Future Directions in Computer Arithmetic: Panel

Miloš D. Ercegovac
University of California at Los Angeles

June 26, 2018

- **The best way to predict the future is to invent it** - Alan Kay, Turing Award Winner;

- Young researchers can and will do it - but where will they come from?

Views and Perspectives

- Computer arithmetic has had an amazing run:
  - From the Stibitz relay adder, assembled on a kitchen table in the late 30s, to the recent Google TPU IC with 64K 8 by 8 multipliers
  - Via many novel algorithms, number representations, and clever designs we made relevant contributions while remaining tiny compared to other fields.
  - Many solutions introduced by the ARITH community made it to the mainstream processors and systems.
  - Conferences like NIPS, ICML, WWW, VLDB, SIGKDD, SIGGRAPH, CVPR, etc. consistently draw very large audiences.

Job well done?

- Yes, but we have a growing problem of sustainability: submissions and attendance are shrinking
- Are we destined to follow a gradual underflow? Is there a format that can help us? Later.
- Issue: Research area size matters – too small to be recognized and have an impact factor. Not encouraging to young faculty striving for tenure
- Without faculty in arithmetic, how do we continue? Where/how do we educate our future researchers?
- Where would young researchers come from? Where would they learn arithmetic?
- Arithmetic courses in US universities have been vanishing: only a few leading CS/ECE departments retain arithmetic courses – for now.

- At UCLA: A graduate course in machine learning: 150+ students; in architecture: 40; in arithmetic: 15.
- To get students, I switched focus on arithmetic design explorations in popular areas (e.g., accelerators, neural networks, and approximate computing). Clearly, arithmetic alone is not attractive to grads.
- How do we cultivate and expand arithmetic knowledge if academia is not interested/supportive? Must combine with another area.

A Few Observations

- Opportunities for the growth are there. Perhaps we could define an arithmetic roadmap for next X years?
- Research trends in accelerators, ML and NN require massive use of arithmetic and optimization of memory organization/access.
- A new world for arithmetic: working with ML and NNs is similar to alchemy - a lot of trial and error. How to optimize arithmetic?
- High level synthesis could help deal efficiently with arithmetic developments
- Higher-order arithmetic algorithms (compound, composable): more

compute power, internal flexibility in representation, reduced standby activities, reduced interface with storage.

- Integration of storage and processing (PIM approaches) heating up: new memory technologies may have an effect on arithmetic research

- Flexible, application-appropriate formats are used: standardized formats considered wasteful and unnecessary. Big data is pretty noisy and this fact can be exploited in making arithmetic efficient.

- Applications using very low precision becoming common. Recent examples

  P. Judd, et al., *Stripes: Bit-serial deep neural network computing,* in MICRO, 2016.

  – per-layer selection of the precision

  H. Sharma et al., *Bit Fusion: Bit-Level Dynamically Composable Architecture for Accelerating Deep Neural Networks,* in ISCA, 2018.

  – 1 by 1, 2 by 2, 4 by 4, and 8 by 8 multiplications

- Analog schemes are being considered: physics to the rescue!

- Resistive RAM cell with conductance
  ~ neural network weight: w_{ij}
- Word line voltage
  ~ neural network input: x_i
- Cell current
  ~ $x_i \cdot w_{ij}$ (Ohm's law)
- Bit line current
  ~ $\sum_{i=0}^{N} x_i \cdot w_{ij}$ (Kirchhoff's law)

  In-memory convolution function

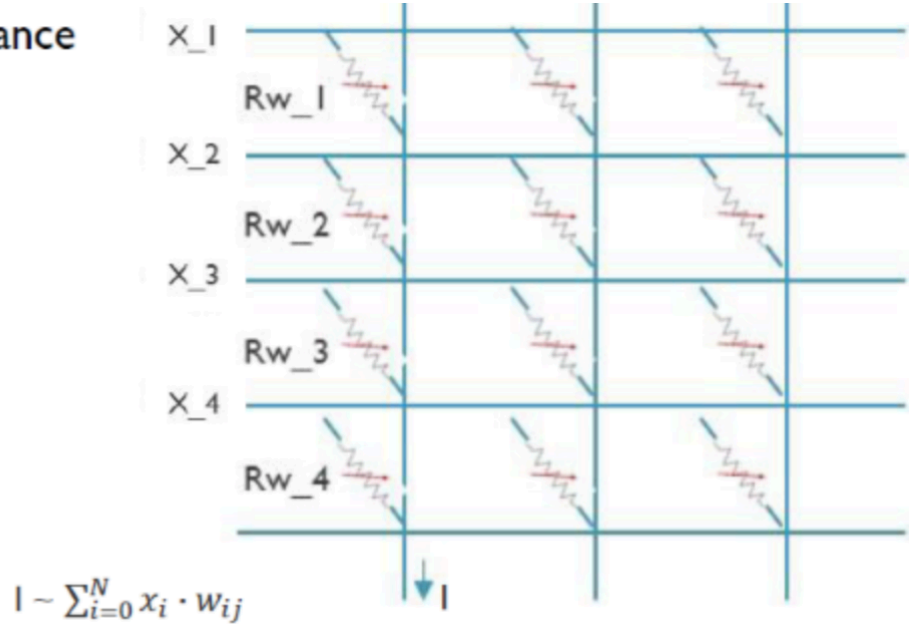

Figure 1: Analog Array.

– Startup Mythic (Austin, Texas) performs neural-network jobs inside a flash memory array, working in the analog domain to reduce power consumption.

– Imec 40-nm Low-Energy Neural Network Accelerator (LENNA) does computing and storing binary weights in relatively compact MRAM cells:

  the goal – a deep-learning inference chip using single-bit data-type. There goes the format dilema.

What Could Be Done for ARITH to Thrive?

- Symposium needs a larger, stable submission pool and attendance
  - Expand sessions on active arithmetic-related area (now ML and neural networks). Recent example: Asilomar added Machine Learning and submissions jumped 20%
  - Expand industrial tracks to attract more researchers/developers
  - Introduce poster sessions to increase attendance and enable meaningful many-to-many interaction.
  - Conventional presentations are limited to one-to-many, with superficially short interaction time: reduce to the best work.

– All contributions published as papers.

- Increase the number of invited papers bringing in established researchers with high-quality record in related areas. It would be a good way to expand the scope, raise the quality, and increase the visibility of ARITH to other areas

- Make session chairs do more work: Begin each session with a 10-minute theoretical minimum related to the papers and the key issues. Start discussion after a presentation with a prepared, well-thought out question

- Consider introducing tutorials on Sunday

- Consider making online tutorials on the ARITH web

On Reviewing and Getting More

- Conference Reviewing Considered Harmful, Thomas Anderson, OS, University of Washington, CACM 2009

  Conference reviewing, as it is currently practiced today, is harmful in two ways.

  1. Conference program committees spend an enormous amount of time on what ends up for many papers being close to a random throw of the dice.

  2. Worse, conference reviewing encourages misdirected effort by the research community that slows down research progress.

  Authors often think reviewers are random or biased; reviewers often worry authors are intentionally gaming the system.

Widely cited papers, are (i) early, (ii) left ample room for others to innovate, and (iii) was in a research area that had a low barrier to entry for other researchers. Only some of those three characteristics could be considered inherently valuable.

There is a heavy-tailed Zipf distribution of merit for conference submissions: the aggregate value of the rejected papers may be comparable to or even larger than the aggregate value of the accepted ones.

- Publish Now, Judge Later, Doug Terry, DB, UCI, CACM 2014.

  Accept any paper that extends the current body of knowledge.

  A conference publication is not the final publication of a research result, but its first publication.

  Through discussions and follow-on journal publication, the community will eventually reach judgment on the significance of the result.

- Jeff Naughton (DB), U Wisc, points out the self-reinforcing role modeling in reviewing: young CS authors who receive nasty reviews may internalize nastiness as the norm.

Thank you - please let's do something