"The Next Challenge: Energy Efficient Approach in ML Architecture"

Professor Uri Weiser
Viterbi Faculty of Electrical Engineering
The Technion
Israel

July 1st 2019

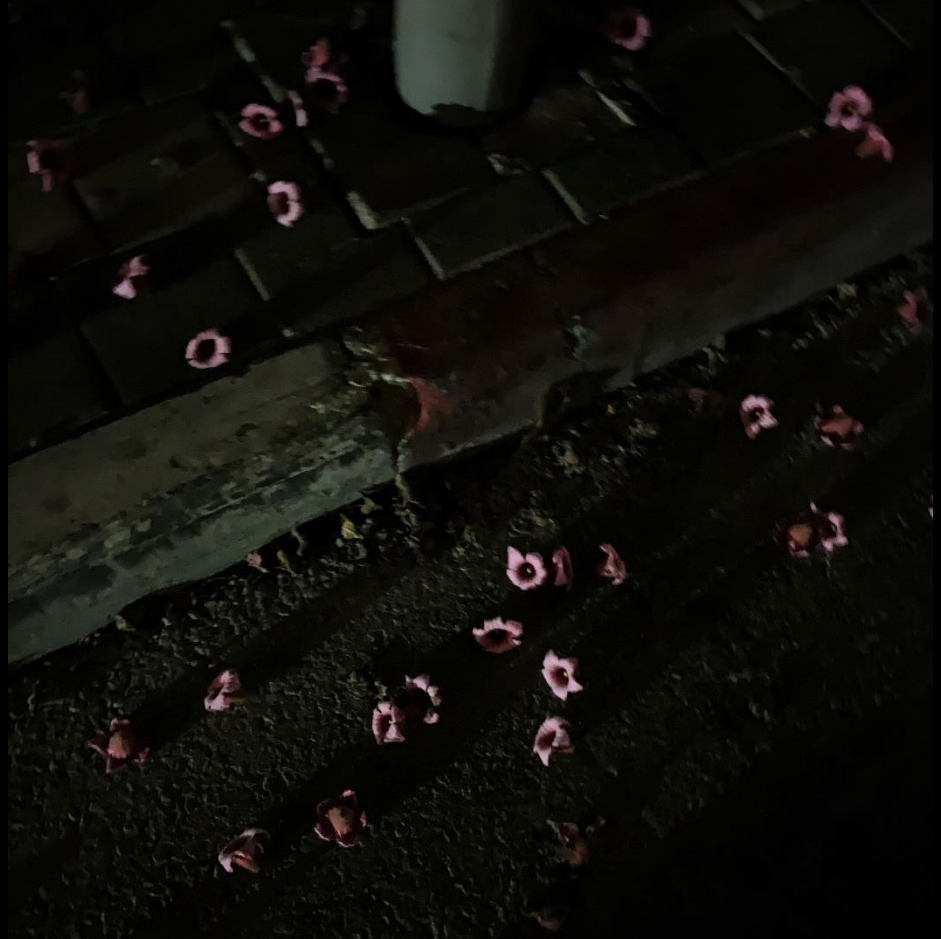
Contributors to the research: Leeor Peled, Daniel Raskin, Gil Shomron, Leonid Yavits, Moran Shkolnik, Avi Baum,

To Yale

5 years passed since Yale@75,

OK but why do you have to drag us with you?

Interesting how you keep staying in the center…

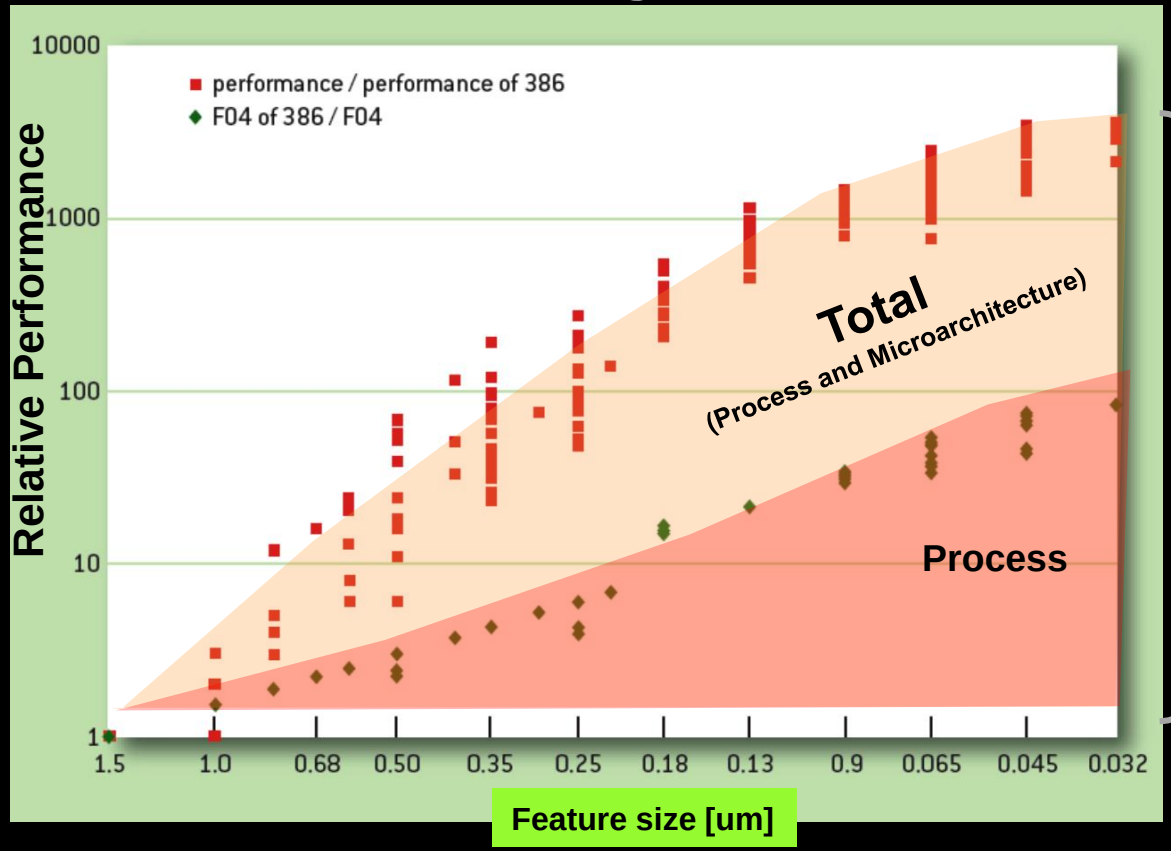


Beauty comes shining through not only when blooming

Agenda:

- Technology environment
  - Process is slowing down
  - Big Data
    - Funnel
  - Killer apps ➔ ML
- Efficient ML BASICS
  - Energy:
    - Amdahl and MA (divide effectively our limited resources)
    - SMT – is this a biggy?
    - Pipeline – why?
    - Map applications to HW
    - Prediction – no validation is necessary
- Conclusions

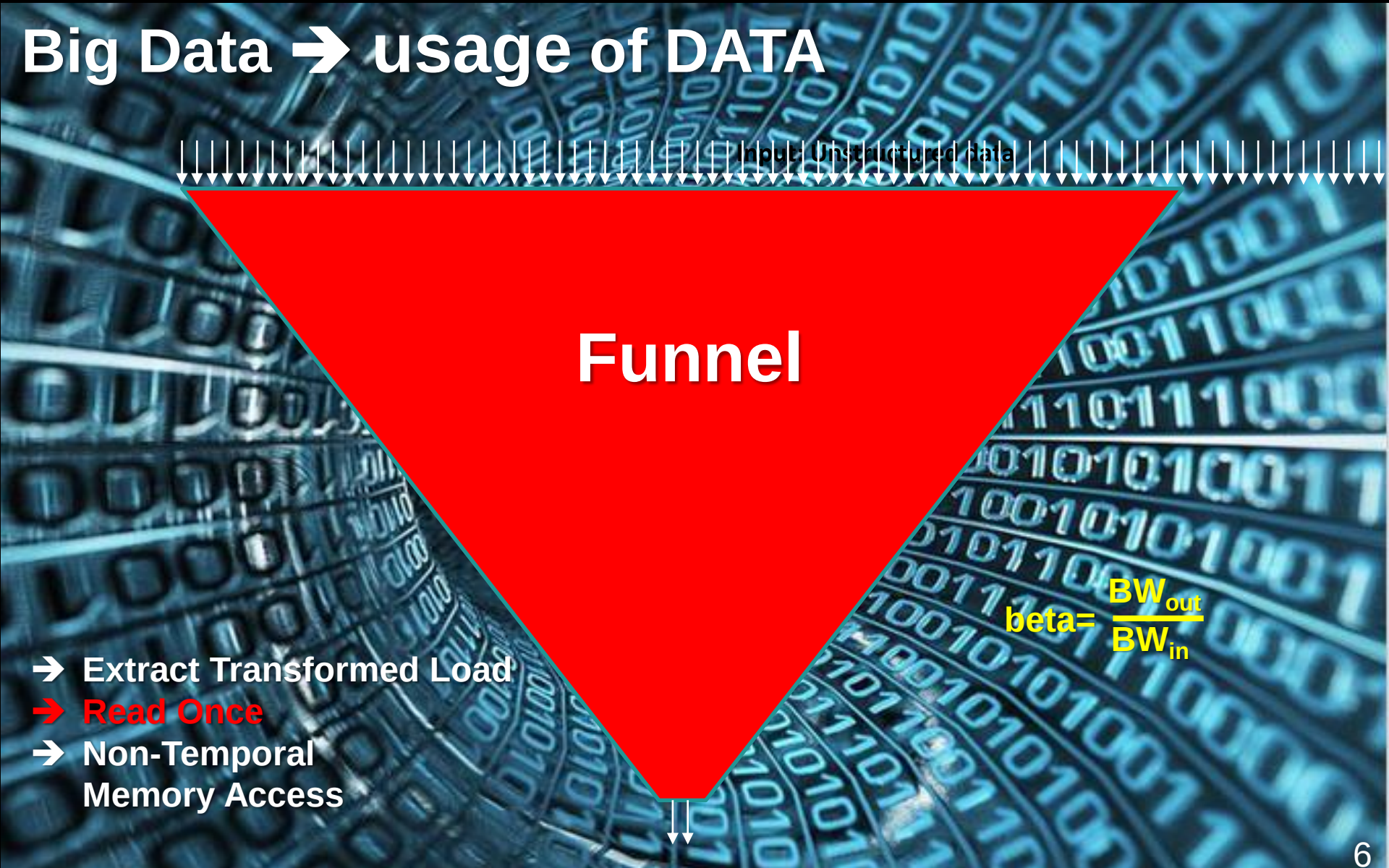
Performance History



We (the architects) did an "OK-" job

*ACMqueue April 6, 2012, Processors, Volume 10, issue 4 CPU DB: Recording Microprocessor History, Andrew Danowitz, Kyle Kelley, James Mao, John P. Stevenson, Mark Horowitz, Stanford University

Big Data ➔ usage of DATA



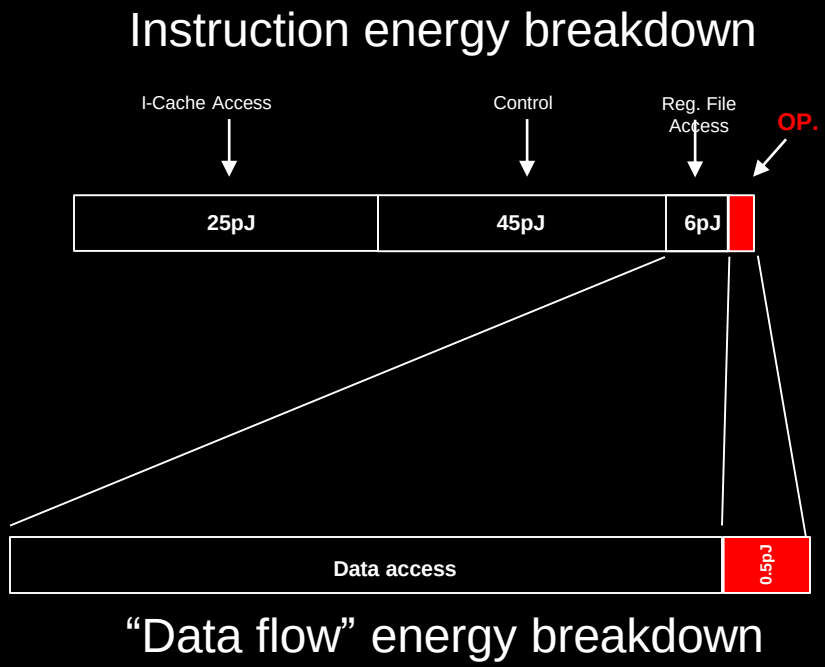
- ➔ Extract Transformed Load
- ➔ Read Once
- ➔ Non-Temporal Memory Access

Killer applications*

- ML is one!?
- Funnel (in most of the cases)
  - Input: huge amount of data
    Output: small amount
  - Many simple operations

*Applications you can not effectively execute on current HW (Dr. Andy Grove)

Energy in “Data Flow” architecture



Now Read Once counts!

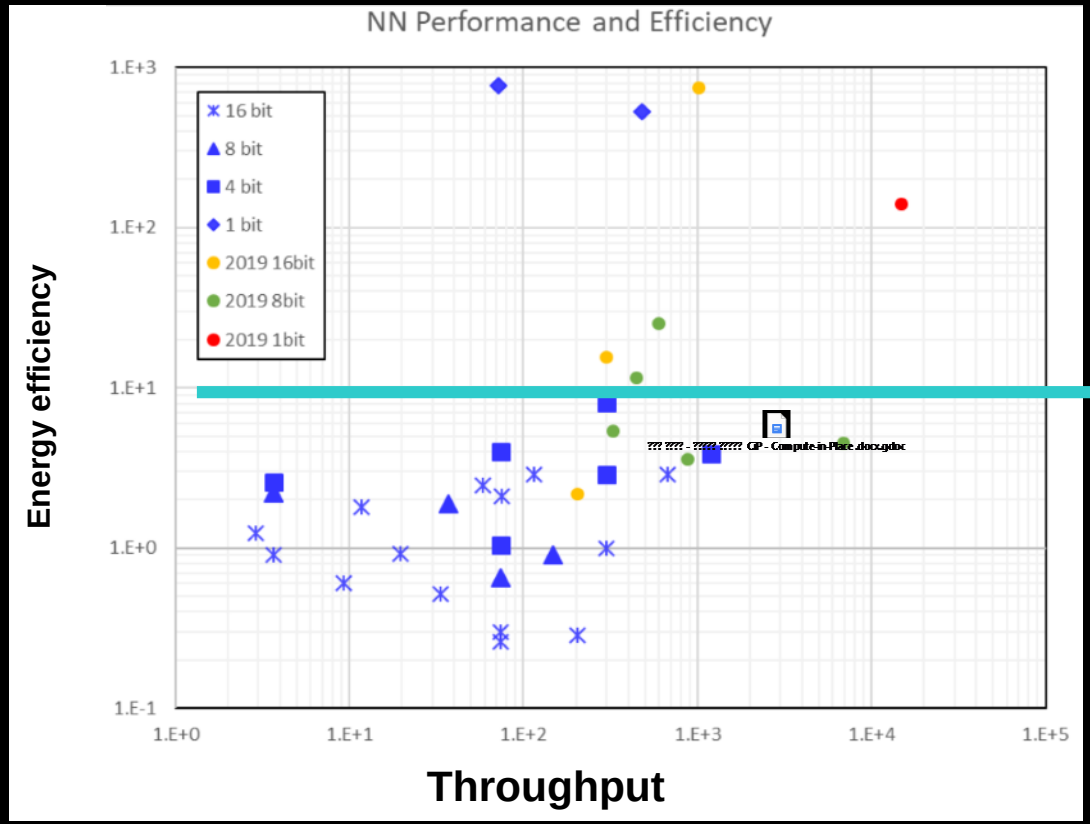
Efficient ML I Accelerator

- **Energy ➔ Performance**
  - **Map applications to HW ➔ Graph mapping; data flow**
  - **Efficient mapping**
    - **Co-design HW structure and smart compiler in specific application environment**
    - **Almost no flow control**
    - **Statistical results – no need to validate execution**

Efficient ML II Balanced design and energy reduction

- Energy ➔ Performance
  - System vs. Accelerators: It is Amdahl again!
  - Energy reduction
    - Reduction in Computing (MACs op.)
      - Pruning
      - Prediction
    - Reduction in data access and movement
      - Pipeline
    - Efficient usage of the Hardware resources
      - Multi-Amdahl (divide effectively your limited resources)
      - SMT

Efficient ML II: Reduction in Computing



0.1pJ/OP

1

TOPS/W drop due to inefficiency (e.g. data movement, DRAM repeated accesses…)

2

Energy efficiency α energy/OP

Efficient ML II: Reduction in Computing (1)

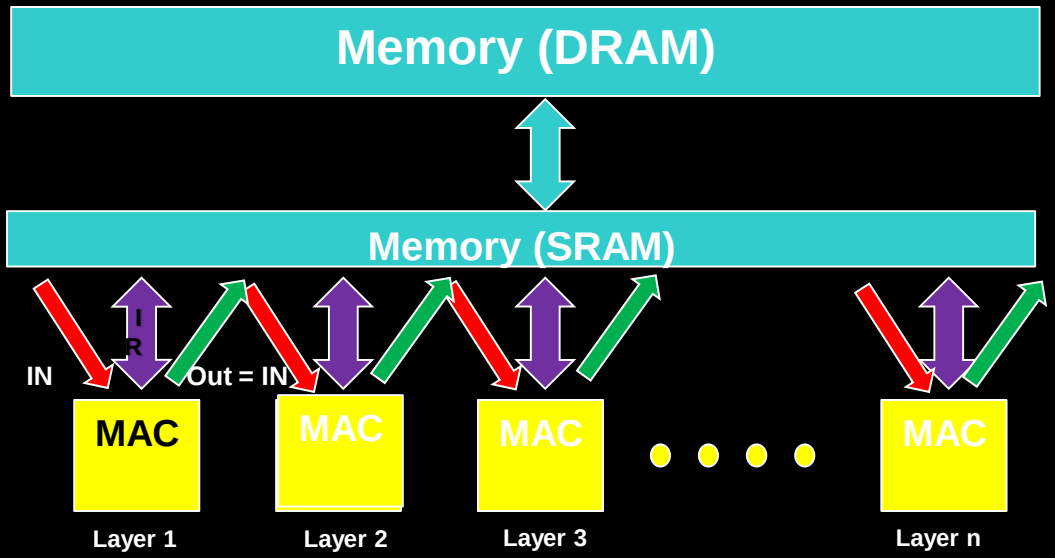
- Reduction in Computing → reduce # of operations via
  - Pruning
    - Well known techniques
  - Value Data (Prediction)
    - ML are statistical → no need to validate execution

G. Shomron, U. Weiser, "Spatial Correlation and Value Prediction in Convolutional Neural Networks"
IEEE Computer Architecture Letters (CAL) Journal January 2019

Efficient ML II: Reduction in Data Accesses (2)

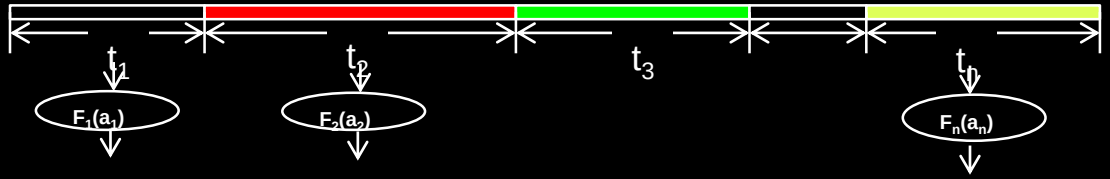
- Reduction in Data access and movements
  - Pipeline execution

➔ Data stays on die



Efficient ML II: Efficient usage of HW

- **Multi-Amdahl*** **(divide effectively your limited resources)**



Optimization using Lagrange multipliers

Target under a constraint A

F'= derivation of the accelerator Function

e.g. efficient resource division (e.g. SRAM)*

$$t_i F_i'(a_i) = t_j F_j'(a_j)$$

- **SMT****
  - **Resources needs are known ahead of time…**

*T. Zidenberg, Isaac Keslassy, U. Weiser, "Optimal Resource Allocation with MultiAmdahl" IEEE MICRO Journal August 2013

** Technion EE, Advanced Microarchitecture course's Exam (winter 2019)

** *G. Shomron, T. Horowitz, U. Weiser, "SMT-SA: Simultaneous Multithreading in Systolic Arrays" IEEE Computer Architecture Letters (CAL) Journal July 2019

Conclusions

- **Opportunities:**
  - **Map application to HW**
  - **Reduce energy per operation?**
  - **Reduce # of operations**
  - **Reduce data movement and memory access**
  - **Efficient usage of HW**
- **We’re gonna have fun**
- **Open field, lots of ideas, many researchers**
- **Opportunities**
- **New passionate energy in the community**
- **Back to the “big impact” era…**

Thank You