COMPUTE SYSTEMS GROUP

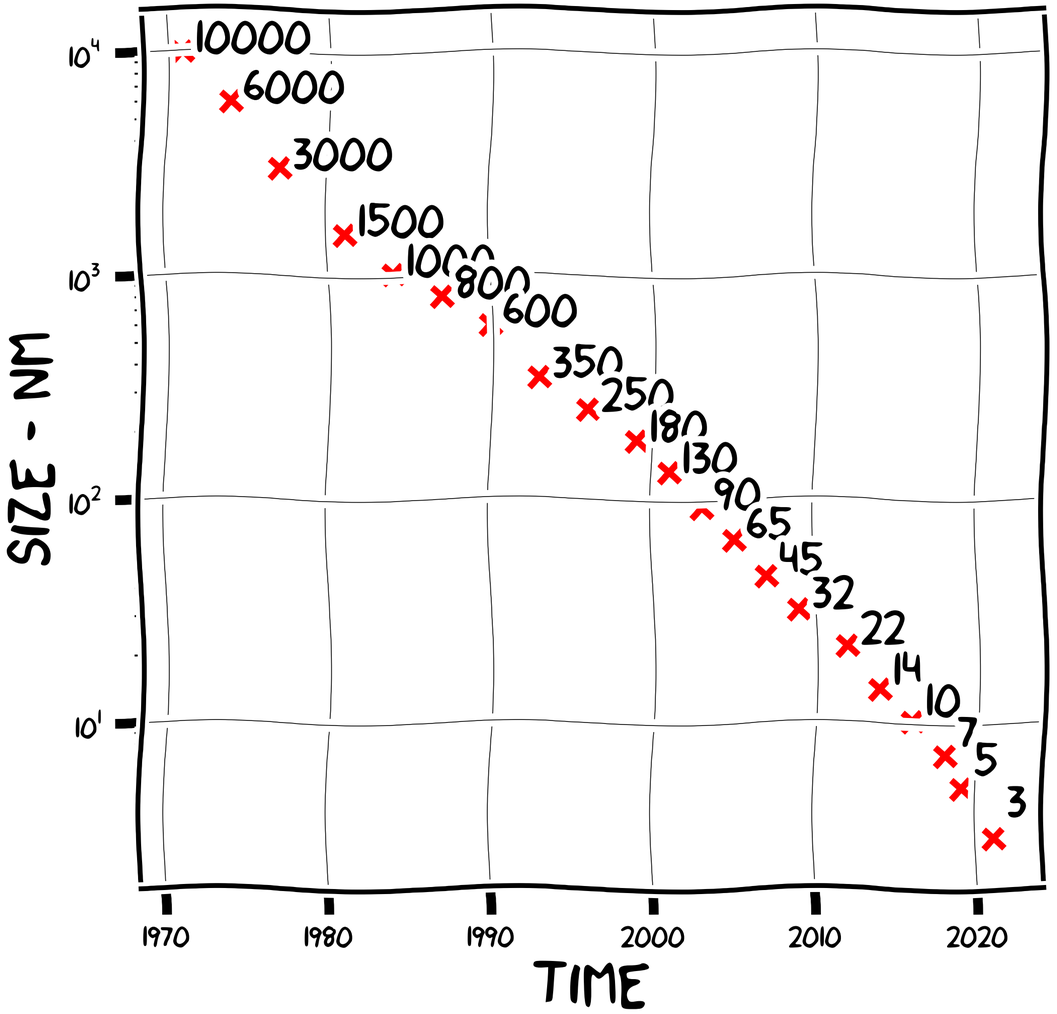
Holger Fröning & Andrea Seeger
Felix Zahn, Günther Schindler, Lorenz Braun, Matthias Hauck, Dennis Rieber, Laura Promberger, Bernhard Klein, Jonas Dann, Vahdaneh Kiani, Kevin Stehle

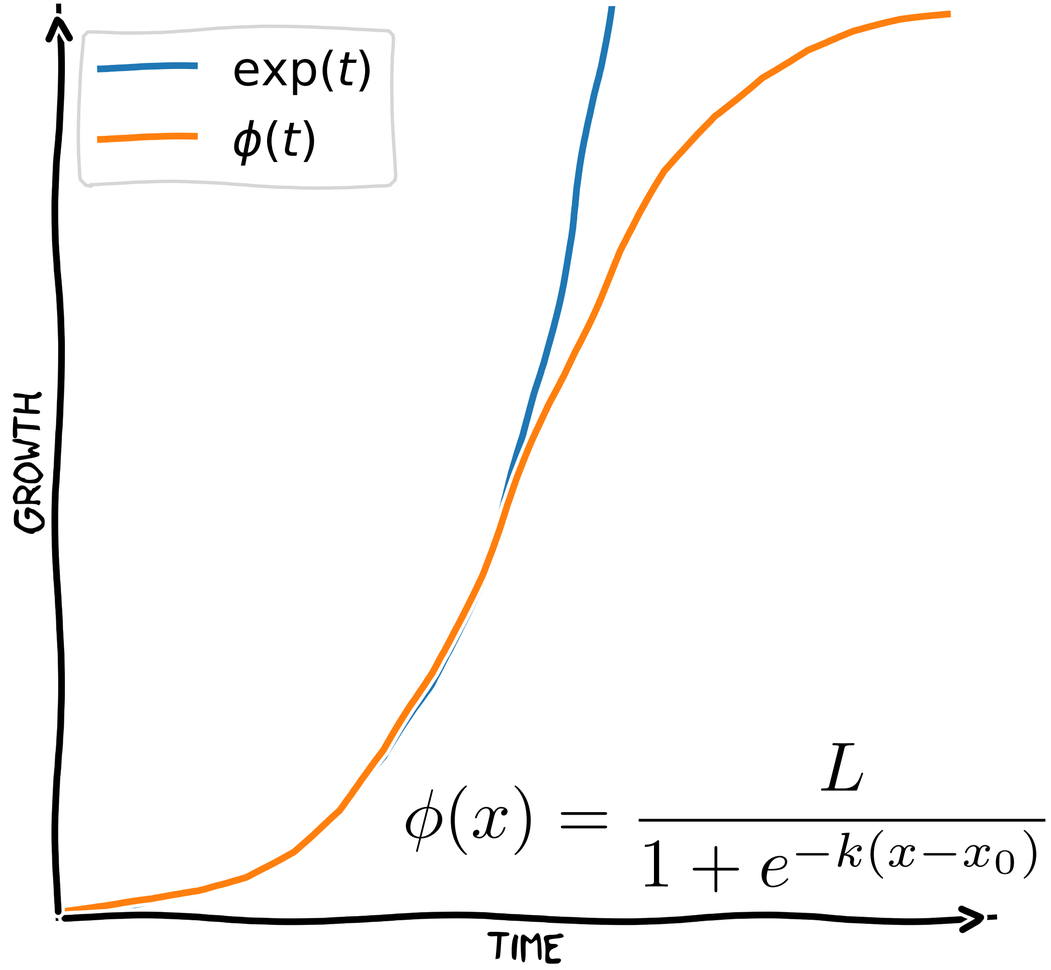
ZITI Inauguration, Feb 6, 2020

ABOUT US

- Professor at Heidelberg University, Germany
  - Currently advising 9 PhDs, 4 being external (3 on ML)
  - PhD from Mannheim University (Networks)
  - Post-Doc, TU Valencia & Heidelberg University (FPGAs, ASICs)
  - Visiting professor, TU Graz (2015 - first contact with ML)
  - Visiting scientist, NVIDIA Research (Santa Clara, CA) (2016)
- Performance, energy efficiency & programmability for future & emerging technologies
  - Computer architects with a focus on low-level software layers
  - High-performance computing, machine learning & data analytics
  - Bipolar computing landscape: super small & super big

MOORE'S LAW IS ALIVE - BUT EXPONENTIAL GROWTH IS AN ILLUSION



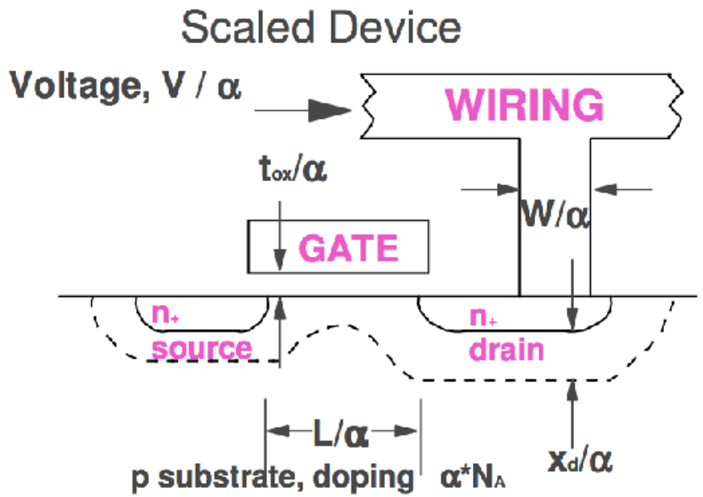


DENNARD SCALING

$$P = afCV^2 + VI_{leakage}$$

Results of classic scaling (scaling factor: α)

- Higher density (α^2)
- Lower voltage, capacity (each 1/α)
- Higher speed (α)
- Active power density (1)



Classic Dennard
Oxide: t_{ox} /α
Wire width: W/α
Gate width: L/α
Diffusion: x_d/α
Substrate: α * NA
Voltage: V/α
Current: I/α

Pradip Bose, IBM Research, 2014

END OF DENNARD SCALING

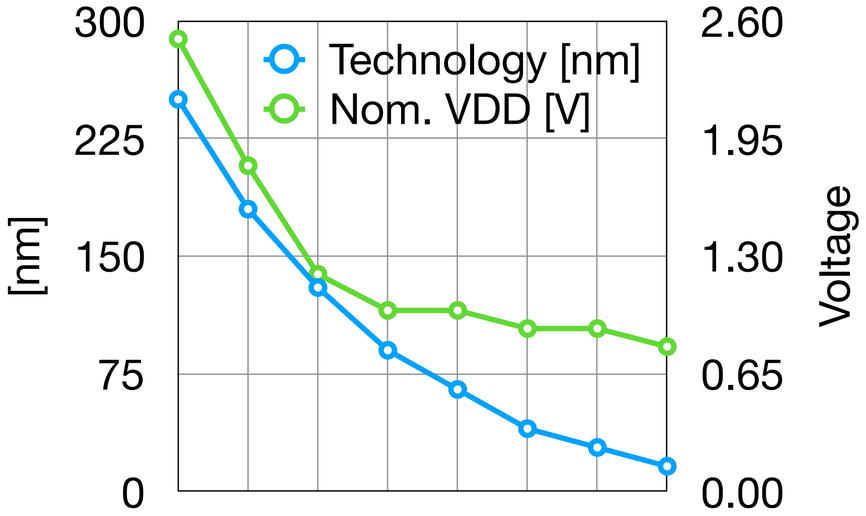
$$P = afCV^2 + VI_{leakage}$$

Results of classic scaling (scaling factor: α)

- Higher density (α^2)
- Lower voltage, capacity (each 1/α)
- Higher speed (α)
- Active power density (1)

Reality today: Post-Dennard

- Wiring & leakage power became a major issue
- Voltage scaling no longer possible
  - $V > V_{TH}$; $V_{TH} > 0.25\text{-}0.3V$
- Power no longer remains constant!
- Power budget limits device count (Moore vs. Dennard)



Classic Dennard
Oxide: t_{ox} /α
Wire width: W/α
Gate width: L/α
Diffusion: x_d/α
Substrate: α * NA
~~Voltage: V/α~~
Current: I/α

Pradip Bose, IBM Research, 2014

PERFORMANCE SCALING

CLASSICAL DENNARD SCALING

$$Perf(\frac{ops}{s}) = \underbrace{\frac{Instructions}{cycle}}_{\propto\ PipelineCount \cdot PipelineDepth} \cdot \underbrace{frequency}_{\text{scales with feature size}}$$

POST DENNARD SCALING

$$Perf(\frac{ops}{s}) = \underbrace{Power(W)}_{\text{fixed}} \cdot Efficiency(\frac{ops}{Joule})$$

REGIME I

operator cost —> *Specialization* —> *heterogeneity and asymmetry*

GPU COMPUTING

PROGRAMMING COMPLEXITY

NUMA EFFECTS & LATENCY HIDING

\+ *data movement cost*

REGIME II

Partly by Bill Dally, Sudha Yalamanchili (UCAA Workshop, 2012)


NVIDIA TESLA

GPU MEKONG PROJECT (BMBF)[1]

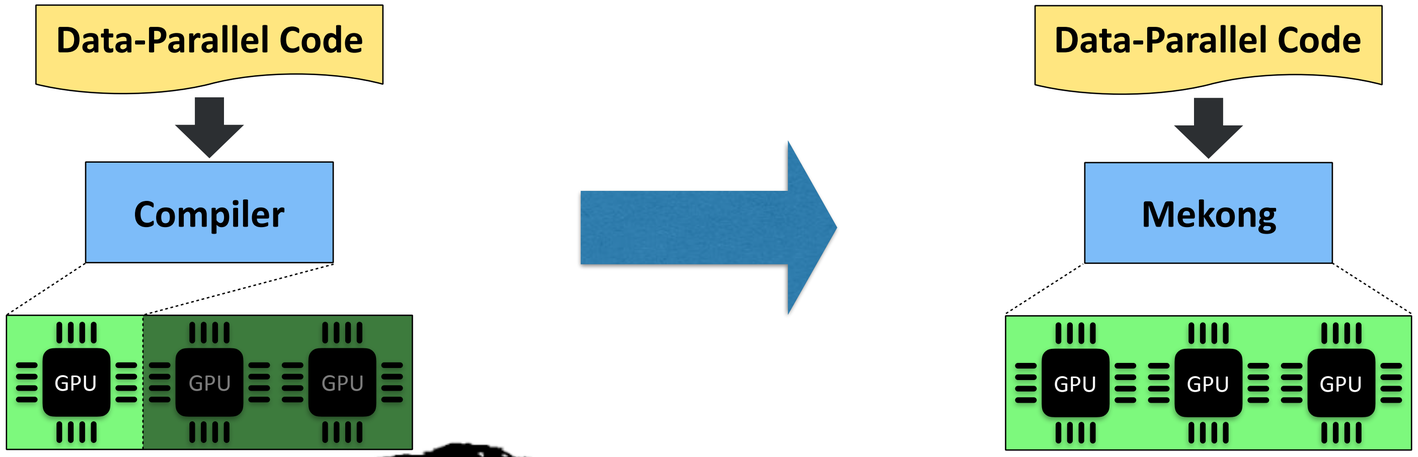
Automatically transform a single-device CUDA program into a multi-device program

- Key: automated partitioning and creation of communication tasks

Code analysis/code generation at compile time

- Key for good data partitioning is memory access pattern

Tool stack: LLVM-based code analysis and transformation



[1] *Received a Google Faculty Research Award in 2014*

OTHER RESEARCH EFFORTS

GPU Mangrove: predictive models for time and power of GPU kernels (part of Mekong)

ML Compiler: mapping DNN operators to special hardware primitives (Bosch)

Graphite: Multi-query support for graph computations (SAP)

ATLAS/LHCb: data acquisition systems for high-energy physics experiments (CERN)

CCUDA: co-scheduling for maximized GPU utilization (Mashhad University, Iran)

IGART: Real-time scheduling of Multi-GPU Systems for Image-guided Adaptive Radiation Therapy (HIDSS4Health)

PERFORMANCE SCALING

CLASSICAL DENNARD SCALING

$$Perf(\frac{ops}{s}) = \underbrace{\frac{Instructions}{cycle}}_{\propto\ PipelineCount \cdot PipelineDepth} \cdot \underbrace{frequency}_{\text{scales with feature size}}$$

POST DENNARD SCALING

$$Perf(\frac{ops}{s}) = \underbrace{Power(W)}_{fixed} \cdot Efficiency(\frac{ops}{Joule})$$

REGIME I

$$\underbrace{\text{operator cost}}_{\text{Specialization —> heterogeneity and asymmetry}}$$

GPU COMPUTING

PROGRAMMING COMPLEXITY

NUMA EFFECTS & LATENCY HIDING

+

REGIME II

$$\underbrace{\text{data movement cost}}_{\text{3 operands x 64bit/operand}}$$

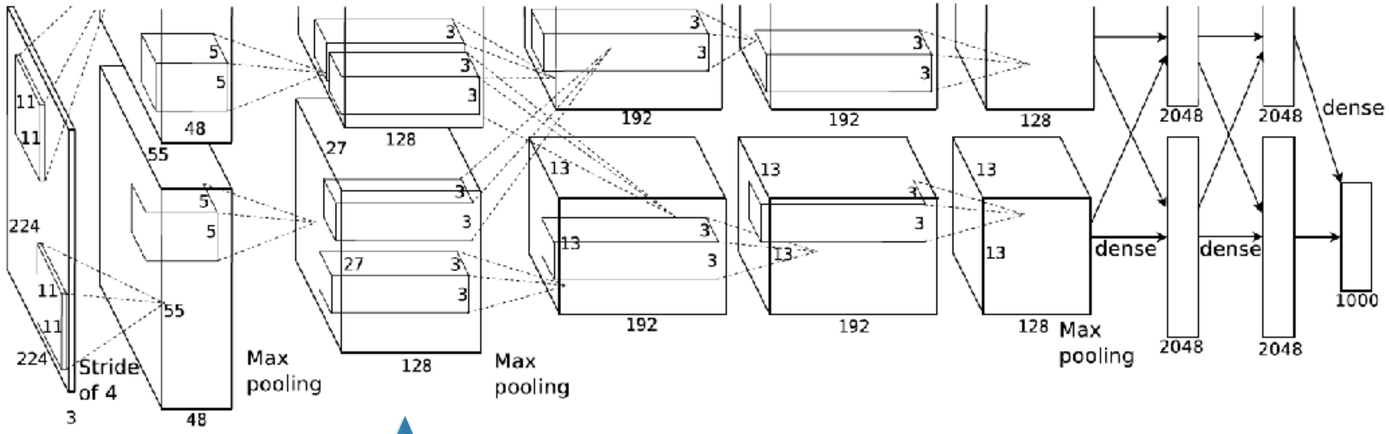
$$Energy = \#bits \cdot dist[mm] \cdot energy_{per\ bit, per\ mm}[\frac{J}{mm}]$$

SPATIAL COMPUTING

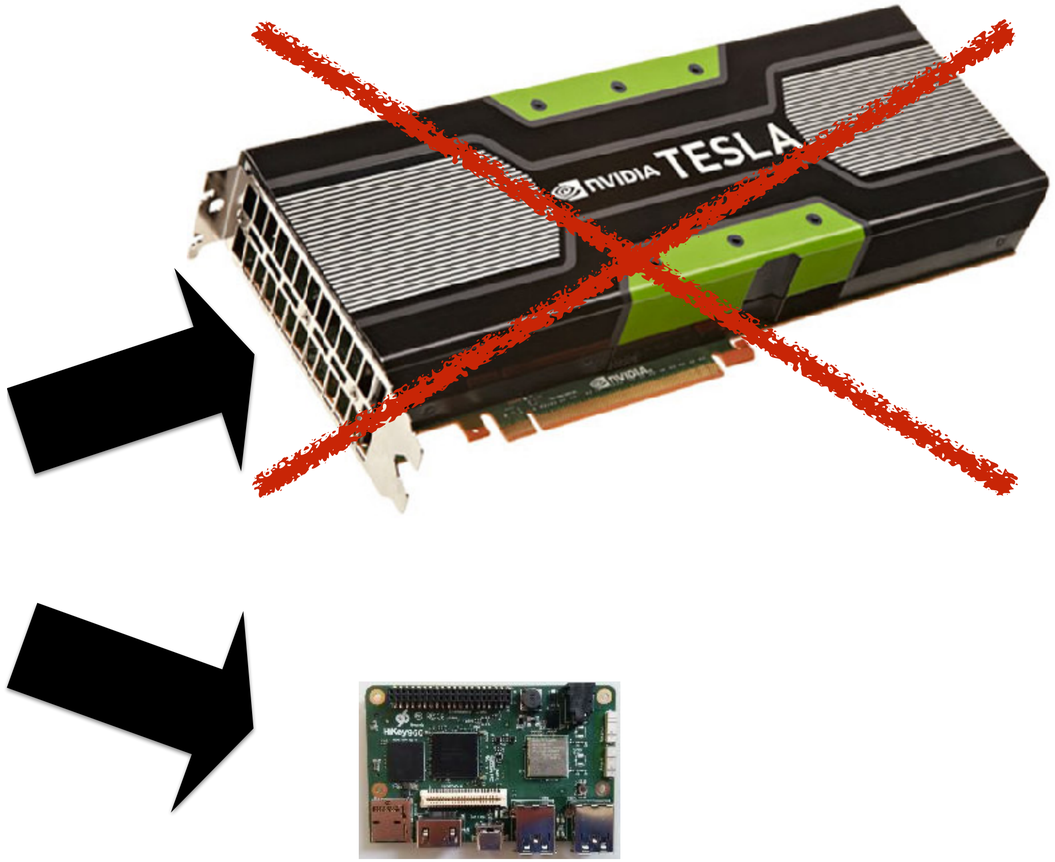
LOCALITY

PREDICTABILITY

Partly by Bill Dally, Sudha Yalamanchili (UCAA Workshop, 2012)

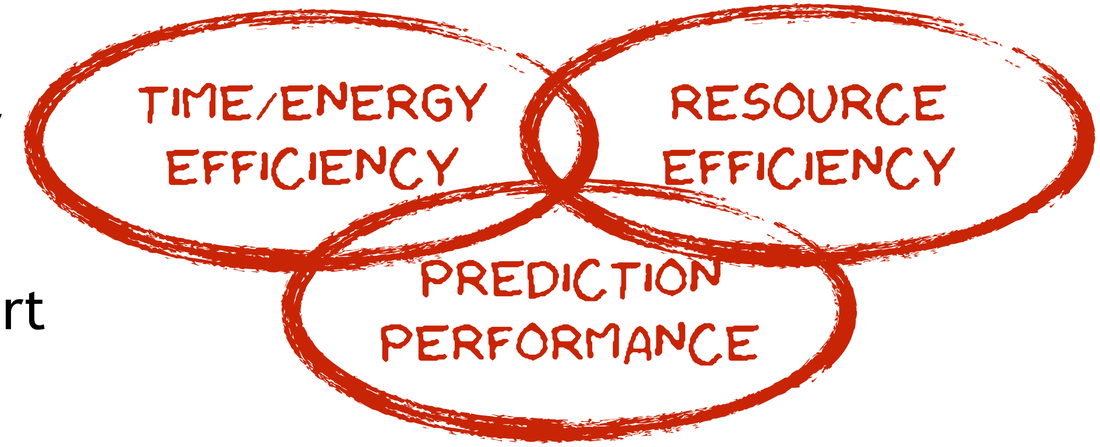


60M parameters, 724M multiply-accumulate operations

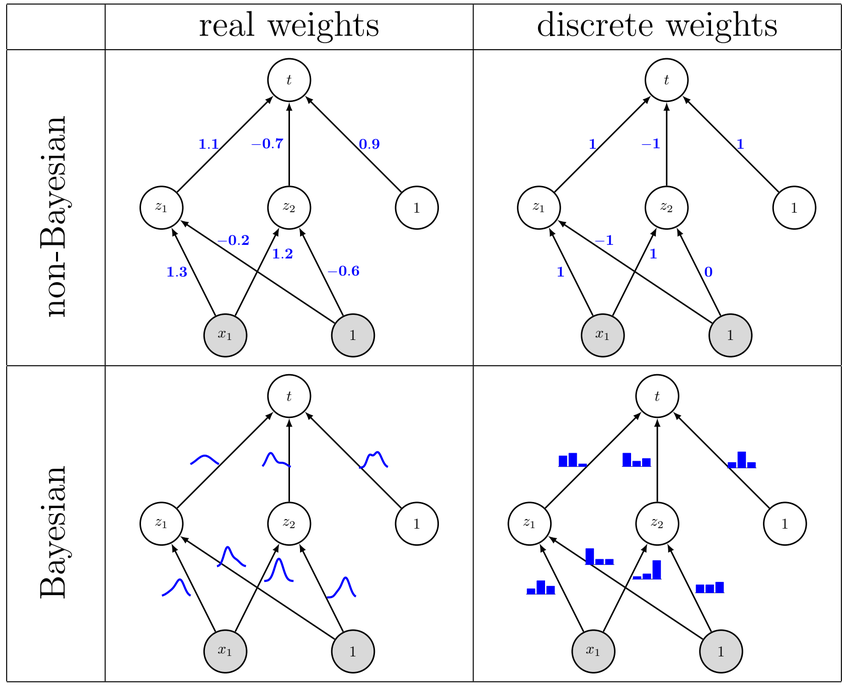


Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2017. ImageNet classification with deep convolutional neural networks. Commun. ACM 60, 6 (May 2017), 84–90. DOI:https://doi.org/10.1145/3065386

DEEPCHIP

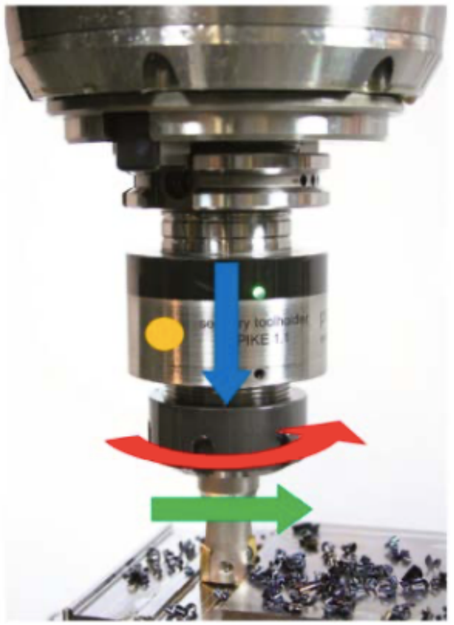


- DL-based speech & image processing for resource-constrained devices
  - No accuracy loss compared to state-of-the-art
  - Reduced precision (quantization), sparsity (pruning) and asynchrony
  - HW-aware compression and mapping methods for ARM, FPGA, etc.
- Collaboration with SPSC group @ TU Graz
- Next: robustness & dependability
  - From toy problems to real-world problems
  - Noice, uncertainty, limited data, …



http://www.deepchip.org

OTHER RESEARCH EFFORTS



Non-parametric and parametric models for industrial condition monitoring (FWF COMET)

- Code generation for quantized neural networks

Energy-proportional interconnection networks (Carl-Zeiss Stiftung)

- Power-saving states for networks

FPGA as memory access pattern accelerator (SAP)

- Predicting, prefetching, unstring

WRAPPING UP

WRAPPING UP

Moore's law will deliver, Dennard scaling is failing

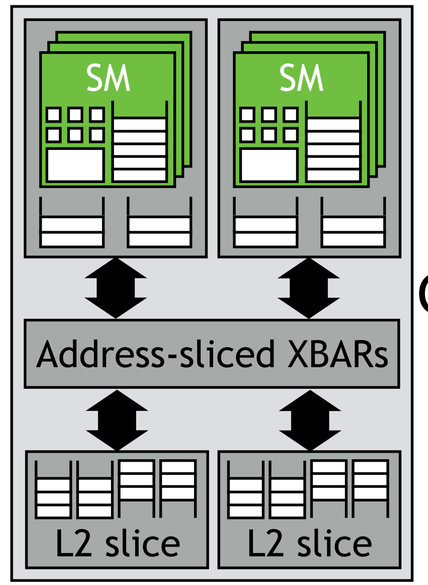
- Anticipated growth in device count for another 5-10 years
- No power budget, though

Post Dennard Regime 1 => **structure**

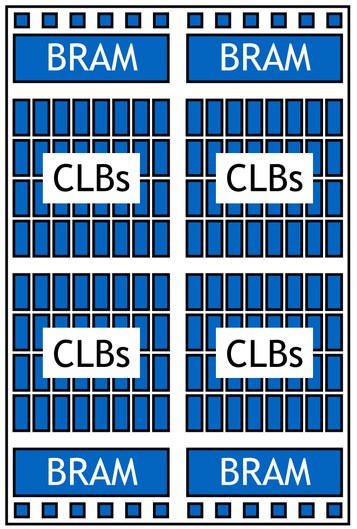
Post Dennard Regime 2 => **locality & predictability**

Our main research concept: **app+data=architecture**

- In spirit of [1]
- picoJoule replaces nanosecond (there be dragons)



GPU



FPGA

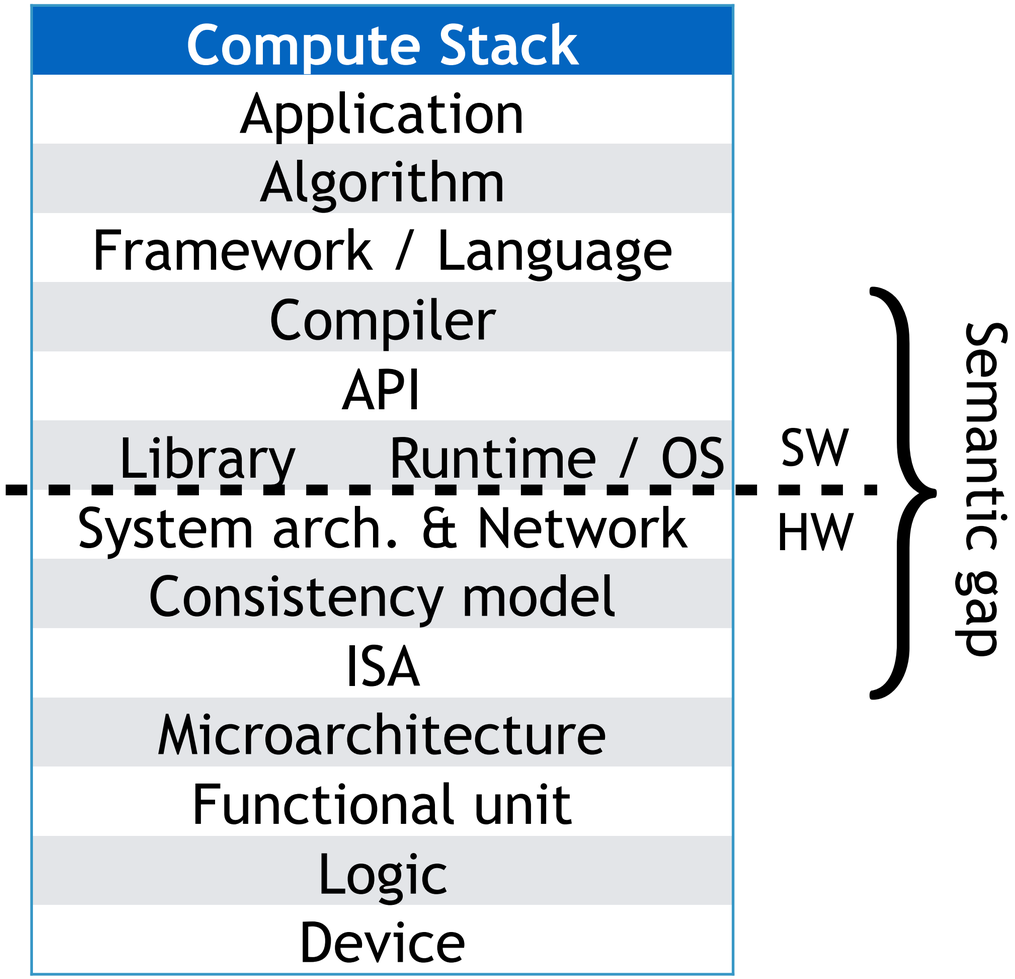
[1] N. Wirth. *Algorithms + Data Structures = Programs*. Prentice-Hall, 1976.

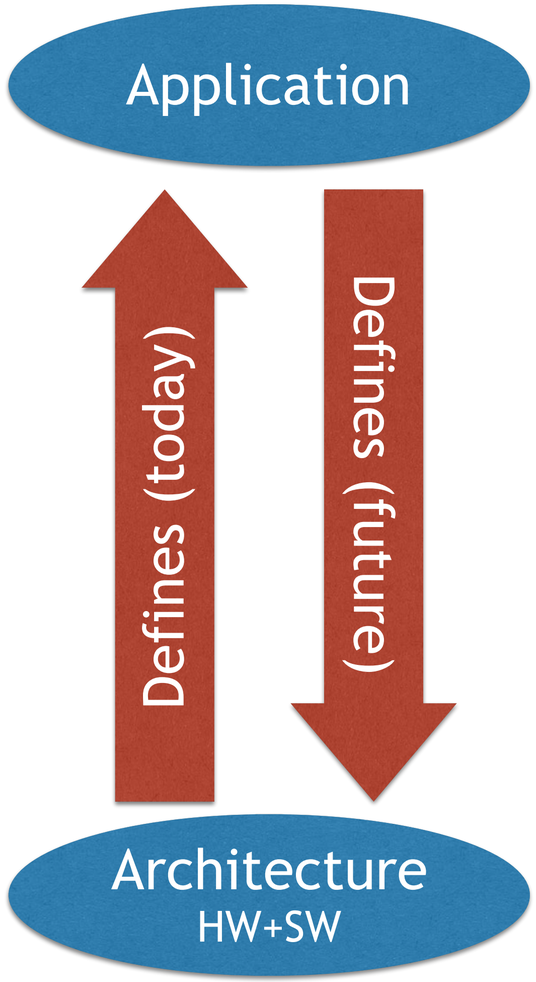
APP + DATA = ARCHITECTURE

Heidelberg University
Natural & formal sciences
Applied computer sciences

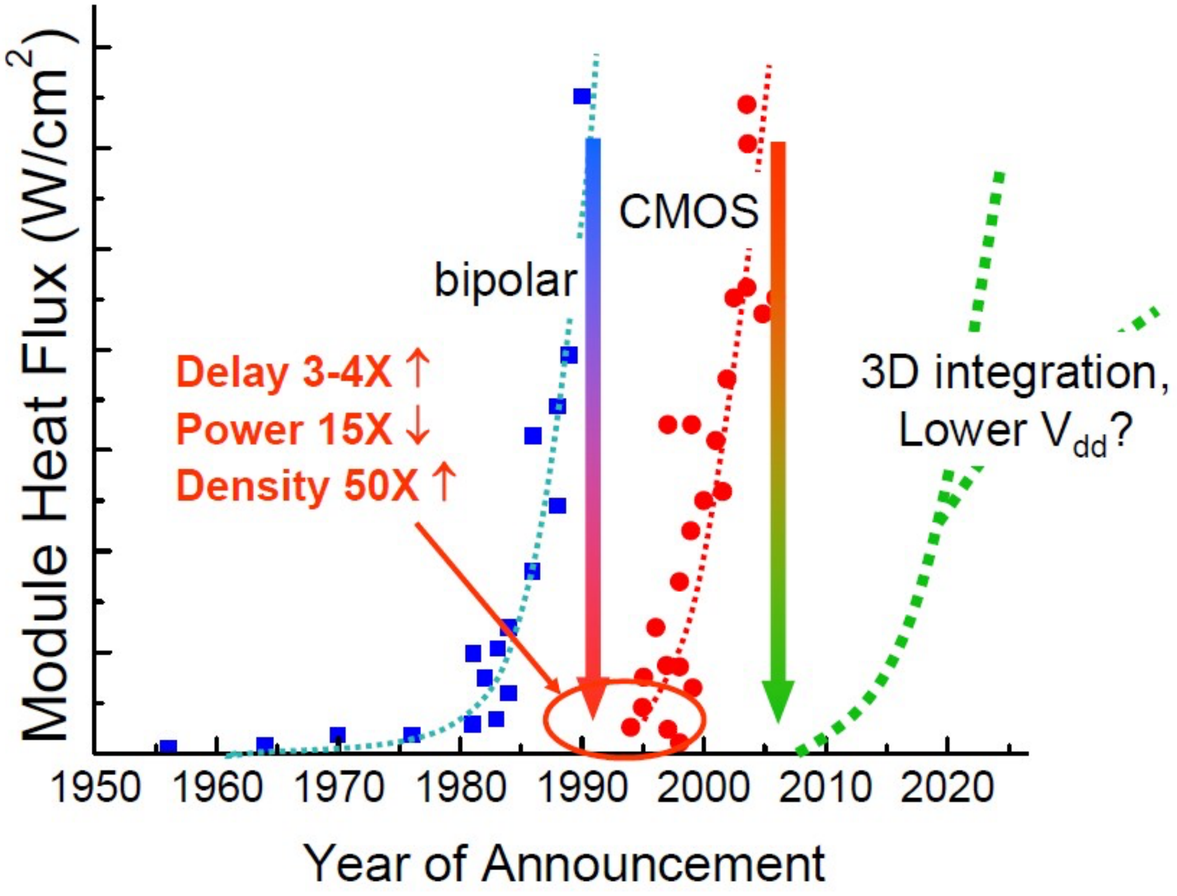
Bridging the gap
Specialization & heterogeneity, HW/SW co-design, beyond-CMOS, programmability

Heidelberg University
(Computational) physics
EINC/CAM & industry





HISTORY REPEATS - OR NOT?



After: R. Schmidt and B. D. Notohardjono. 2002. High-end server low-temperature cooling. IBM J. Res. Dev. 46, 6 (Nov. 2002), 739-751. DOI=http://dx.doi.org/10.1147/rd.466.0739

Candidates for Beyond-CMOS :)

- Approximate/near-threshold-voltage computing
- Cryogenic computing
- Quantum computing
- Neuromorphic computing
- ...

No general-purpose replacement :(

Partially or completely disruptive to the compute stack :(