental results show that the proposed method is effective.

IOWA STATE UNIVERSITY
College of Engineering

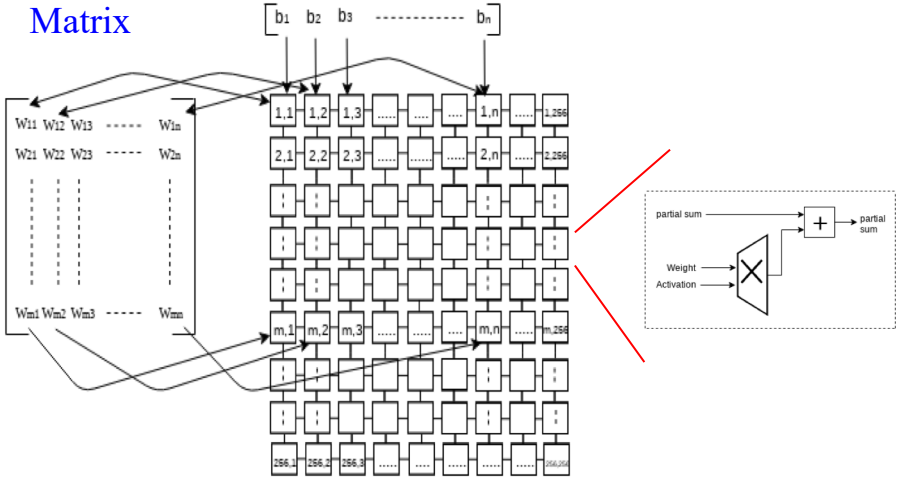
Impact of Structural Faults on Neural Network Performance

Krishna Teja Chitty-Venkata and Arun Somani
Iowa State University, Ames, IA, USA

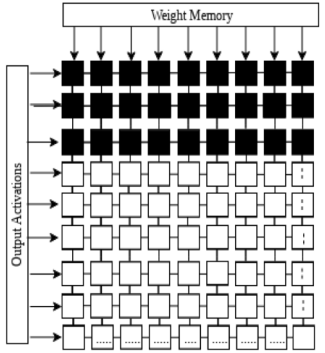
Dependable Computing and Networking Laboratory

Systolic Array

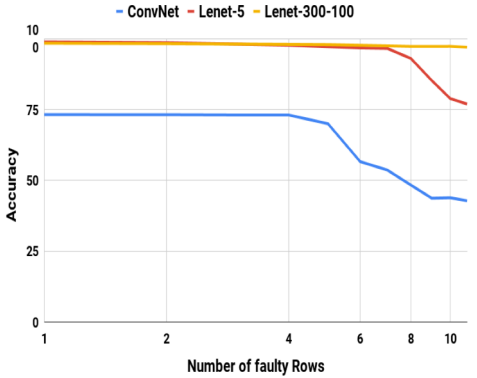
Weight Matrix



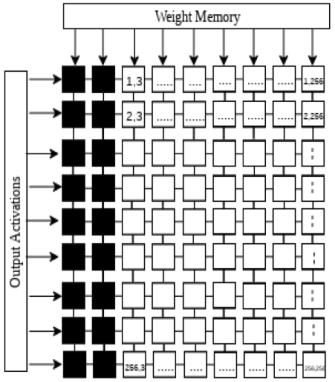
Row Faults

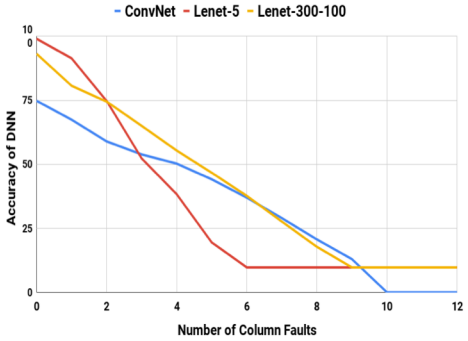


Impact of Faults



Column Faults





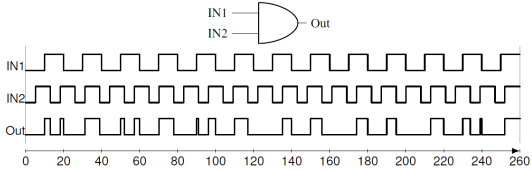
Row Faults: DNNs are resistant to row faults till an extent
Column Faults: They have have very high impact on network accuaracy (if the column is used)

Mitigation Stratgies: Matrix Transpose, Array Reduction

Energy-Efficient Near Sensor Convolution using Pulsed Unary Processing

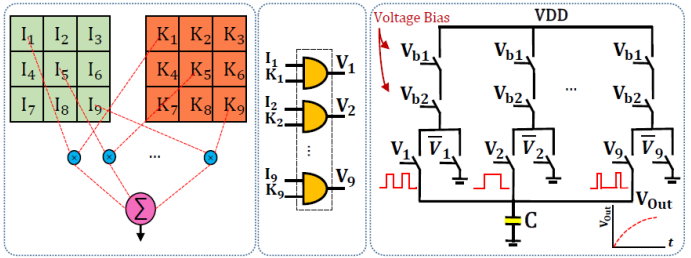
M. Hassan Najafi, S. Rasoul Faraji, Kia Bazargan, and David Lilja

- **Near-sensor convolution has many applications in IoT**
  - Conventional **fixed-point** designs**: Fast** and **accurate,** but **complex** and **costly**
  - **SC-based** designs: **Low-cost**, but **low accuracy**, **long latency**, **high energy consumption**
  - Analog-to-digital (**ADC**) and analog-to-stochastic (**ASC**) converters are also **costly**
    - ❖ **Traditional designs inefficient for near-sensor processing**
- **Pulsed Unary Processing**
  - Introduced recently for high-performance processing using stochastic logic
  - Inputs represented using PWM signals, duty cycle determines the value



Example of multiplication using PWM signals

- **Proposed Design**
  - **Input data** are converted to **inharmonic PWM signals**
  - **AND** gates are used to do the **multiplications**
  - **Outputs** of AND gates are accumulated using an **active integrator**



- **Evaluation**
  - **Significant improvement in the hardware area cost, power, and energy consumption**
  - **Avoiding costly ADCs**

An Efficient Application Specific Instruction Set Processor (ASIP) for Tensor Computation

Huang Weipei, CHEUNG Chak-Chung Ray, YAN Hong

Department of Electrical Engineering
香港城市大學
City University of Hong Kong

- **Definition -** a processor which have their instruction sets augmented and tailored towards a specific application.
- **Advantage** – programmable, more flexible, lower cost and time for implementation shortened; higher power efficiency
- **Current application-specific processors – TPU** invented by Google; China **Cambricon** released the Cambrian processor for computer vision and AI

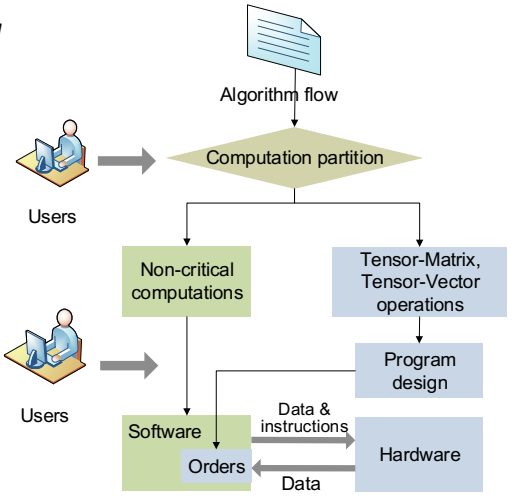


Fig. Typical design flow on ASIP system

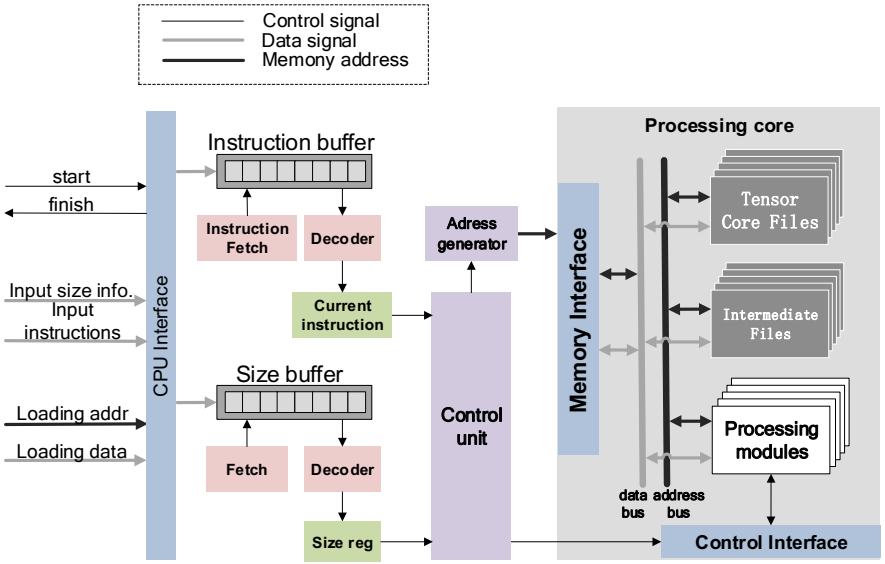


Fig. ASIP architecture for tensor processing

TABLE I Resource utilization

	Utilization	Utilization(%)
LUT	26317	8.67
LUTRAM	930	0.71
FF	35464	5.84
BRAM	157.5	15.29
DSP	62	2.21

Running at 111MhZ
Power consumption: 1.87W
~14ms for CP decomposition of 40*40*1200 tensor

筑波大学
University of Tsukuba

University of Tsukuba
Center for Computational Sciences

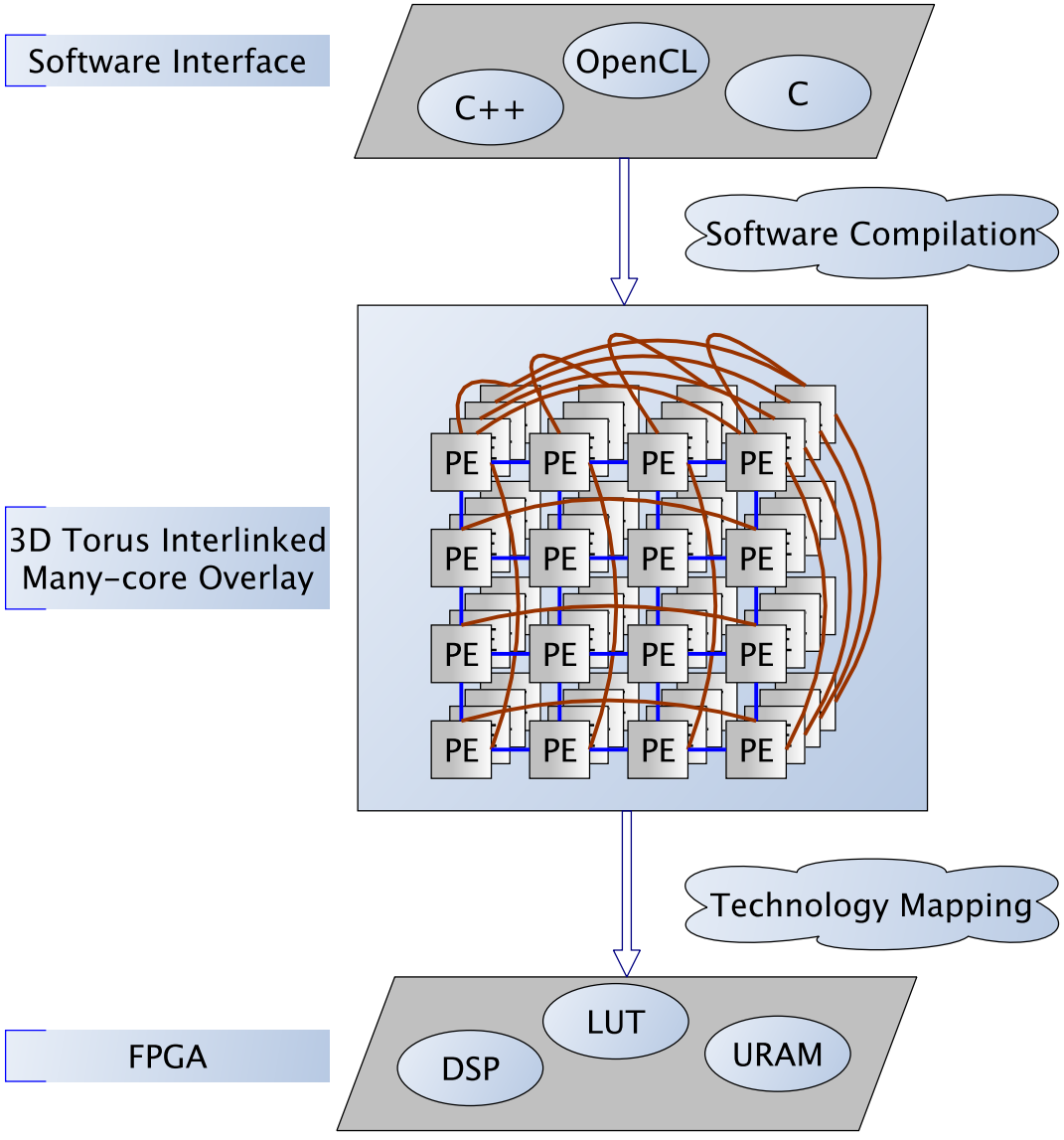
MITRACA: Manycore Interlinked Torus Reconfigurable Accelerator Architecture

The 30th IEEE International Conference on
Application-specific Systems, Architectures and Processors

Riadh Ben Abdelhamid, Yoshiki Yamaguchi and Taisuke Boku

University of Tsukuba, Japan
Graduate School of Systems and Information Engineering
Department of Computer Science

July 15, 2019

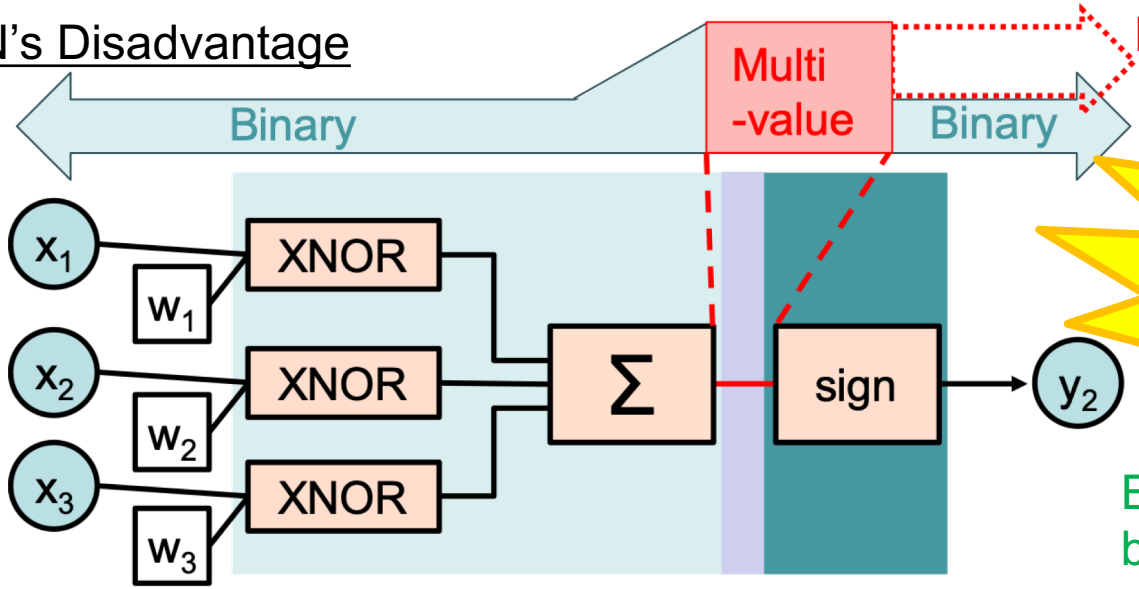


Overview of the proposed overlay architecture

#103 DeltaNet: Differential Binary Neural Network

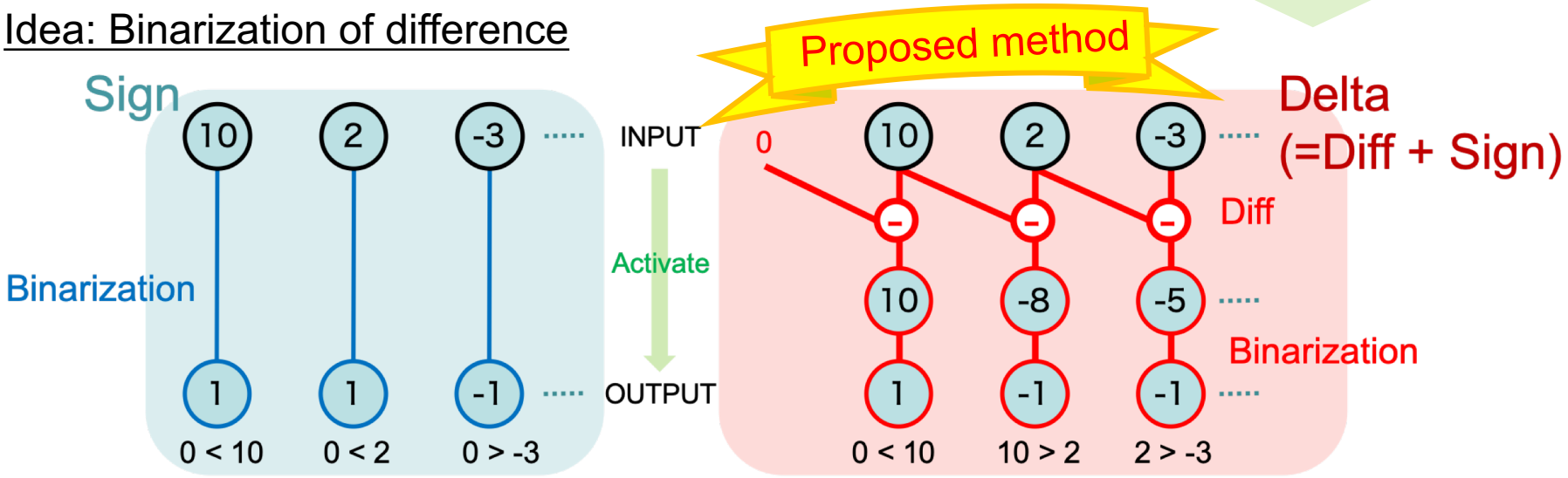
Yuka Oba*, Kota Ando*, Tetsuya Asai*, Masato Motomura†, and Shinya Takamaeda-Yamazaki*‡
(*Hokkaido University, †Tokyo Institute of Technology, ‡JST PRESTO)

BNN's Disadvantage



Effective use of information before binarization.

Idea: Binarization of difference



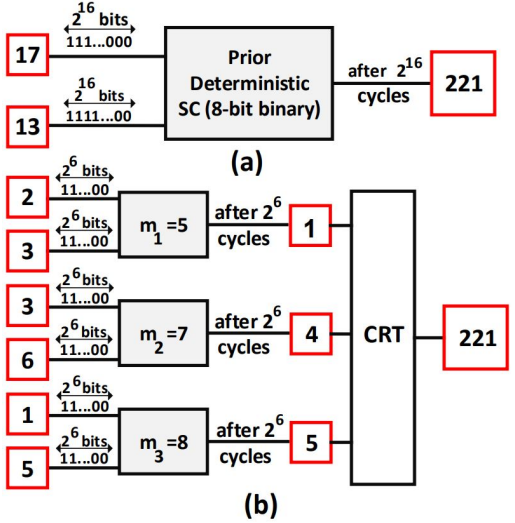
Comparison only with 0.

Leave the magnitude relationship.

Using Residue Number Systems to Accelerate Deterministic Bit-stream Multiplication

Kamyar Givaki, Reza Hojabr, M. Hassan Najafi, Ahmad Khonsari, M.H. Gholamrezayi, Saeid Gorgin, Dara Rahmati

- **Deterministic methods of SC**
  - Conventional SC: **low accuracy,** long latency, and high energy consumption
  - Recently proposed Deterministic methods of SC: **Fully accurate** but still long latency and high energy consumption
  - Guarantee that each bit of the first bit-stream sees each bit of the second bit-stream exactly once
    - ❖ **Conventional SC designs are not suitable for applications that require accurate computation**
- **Three Deterministic methods of SC**
  - Rotation of bit-streams
  - Clock dividing bit-streams
  - Using bit-streams with relatively prime lengths
- **Proposed Design**
  - **Combines** the idea of deterministic SC with Residue Number System **(RNS)**
    - **Residues have lower bit-widths -> exponentially shorter bit-streams**
  - The final results are in the **RNS format -> need to be converted back to binary**
  - **Number generators** can be **shared** between all of the **computation lanes**.
- **Evaluation**
  - **Smaller bit-stream generators**
  - **Significant improvement in terms of computation time and energy consumption compared to previously proposed deterministic methods of SC**



KALRAY

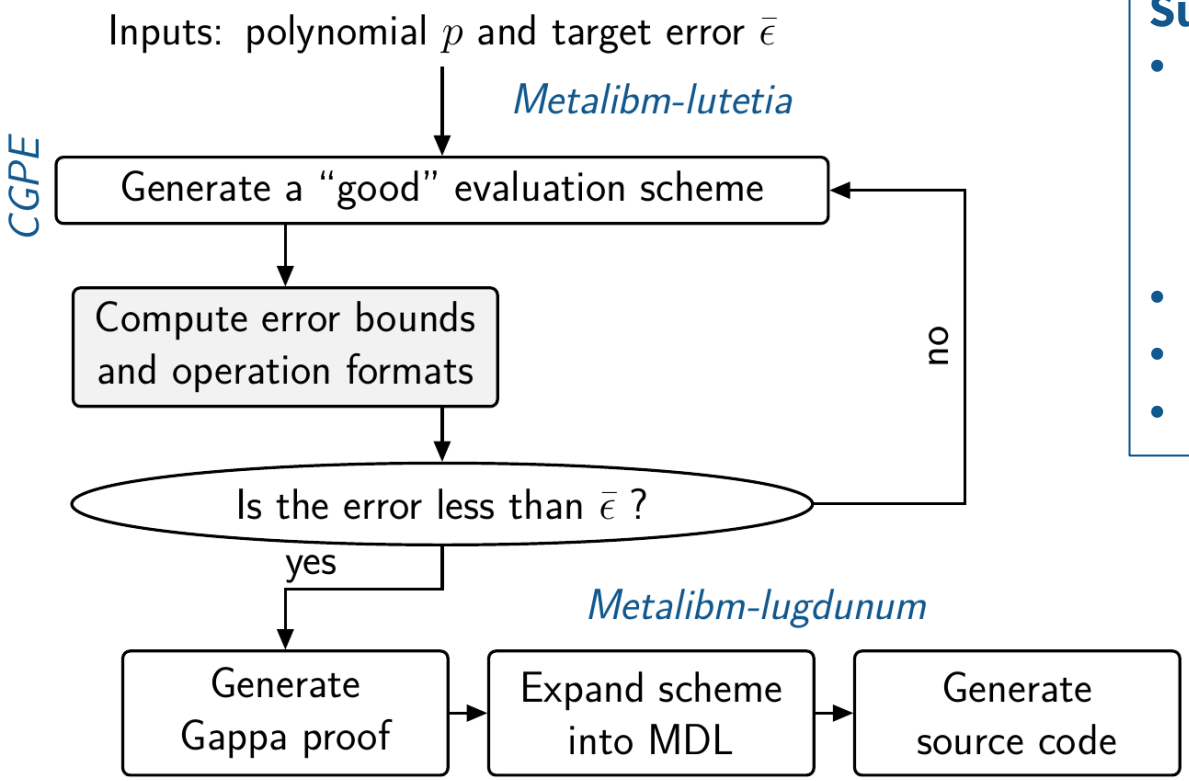
Precision Adaptation for Fast and Accurate Polynomial Evaluation Generation

Nicolas Brunie, *Christoph Lauter, Guillaume Revy*

ASAP 2019, Cornell University, July 15th 2019

www.kalrayinc.com

Application of "computing just right" to polynomial evaluation



Summary:

- Composing various tools
  - **CGPE**
  - **Metalibm-Lutetia**
  - **Metalibm-Lugdunum**
- Exploring imp. space
- Generating implementation
- Generating error certificate

Target	sine	zeros	exp	sinh
10	0.70	0.73	1.07	1.08
55	1.10	-	1.42	0.98
85	-	-	2.73	3.48