Microsoft

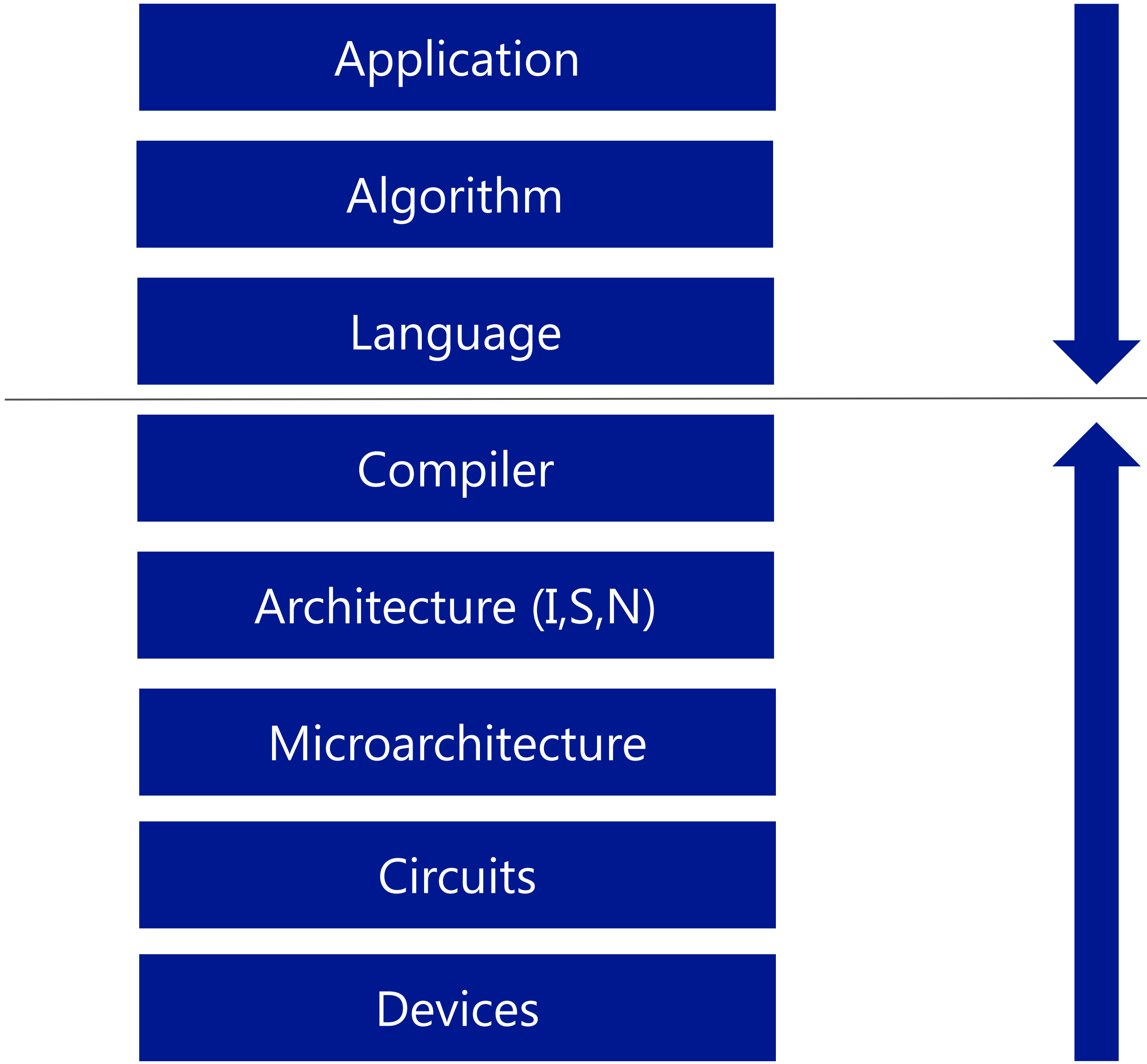
Microsoft Research Faculty Summit 2013

Microsoft

The Beast from Below

Doug Burger
Director, Client and Cloud Apps
Microsoft Research

Microsoft Research
Faculty Summit



Hardware complexity largely hidden via stable abstractions and interfaces

Continuous improvements in all of the lower layers has created consistently large gains in performance

The glory of Moore's Law

The experts look ahead

Cramming more components onto integrated circuits

With unit cost falling as the number of components per circuit rises, by 1975 economics may dictate squeezing as many as 65,000 components on a single silicon chip

By Gordon E. Moore

Director, Research and Development Laboratories, Fairchild Semiconductor division of Fairchild Camera and Instrument Corp.

The future of integrated electronics is the future of electronics itself. The advantages of integration will bring about a proliferation of electronics, pushing this science into many new areas.

Integrated circuits will lead to such wonders as home computers—or at least terminals connected to a central computer—automatic controls for automobiles, and personal portable communications equipment. The electronic wristwatch needs only a display to be feasible today.

But the biggest potential lies in the production of large systems. In telephone communications, integrated circuits in digital filters will separate channels on multiplex equipment. Integrated circuits will also switch telephone circuits and perform data processing.

Computers will be more powerful, and will be organized in completely different ways. For example, memories built of integrated electronics may be distributed throughout the machine instead of being concentrated in a central unit. In addition, the improved reliability made possible by integrated circuits will allow the construction of larger processing units. Machines similar to those in existence today will be built at lower costs and with faster turn-around.

Present and future

By integrated electronics, I mean all the various technologies which are referred to as microelectronics today as well as any additional ones that result in electronics functions supplied to the user as irreducible units. These technologies were first investigated in the late 1950's. The object was to miniaturize electronics equipment to include increasingly complex electronic functions in limited space with minimum weight. Several approaches evolved, including microassembly techniques for individual components, thin-film structures and semiconductor integrated circuits.

Each approach evolved rapidly and converged so that each borrowed techniques from another. Many researchers believe the way of the future to be a combination of the various approaches.

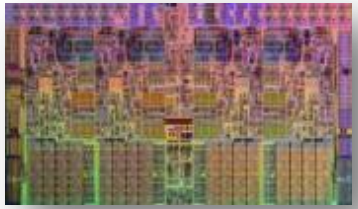
The advocates of semiconductor integrated circuitry are already using the improved characteristics of thin-film resistors by applying such films directly to an active semiconductor substrate. Those advocating a technology based upon films are developing sophisticated techniques for the attachment of active semiconductor devices to the passive film arrays.

Both approaches have worked well and are being used in equipment today.

The author

Dr. Gordon E. Moore is one of the new breed of electronic engineers, schooled in the physical sciences rather than in electronics. He earned a B.S. degree in chemistry from the University of California and a Ph.D. degree in physical chemistry from the California Institute of Technology. He was one of the founders of Fairchild Semiconductor and has been director of the research and development laboratories since 1959.

Electronics, Volume 38, Number 8, April 19, 1965



Intel 4004	**Intel Core i7 980X**	**%/year, Ratios:**
2300 transistors	1.17B transistors	38%, 508000
740 kHz clock	3.33 GHz clock	23%, 4450
10um process	32nm process	15%, 312
10.8 usec/inst	73.4 psec/inst	34%, 147000

Moore’s secret sauce: Dennard scaling

Device or Circuit Parameter	Scaling Factor
Dimension, Tox, L, W	1/k
Doping Concentration Na	k
Voltage (V)	1/k
Current (I)	1/k
Capacitance (eA/t)	1/k
Delay time/circuit (VC/I)	1/k
Power dissipation/circuit (VI)	1/k^2
Power density (VI/A)	1

Historically, k ~= 1.4

[Dennard, Gaensslen, Yu, Rideout, Bassous, Leblanc, **IEEE JSSC**, 1974]

Design of Ion-Implanted MOSFET’s with Very Small Physical Dimensions

ROBERT H. DENNARD, MEMBER, IEEE, FRITZ H. GAENSSLEN, HWA-NIEN YU, MEMBER, IEEE, V. LEO RIDEOUT, MEMBER, IEEE, ERNEST BASSOUS, AND ANDRE R. LEBLANC, MEMBER, IEEE

***Abstract*—This paper considers the design, fabrication, and characterization of very small MOSFET switching devices suitable for digital integrated circuits using dimensions of the order of 1 μ. Scaling relationships are presented which show how a conventional MOSFET can be reduced in size. An improved small device structure is presented that uses ion implantation to provide shallow source and drain regions and a nonuniform substrate doping profile. One-dimensional models are used to predict the substrate doping profile and the corresponding threshold voltage versus source voltage characteristic. A two-dimensional current transport model is used to predict the relative degree of short-channel effects for different device parameter combinations. Polysilicon-gate MOSFET’s with channel lengths as short as 0.5 μ were fabricated, and the device characteristics measured and compared with predicted values. The performance improvement expected from using these very small devices in highly miniaturized integrated circuits is projected.**

Manuscript received May 20, 1974; revised July 3, 1974.
The authors are with the IBM T. J. Watson Research Center, Yorktown Heights, N.Y. 10598.

LIST OF SYMBOLS

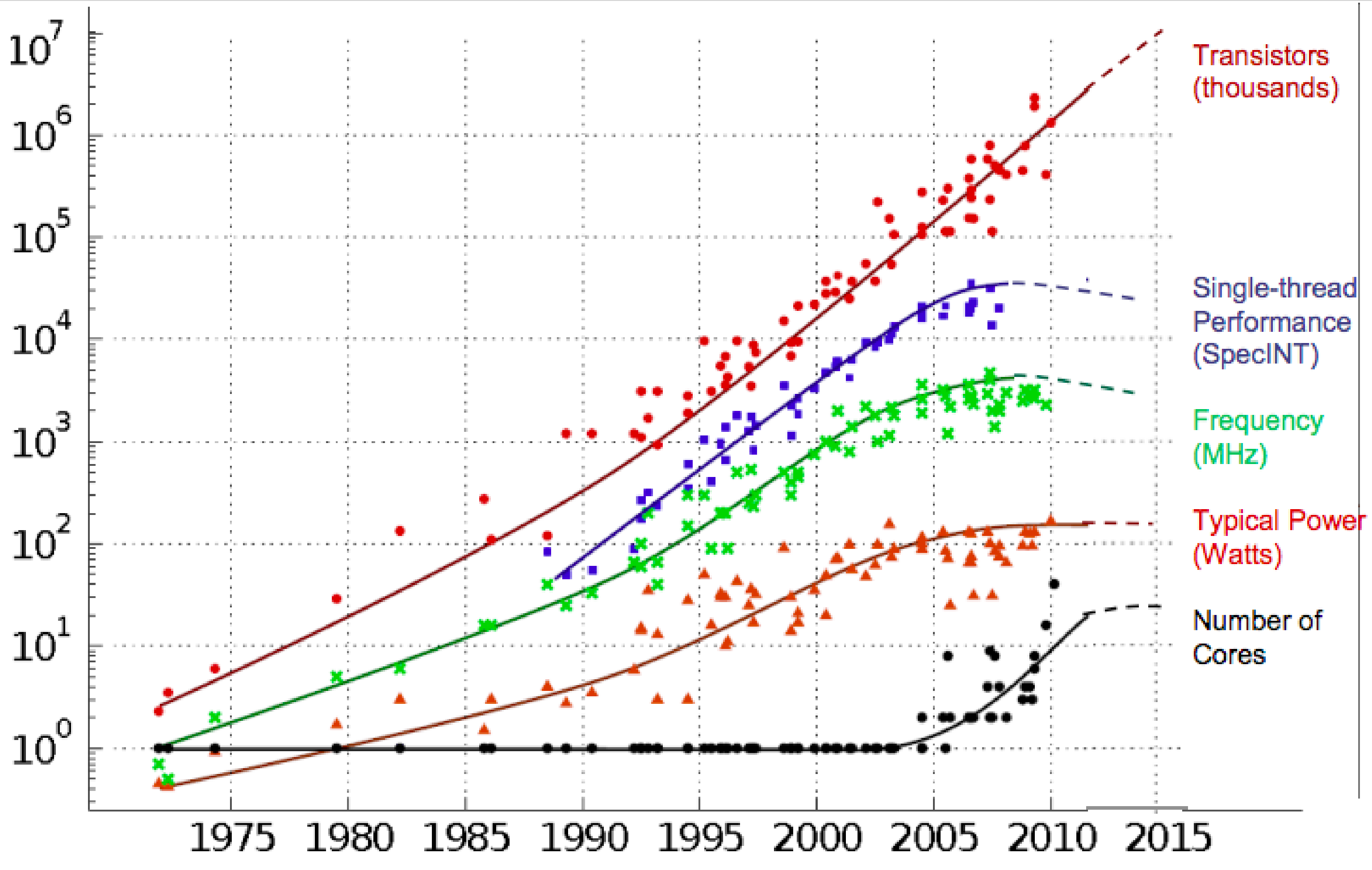
α	Inverse semilogarithmic slope of subthreshold characteristic.
D	Width of idealized step function profile for channel implant.
ΔW_f	Work function difference between gate and substrate.
ϵ_{Si}, ϵ_{ox}	Dielectric constants for silicon and silicon dioxide.
I_d	Drain current.
k	Boltzmann’s constant.
κ	Unitless scaling constant.
L	MOSFET channel length.
μ_{eff}	Effective surface mobility.
n_i	Intrinsic carrier concentration.
N_a	Substrate acceptor concentration.
Ψ_s	Band bending in silicon at the onset of strong inversion for zero substrate voltage.

2x transistor count

40% faster

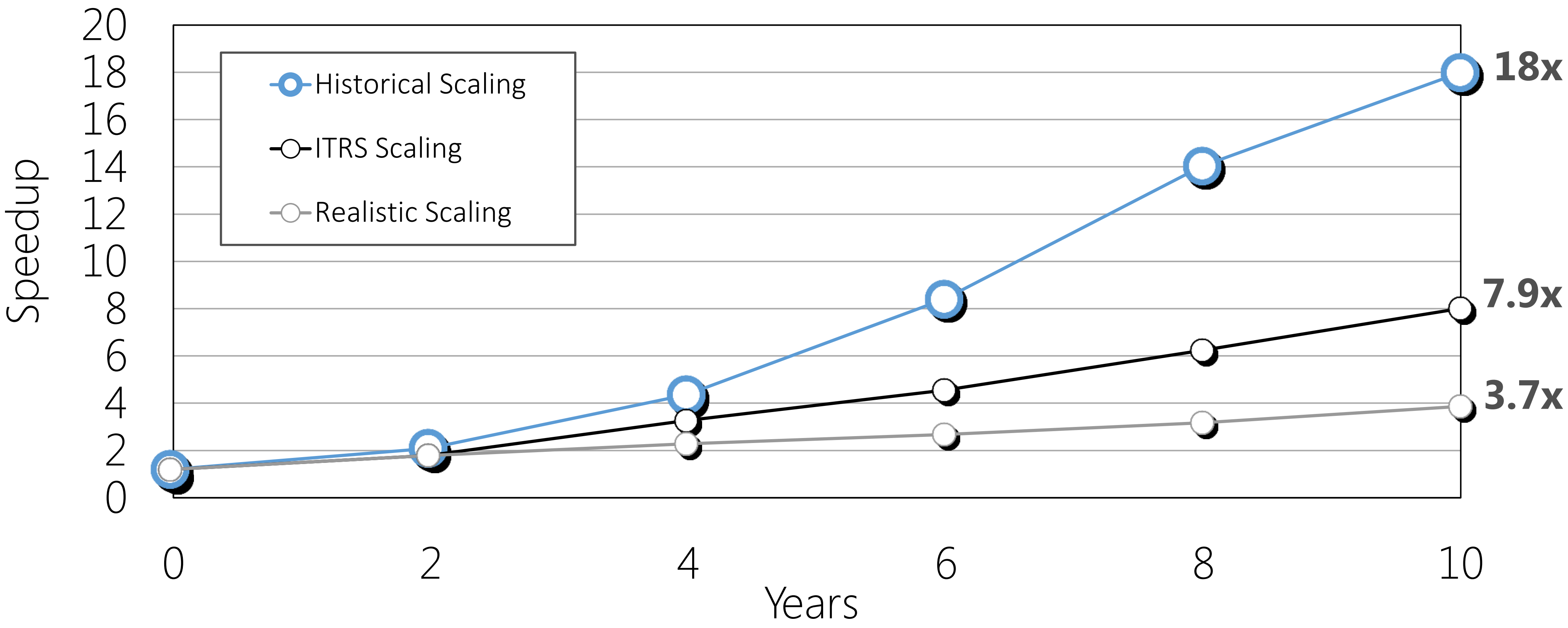
50% more efficient

Dennard scaling is dead



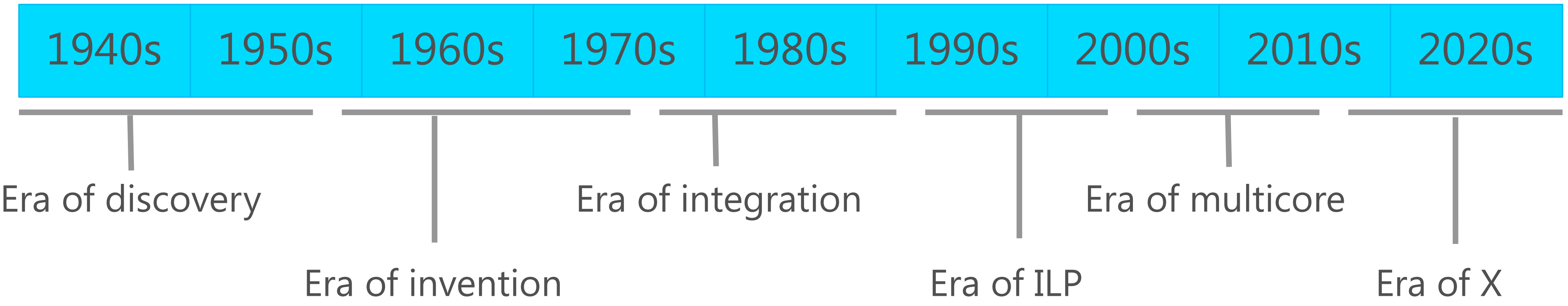
Original data collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond and C. Batten
Dotted line extrapolations by C. Moore

Multicore to the rescue?



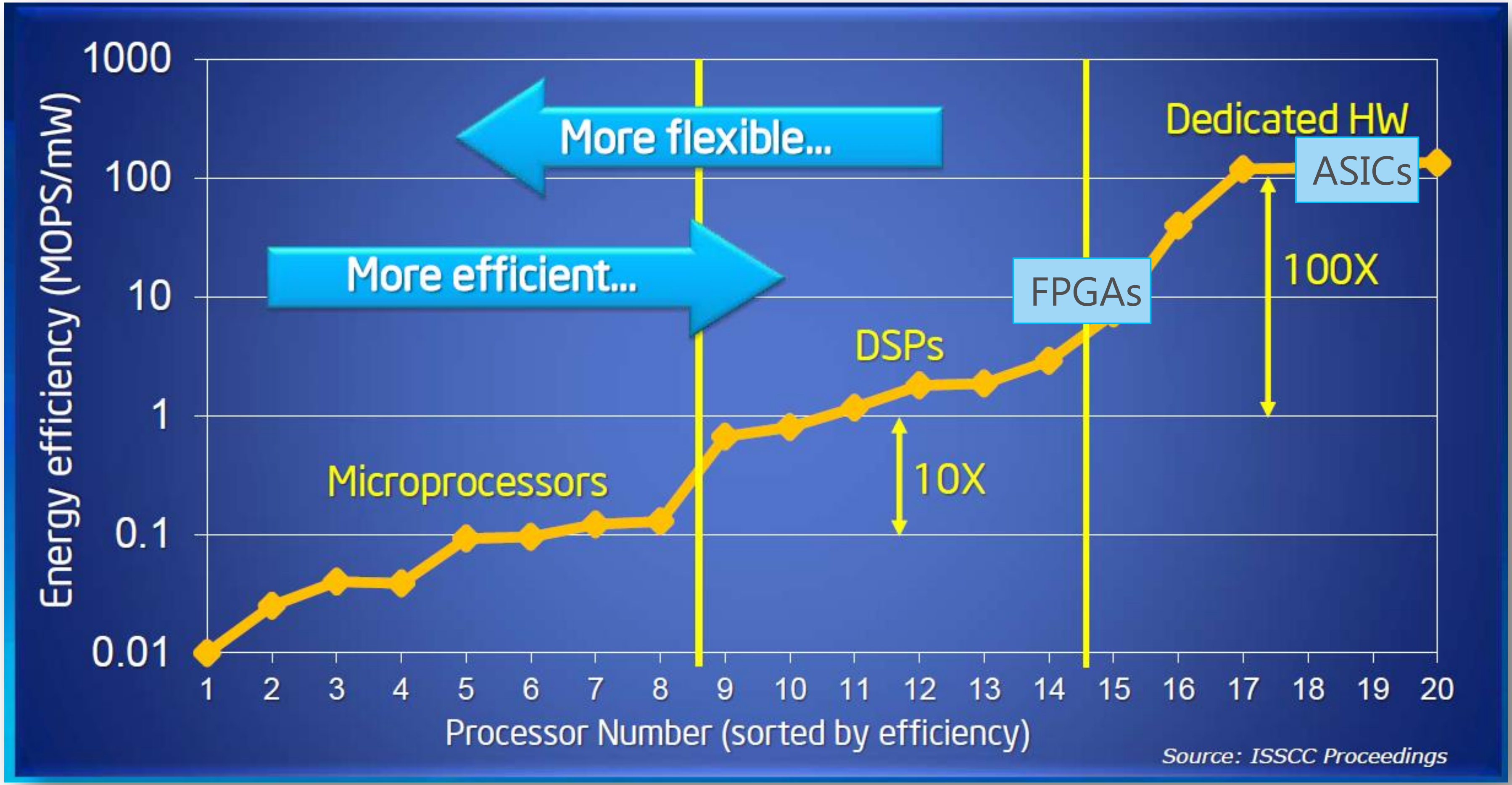
[Esmaeilzadeh, Blem, St. Amant, Sankaralingam, Burger, ISCA 2011]

A brief history of computer architecture



X = {Logic specialization, neural computing, cold computing, ?}

Specialization: A path forward (?)



Source: Bob Broderson, Berkeley Wireless group

More gains the lower you go

Code specialization	10x
Logic specialization	100x
Circuit specialization	1000x
Device specialization	10000x

Logic synthesis as a platform

Large gains in efficiency with direct software-to-logic

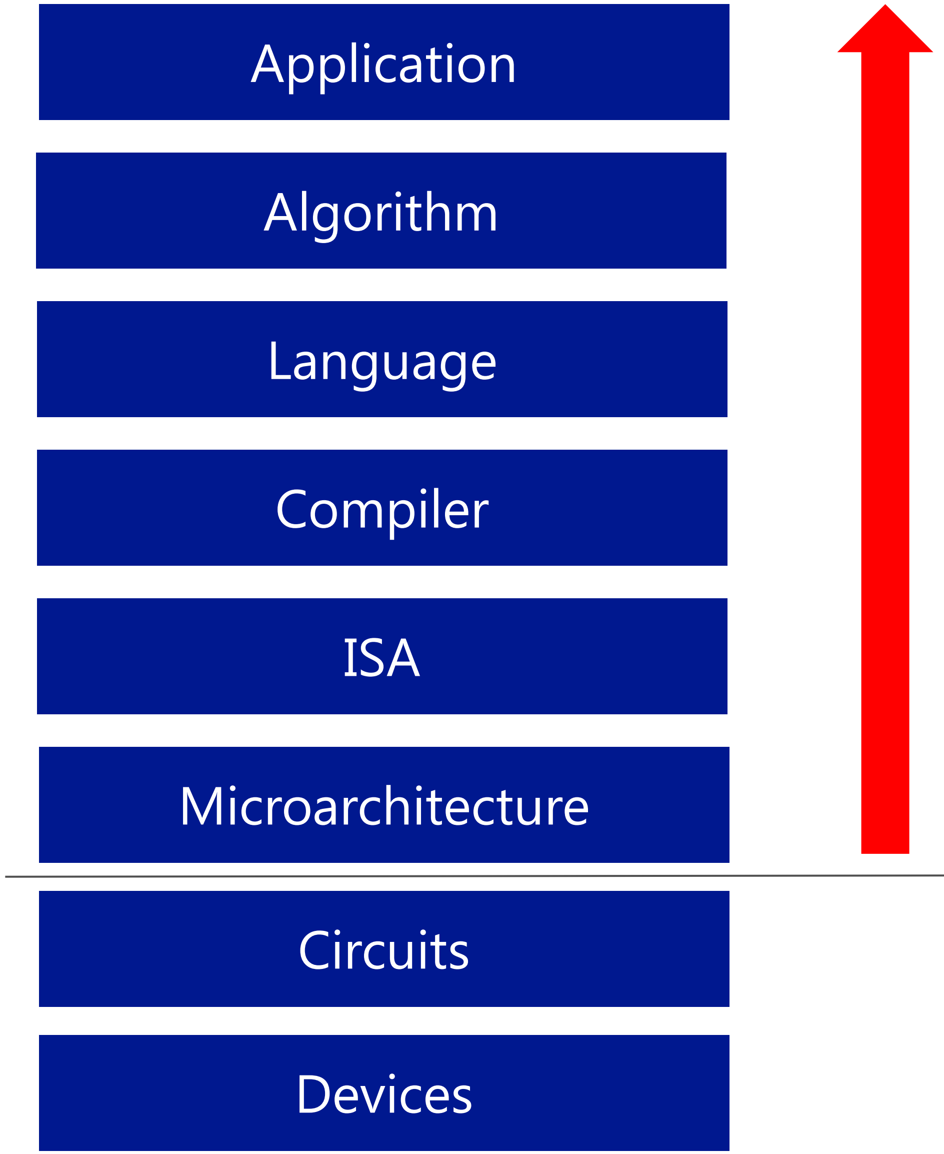
- 100x for FPGA/CGRAs, 1000x for ASICs

Development and compilation is a huge challenge

- Mix of cores, hard IP blocks, and tools to target them (AutoESL, OpenCL)
- Map common operations and flows into libraries, compose them

Will see growing adoption, increased tool investment

- Initially FPGAs in the cloud, CGRAs/ASICS in the client



Logic specialization has great potential in the short term since it is compatible with the higher levels of the stack, but leaves several levels unaddressed

Generality
(CPUs)

Efficiency
(ASICs)

How do we resolve this tension?

An end to Moore's Law

High Volume Manufacturing	2008	2010	2012	2014	2016	2018	2020	2022
Technology Node (nm)	45	32	22	16	11	8	6	4
Integration Capacity (BT)	8	16	32	64	128	256	512	1024

Source: Shekhar Borkar, Intel Corporation

Approximate computing

Changing workloads offer an opportunity

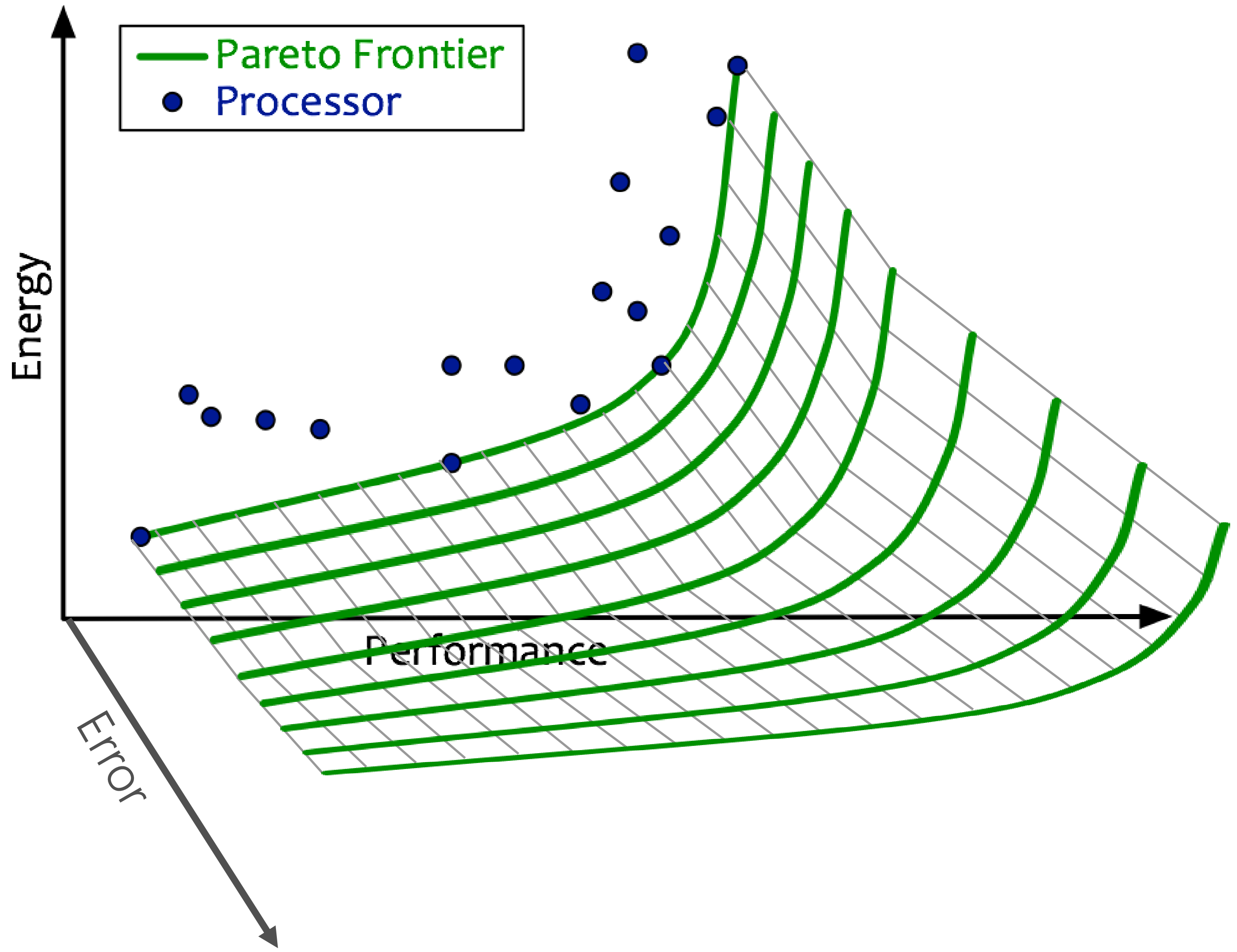
- Large-scale machine learning
- Computer vision
- Bioinformatics
- Mining big data
- Speech and AI

Robust to reduced precision

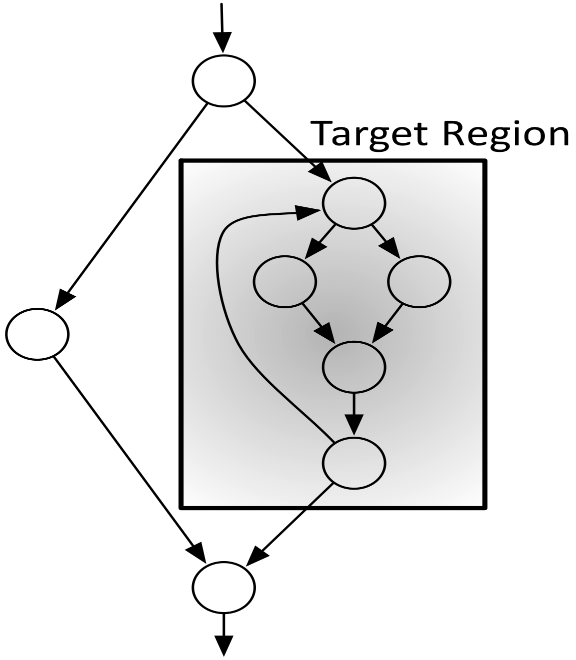
- Need formal semantics to reason about and bound error
- Need to handle dynamic noise and variations

Thanks to Luis Ceze, Hadi Esmaeilzadeh, Adrian Sampson, and others

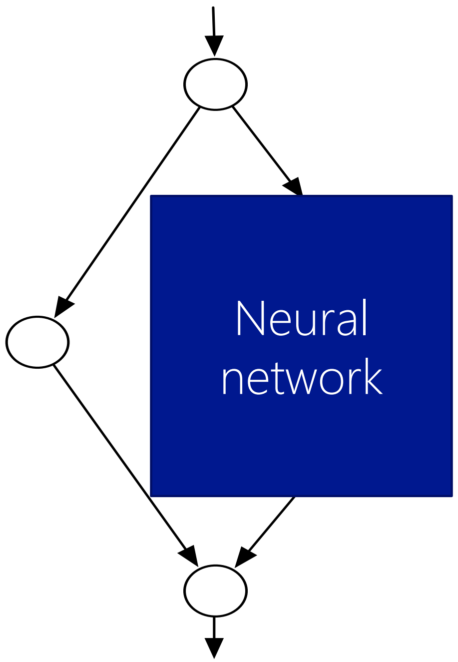
Approximate computing



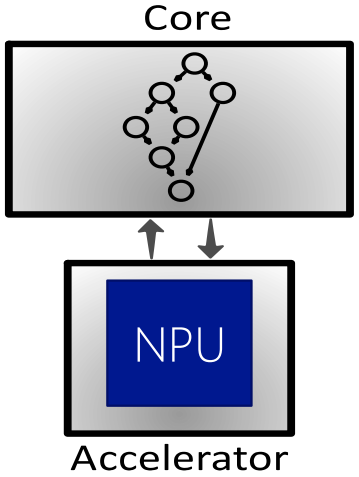
Transforming von Neumann to NNs



Imperative
code



Transformed
code



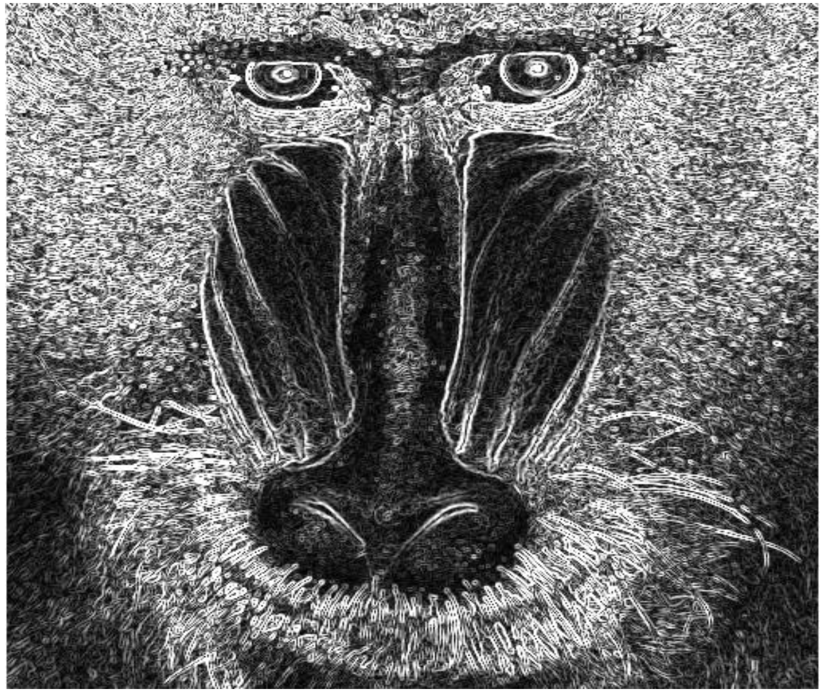
Accelerated
execution

[Esmaeilzadeh, Sampson, Ceze, and Burger, MICRO 2012]

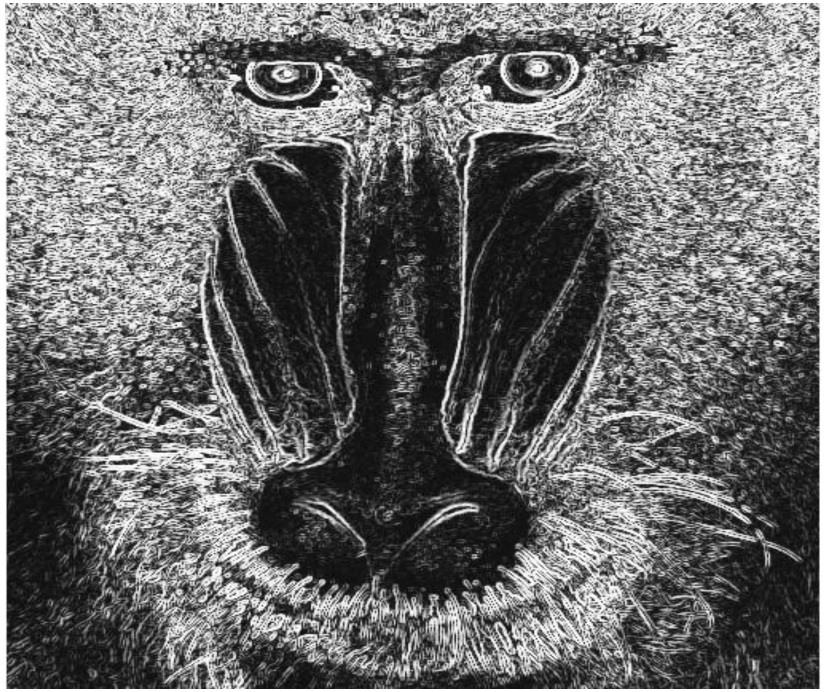
Applications using Neural Transformation

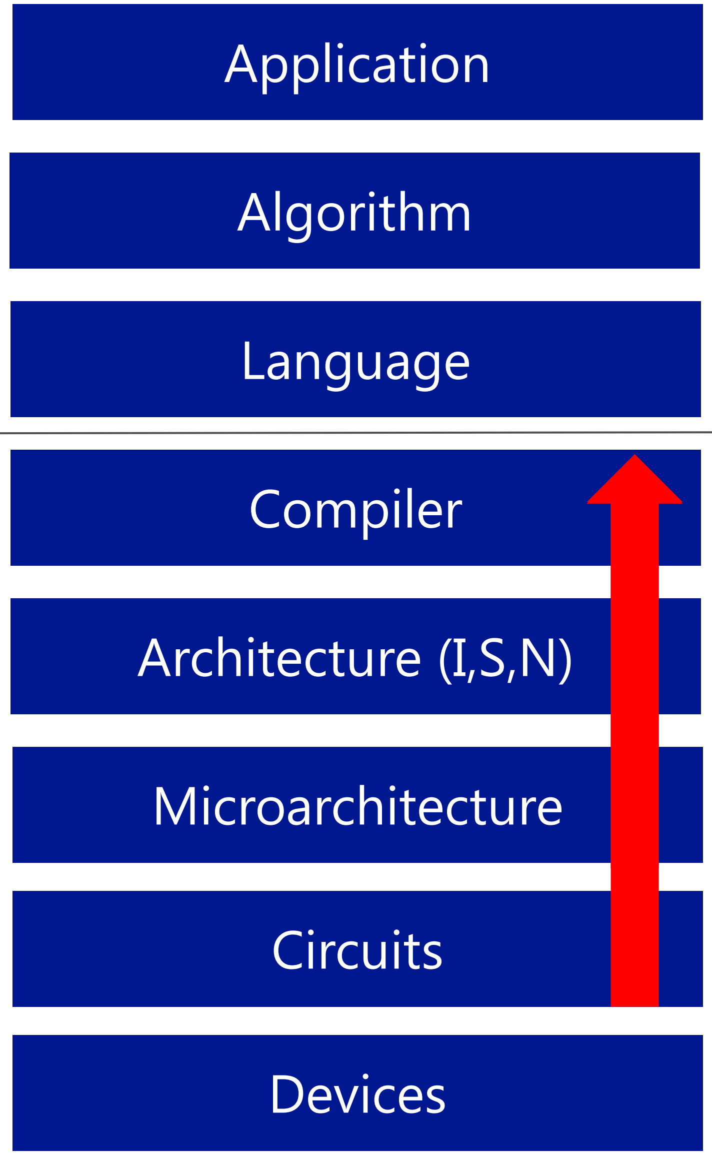
- Signal Processing
  - fft
- Robotics
  - inverse2kj
- 3D graphics
  - jmeint
- Compression
  - jpeg
- Machine Learning
  - kmeans
- Image processing
  - sobel

ORIGINAL CODE

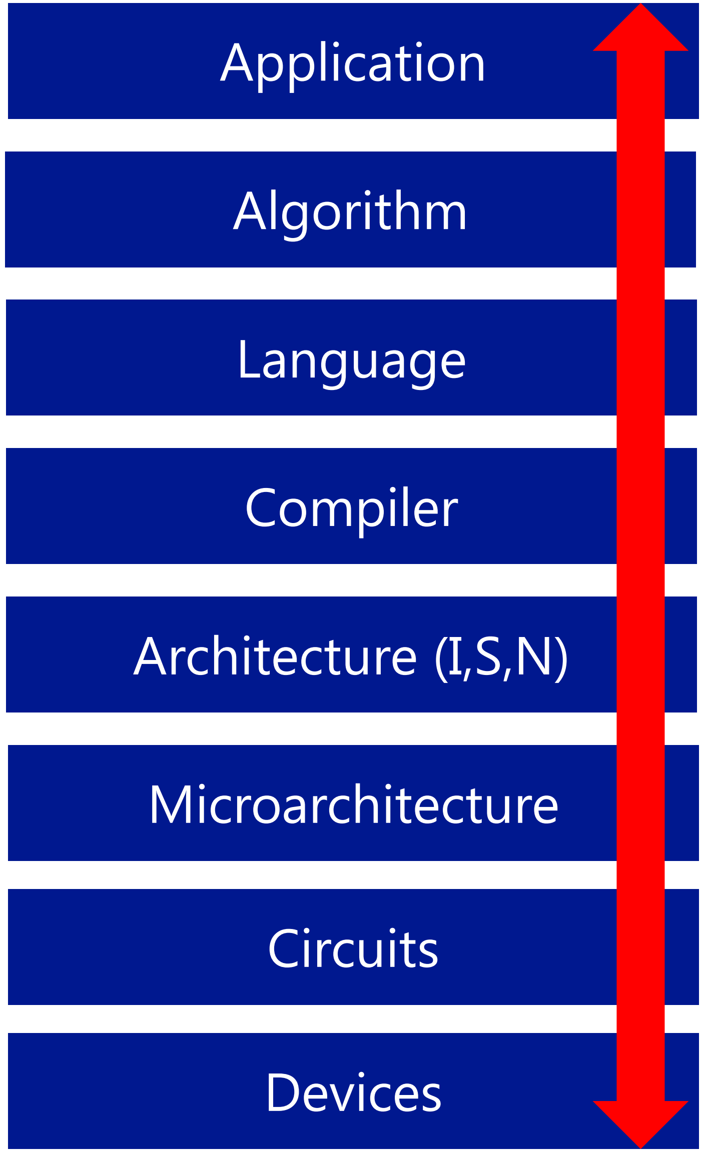


NPU-TRANSFORMED CODE





Gains from NPUs still limited by being hidden behind standard abstractions



New execution models needed:

Hardware synthesis
Neural
Quantum
Cellular automata?
Chemical?
Bayesian?

Neural Networks as a Platform

The seeds of widening efforts are growing

Many types of networks

- Artificial Neural Networks
- Bio-inspired Neural Networks (DNNs)
- Bio-abstracted Neural Networks

A recent quote from Jim Smith

- "Giants are walking the Earth today, but we don't yet know whom they are"

Five predictions for 2025

1. Moore's Law will be dead
2. Hardware/software compilation will be common
3. Neural execution will start to have a complete stack
4. Physics-based computation will be a hot topic
5. Machines will beat humans at many more tasks

MOSFET (N-type) Transistor Operation