Story ……..

…. From a Faraway Time & Place

1990

1990 | | | | | | | | 2013

What Is The X Axis?

- Time?
- Heterogeneity?
- Compute Power?

- Screen Resolution?
- Wireless BW?
- Weight?

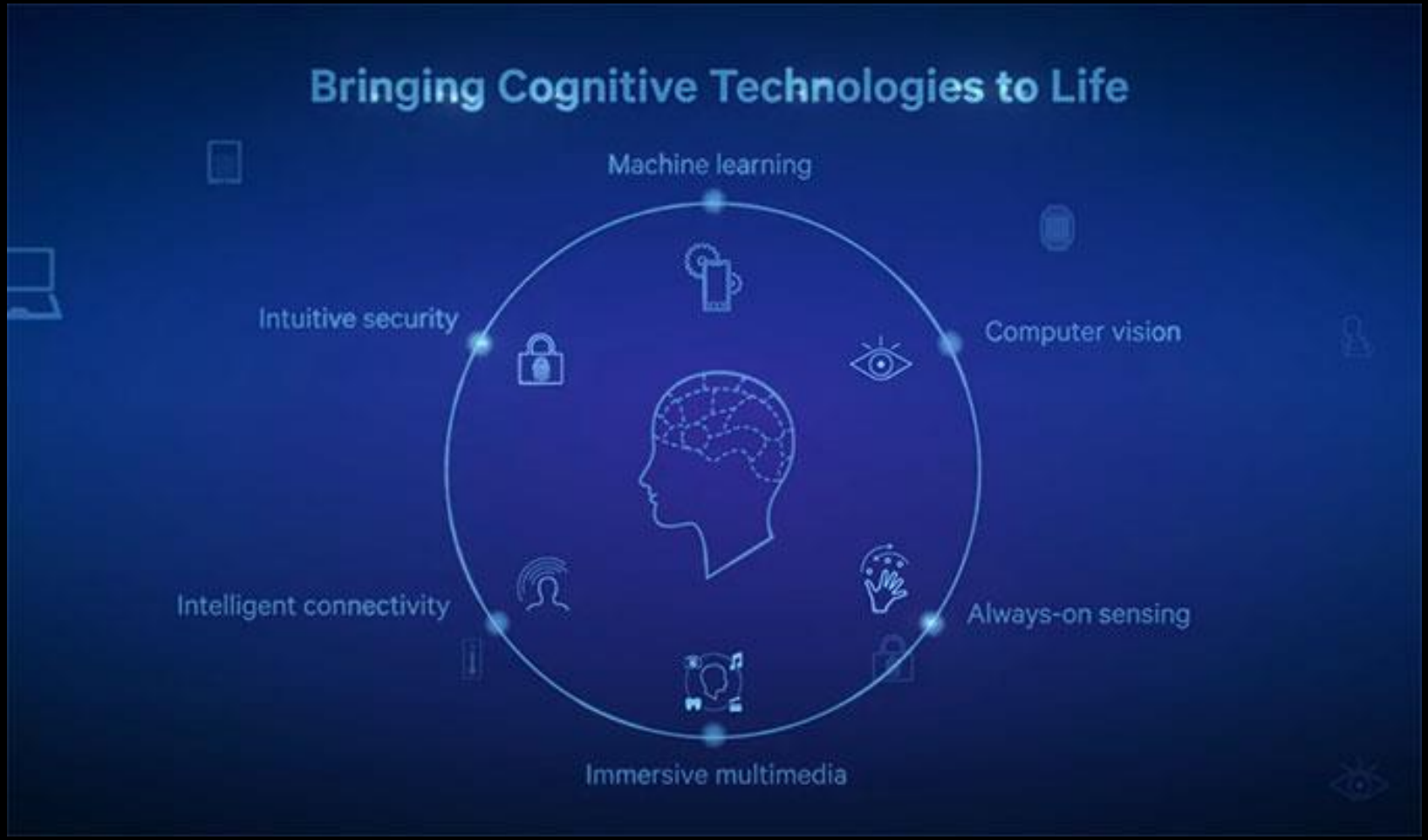
Compare Apples to Apples

Height: 381mm	Height: 115.2mm
Width: 381mm	Width: 58.6mm
Depth: 435mm	Depth: 9.3mm
Weight: 15,800g	Weight: 137g
Price: £1,500	Price: £599
CPU: 500MHz	CPU: 1GHz
RAM: 128MB	RAM: 512MB
Display: 1024 x 768	Display: 960 x 640
Storage: 30GB	Storage: 32GB

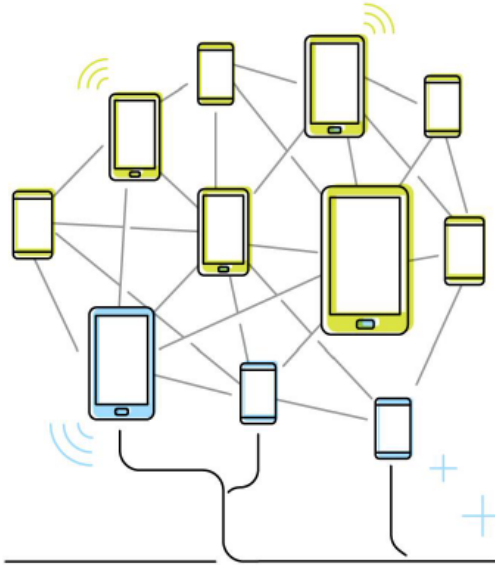
Different Platform

Tomorrow?

Today


Bringing Cognitive Technologies to Life
Machine learning
Intuitive security
Computer vision
Intelligent connectivity
Always-on sensing
Immersive multimedia

Brains at the edge: machine learning goes mobile



Deloitte Global predicts that in 2017 over

300 million smartphones

(more than **a fifth of units sold**)

will have on-board neural network machine-learning capabilities

This will allow smartphones to perform machine-learning tasks

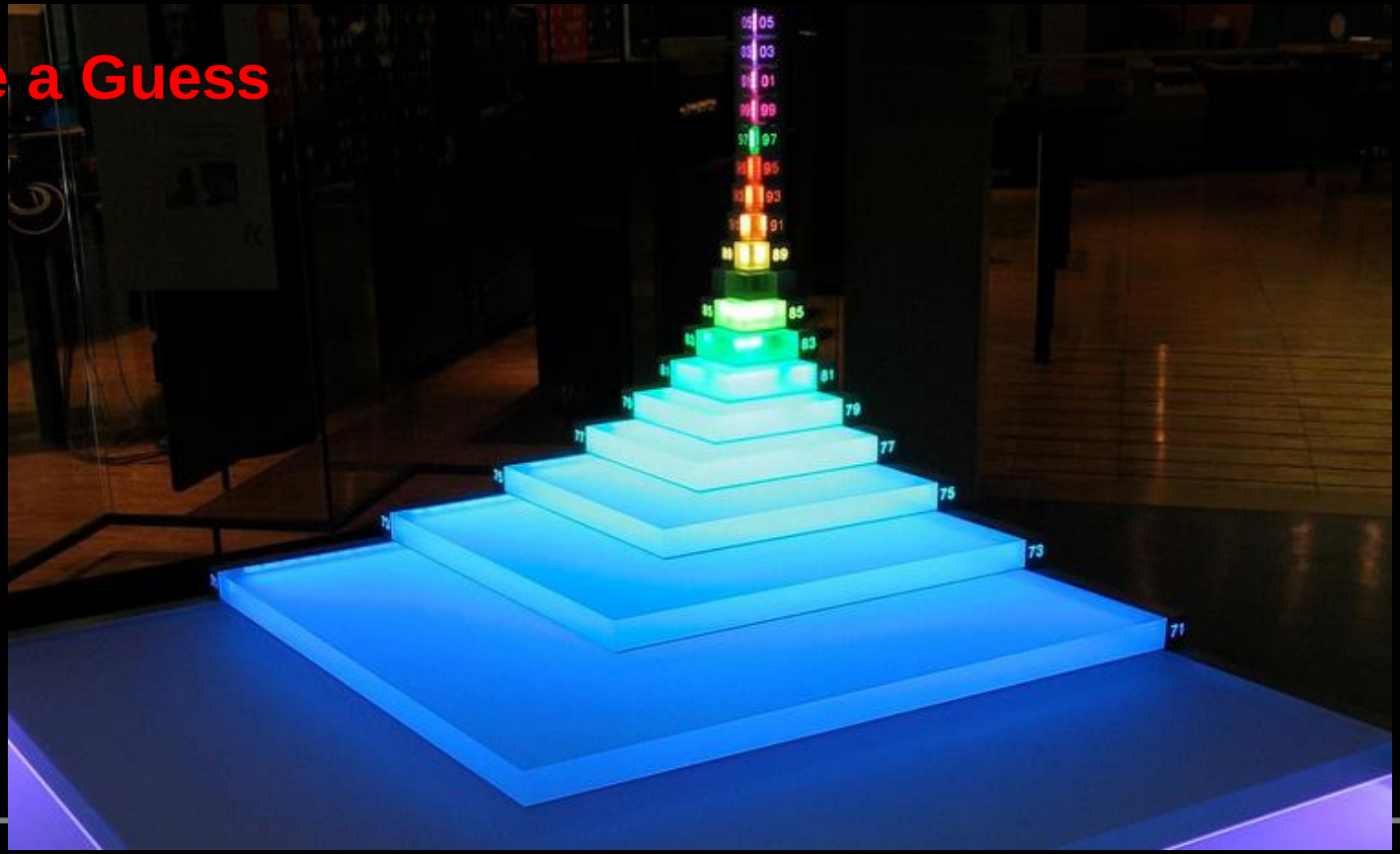
This functionality will enhance applications including:

- indoor navigation
- augmented reality
- language translation
- image classification
- speech recognition
- and many more currently unknown applications

How Did We Come This Far?

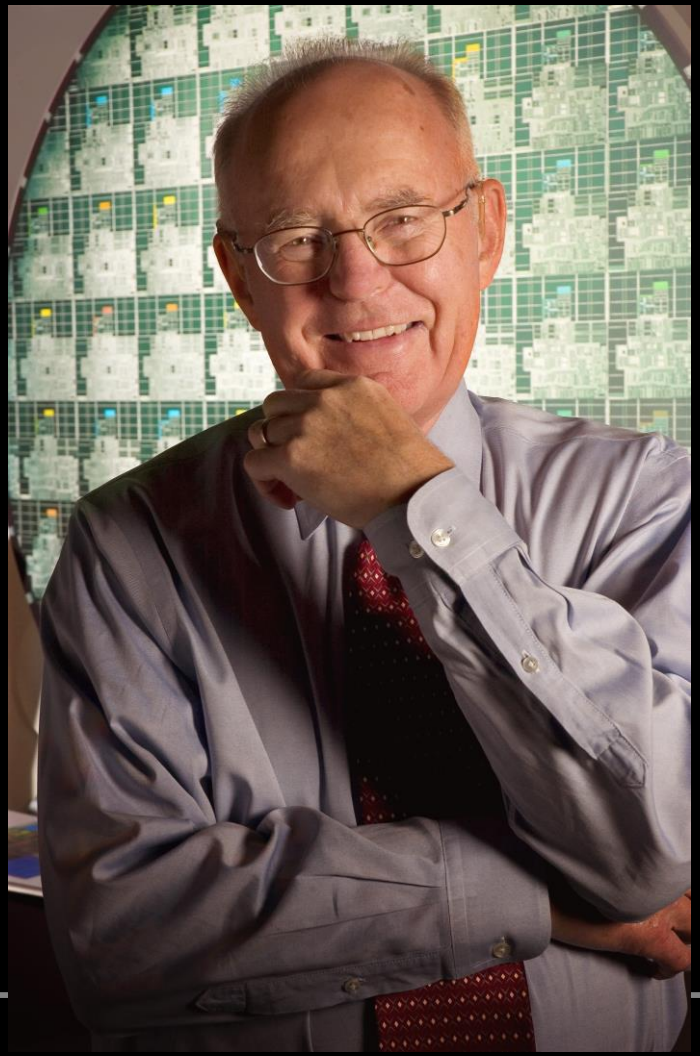
Somewhere in Germany …

Take a Guess

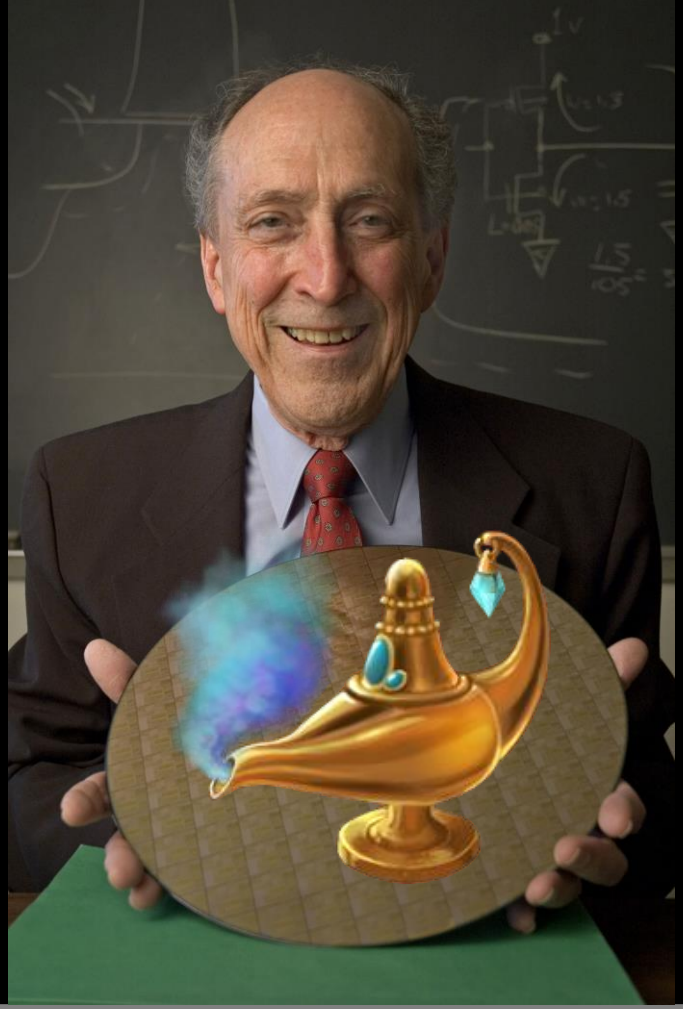


Hint: Gesetz = Law

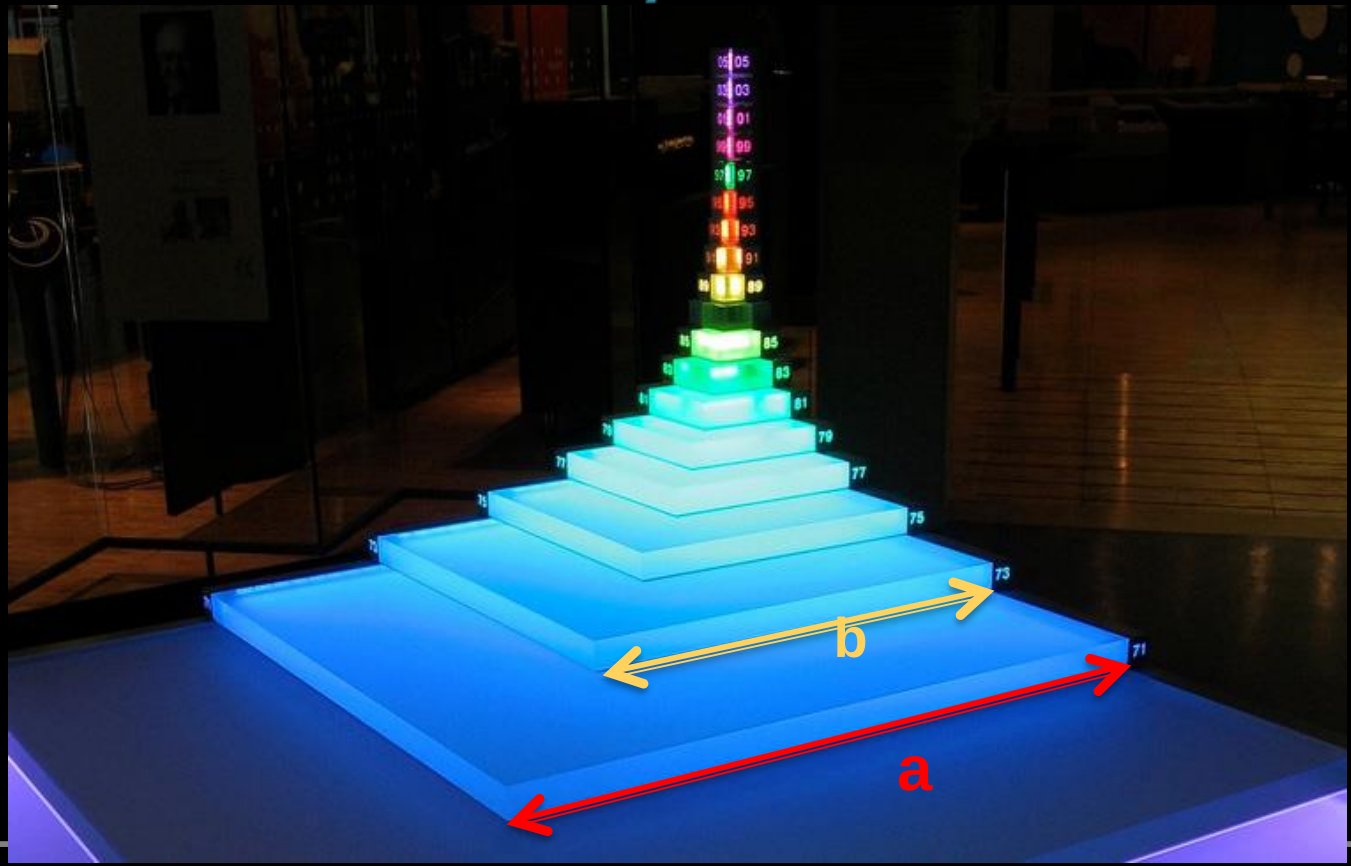
"Frankly I did not expect to be this precise"



Invented DRAM

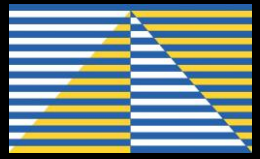


Dennard's Factor α=a/b



Scaling Magic

More Transistors



(μ)Architecture Techniques

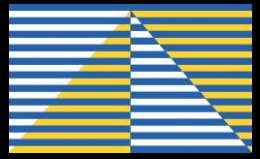
Branch Predictors
OOO Executions
Superscalar
Larger Cache
Pipelining

Faster Transistors

CV^2f → α Scaling Factor → $\alpha^2 (C/\alpha (V/\alpha)^2 \alpha f)$

Scaling Magic

More Transistors



(μ)Architecture Techniques

Higher Performance

CV^2f → α Scaling Factor → $\alpha^2 (C/\alpha (V/\alpha)^2 \alpha f)$

Power Dissipation ≈ CV^2F

- C: Capacitance
  - Wire Length
  - Transistor Size
  - Switching Delay
- V: Supply Voltage
- F: Frequency
  - Clock Frequency

Who Moved My Cheese?

Parameter	Dennard's Factor	500nm→350nm	90nm→65nm
Dimension ***Tox, L, W***	**1/α**	**0.7**	**0.7**
Voltage ***V***	**1/α**	**0.7**	**1**
Current ***I***	**1/α**	**0.7**	**1**
Capacitance ***C***	**1/α**	**0.7**	**0.7**
Delay ***(VC/I) RC***	**1/α**	**0.7**	**0.7**
Power Dissipation ***(VI)***	**$1/\alpha^2$**	**0.5**	**1**
CMOS Power ***CV^2f***	**$1/\alpha^3$**	**0.35**	**0.7**

$CV^2f \Rightarrow \alpha^2 (C/\alpha (V/\alpha)^2 \alpha f) = CV^2f$

$CV^2f \Rightarrow \alpha^2 (C/\alpha V^2 f) = \alpha CV^2f$

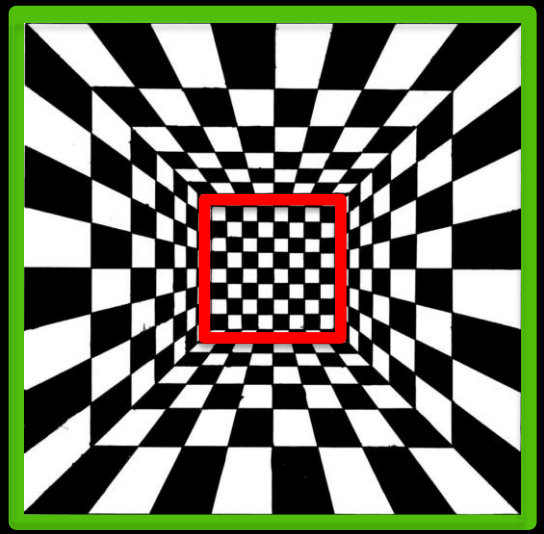
The Four Horsemen of Dark Silicon

Michael Taylor

The Four Horsemen of Dark Silicon

The Four Horsemen of Dark Silicon

α^2 (C/α $V^2\alpha f$)



The Four Horsemen of Dark Silicon

$$\cancel{\alpha^2}\ (C/\alpha\ V^2 \alpha f) = CV^2 f$$

- *Shrink*
  - Limited Performance
  - Diminished Returns

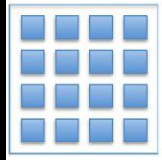
The Four Horsemen of Dark Silicon

$\alpha^2 (C/\alpha (V/d_1)^2 \ f/d_2)$

- ***DIM***
  - Lower Voltage by factor d_1
  - Lower Freq by factor d_2
  - Lower Performance

DIM

22 nm: 16 BCEs

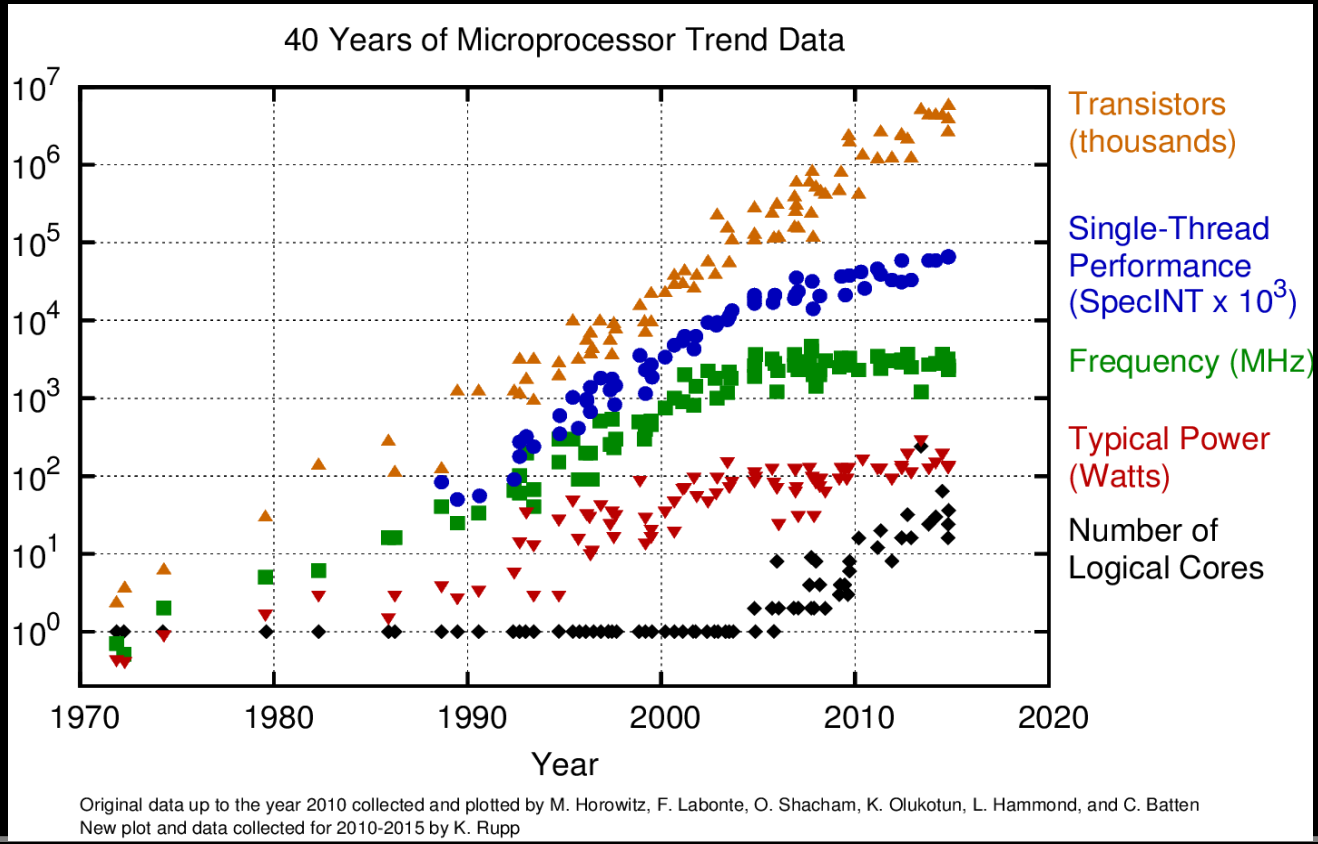


1 000 000 000 transistors

16 RISC cores
@2 GHz
8 MB LLC
Ring or mesh interconnect

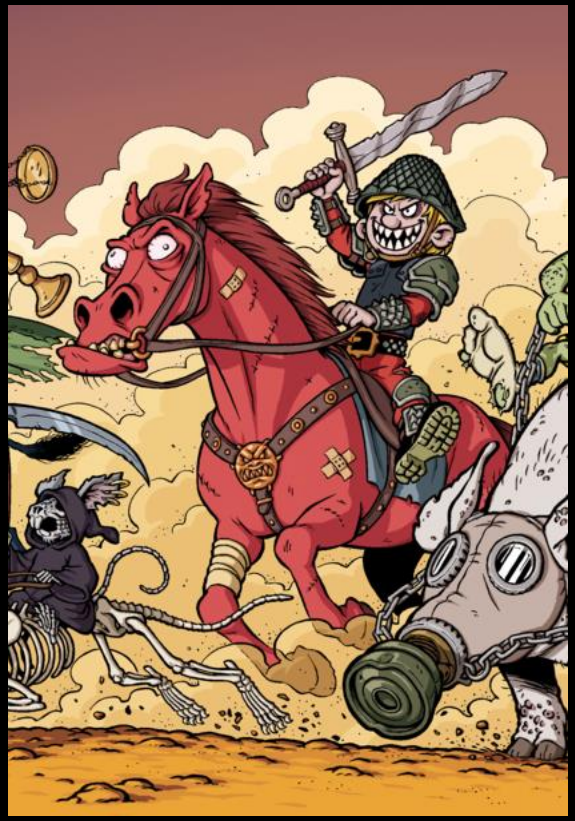
Source:
Christian Märtin

The Rise of Multicore

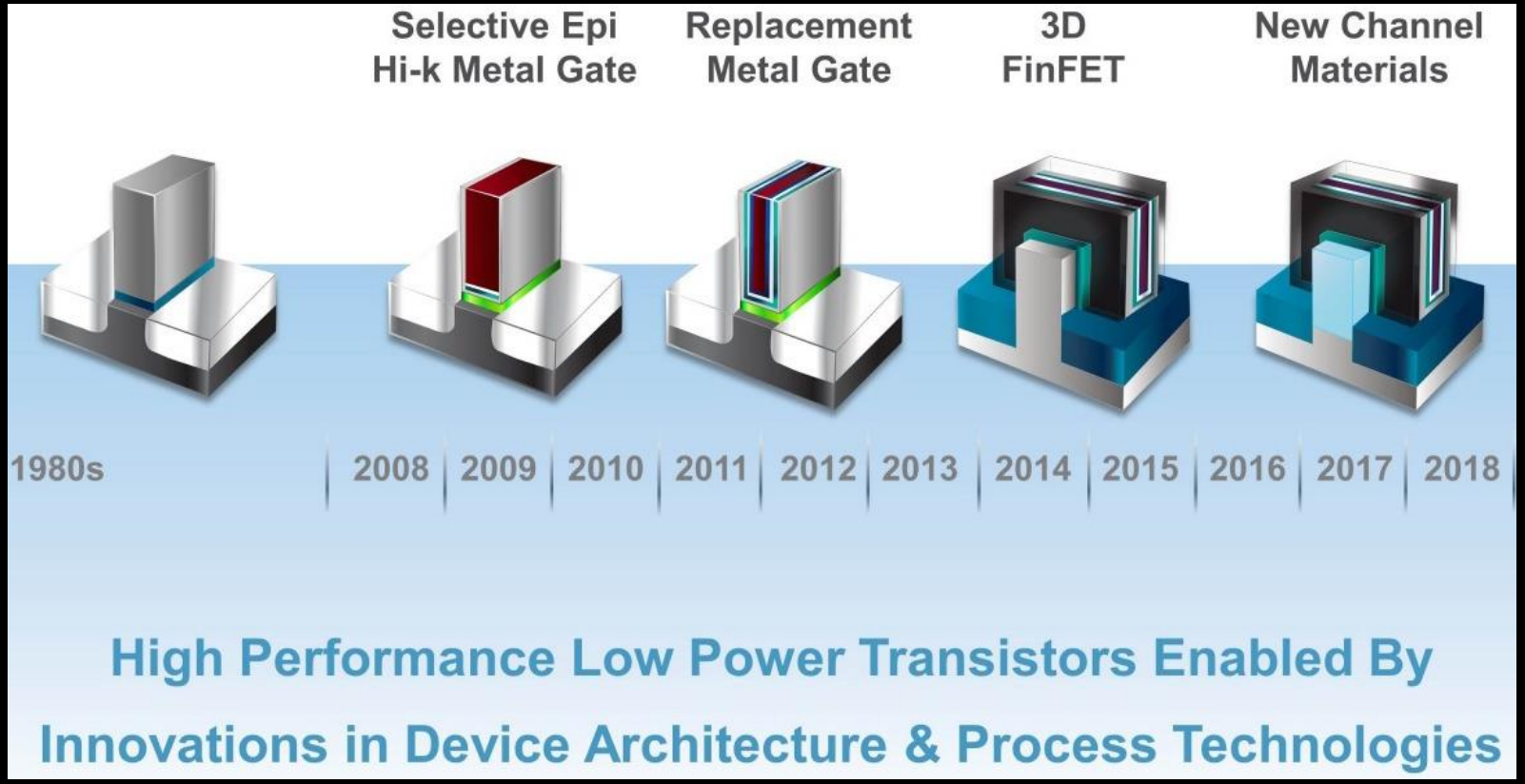


The Four Horsemen of Dark Silicon

- *Dues Ex Machina*
  - Beyond CMOS
- Time to reach the manufacturing scale



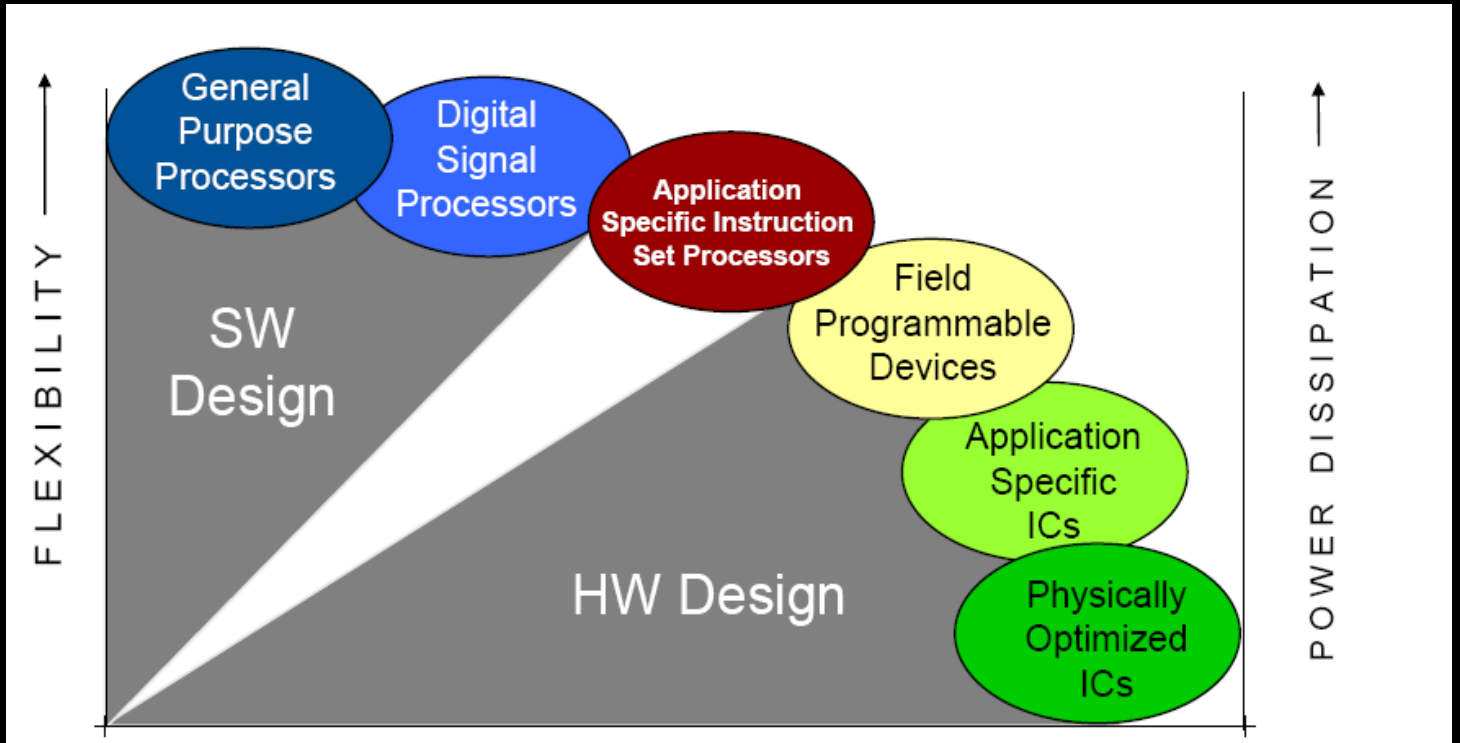
The Transistor Evolution



Source:
Applied Materials®

The Four Horsemen of Dark Silicon

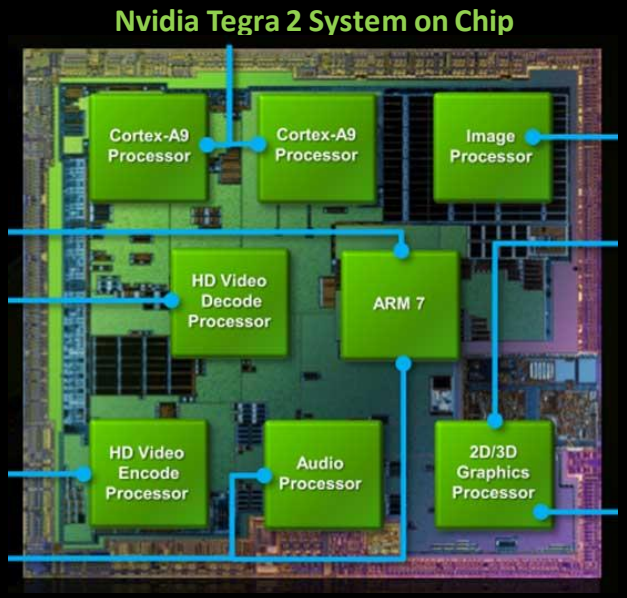
- ***Specialization***
  - Efficient Accelerators
  - Orders of Magnitude better efficiency



Source: T. Noll, RWTH Aachen, via R. Leupers, "From ASIP to MPSoC", Computer Engineering Colloquium, TU Delft, 2006

Heterogeneous Computing

- Opportunity and need for specialization
  - Heterogeneous multi-core system
  - On-chip accelerators
  - GFLOPS/W
  - Energy/(FL)OP

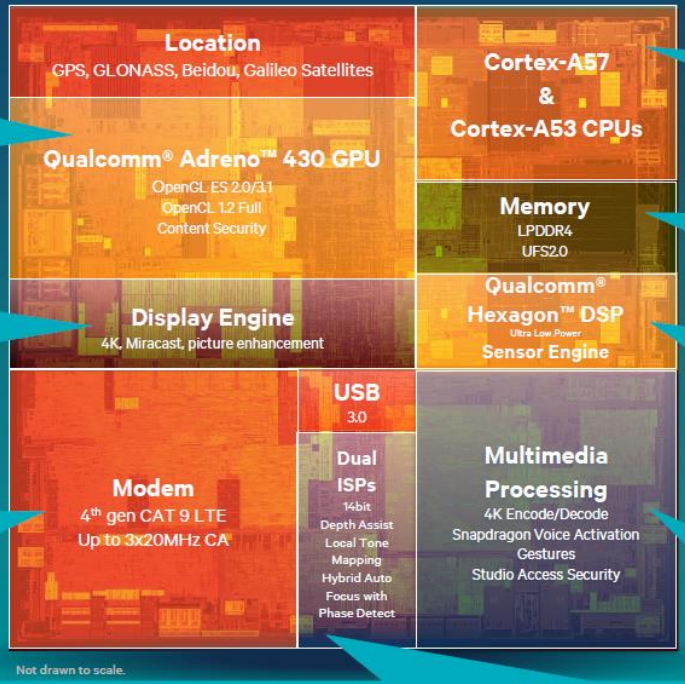


Introducing the Snapdragon 810 Processor

Advanced Graphics & Compute with the Adreno 430 – the best GPU Qualcomm Technologies' has ever made

4K primary & external display support with ecoPix and TruPalette and 3:1 pixel compression

Mobile industry's FIRST announced multi-channel 4G LTE SoC supporting Category 9 Carrier Aggregation



Not drawn to scale.

FIRST Announced ARM®v8-A/64-bit using Cortex®-A57+ Cortex®-A53

Mobile industry's FIRST announced dual channel 1600 MHz LPDDR4 memory

Qualcomm Technologies' FIRST UFS 2.0 Support

Greatly improved power management for DSP/Sensor Engine, Low Power Snapdragon Voice Activation (SVA), 12-channel surround sound decode

Qualcomm Technologies' FIRST hardware implementation of 4K HEVC/ H.265 video encode. HEVC designed to deliver up to 50% better video compression

Qualcomm Technologies' FIRST 14-bit Dual ISP for highest quality, depth enabled photography. Up to 21MP for main camera with depth assist, phase detect, for sharper dual camera user experiences

Qualcomm Adreno and Qualcomm Hexagon are prc

8

Memory ACCESs is >500× Arithmetic Energy

Operation	16 bit (integer)		64 bit (DP-FP)	
	E/op PJ	vs. Add	E/op PJ	vs. Add
ADD	0.18	1.0 ×	5	1.0 ×
Multiply	0.62	3.4 ×	20	4.0 ×
16-Word Register File	0.12	0.7 ×	0.34	0.07 ×
64-Word Register File	0.23	1.3 ×	0.42	0.08 ×
4 K-word SRAM	8	44 ×	26	5.2 ×
32 K-word SRAM	11	61 ×	47	9.4 ×
DRAM	**640**	**3556×**	**2560**	**512 ×**

The Right Tool

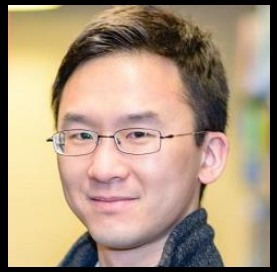
Vision: Machine Learning

- Start with Building Blocks of Machine Learning and Deep Learning
- Guest Lecture: Kian Katanforoosh (Stanford)
  - Machine Learnig to Deep Learning Journey

- Guest Lecture: Hadi Esmaeilzadeh (UCSD)
  - Machine Learning Accelerators on FPGAs

Vision : Linear Algebra, Inference, Numerics

- Linear Algebra Basics
- Matrix Computations and their accelerators
- Neural Network Inference
- Guest Lecture: Yu-Hsin Chen (MIT & Nvidia)
  - Deep Learning Inference



- Guest Lecture: Robert Schreiber (Cerebras)
  - Understanding Numerical Errors

Vision: Training, Scaling Inference

- Training Basics
- Training Tradeoffs
- Precision and Sparsity
- Guest Lecture: Boris Ginsburg (Nvidia)
  - Training of DNNs

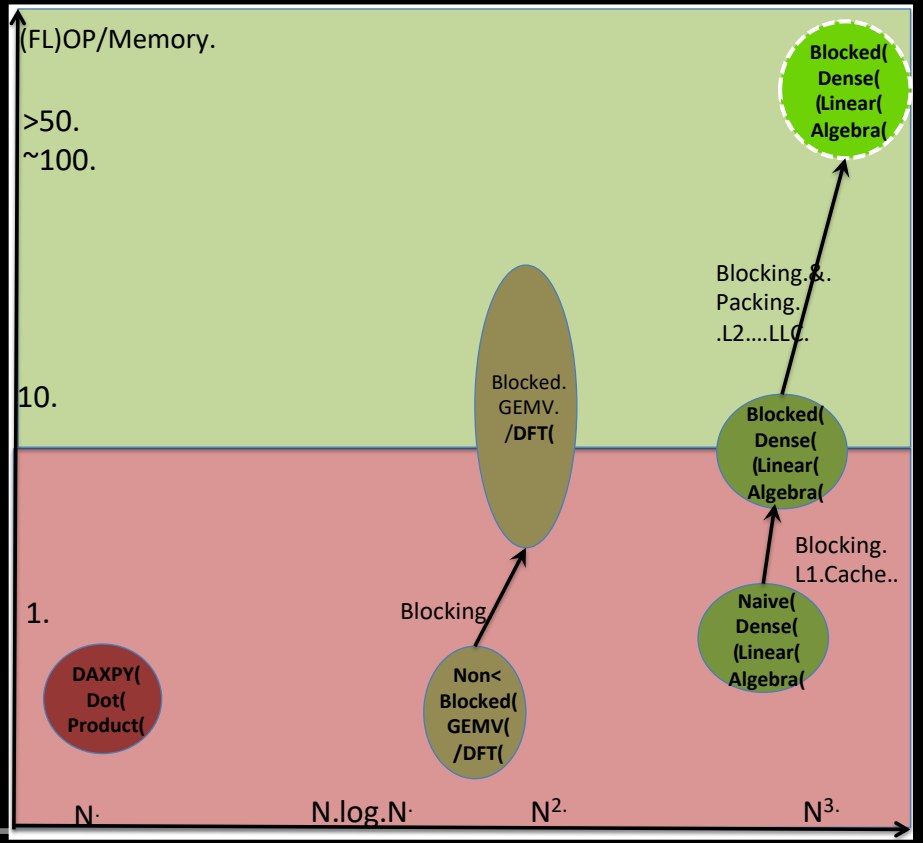
- Guest Lecture: Eric Chung (Microsoft)
  - Real-Time AI at Cloud Scale with Project Brainwave

Vision: Scaling, ML Benchmarks Performance

- Scaling Training
- Distributed Training
- CGRAs and FPGAs
- Guest Lecture: Cliff Young (Google)
  - MLPerf Benchmark
- Guest Lecture: Mikhail Smelyanskiy (Facebook)
  - AI at Facebook Datacenter Scale

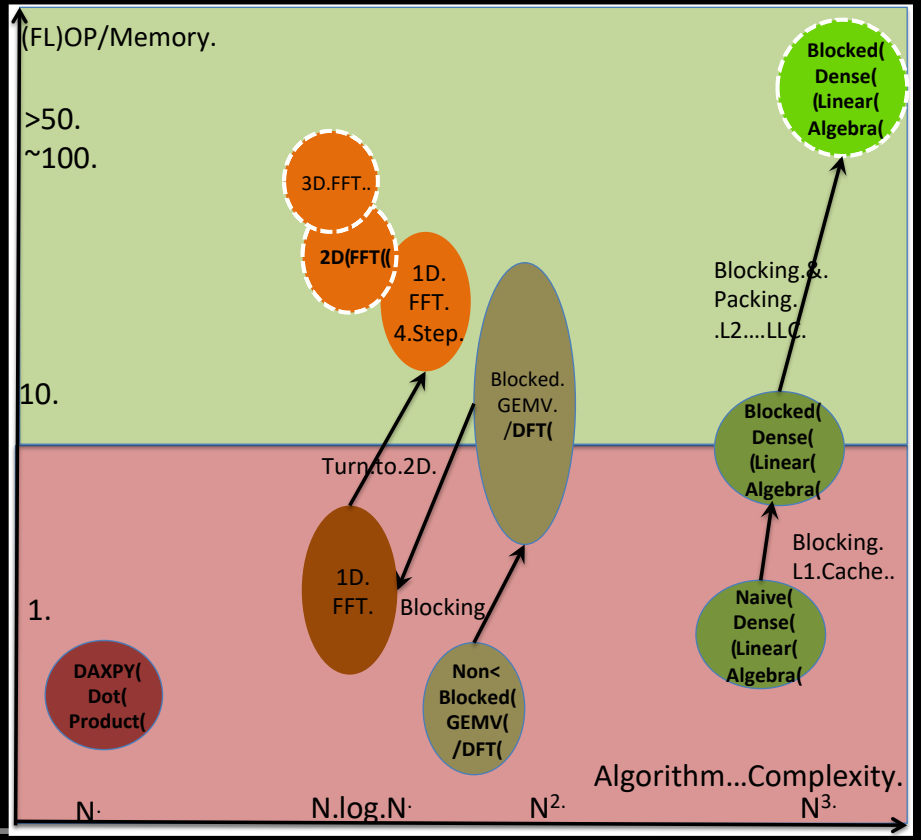
Scientific Computing Applications

- Locality
  - Blocking



Scientific Computing Applications

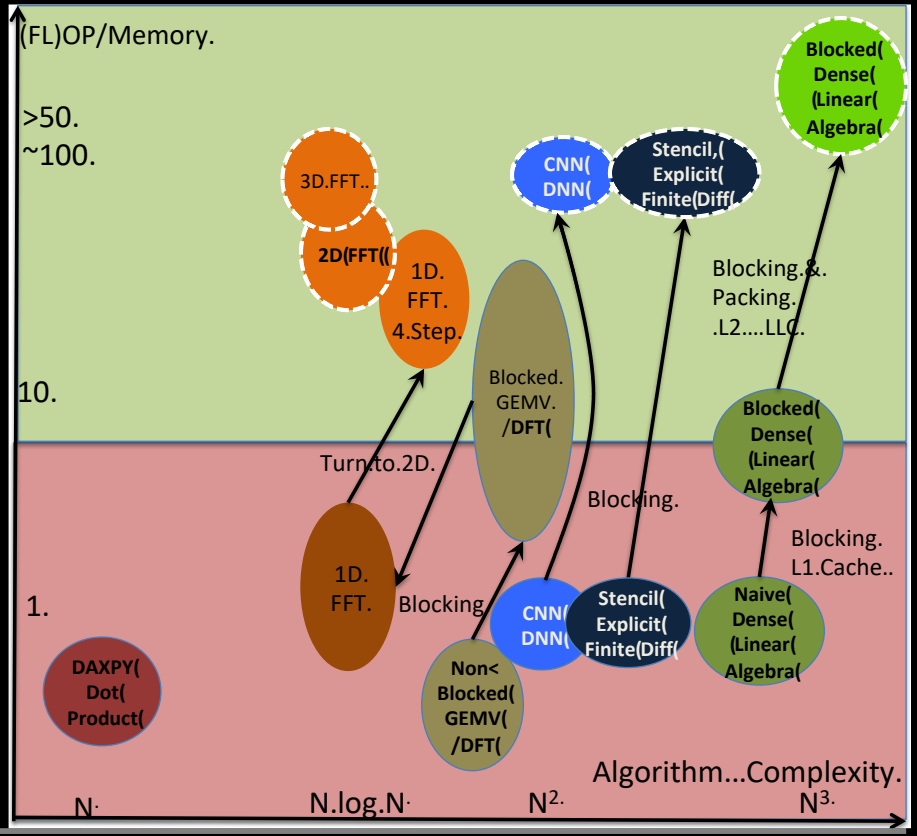
- Locality
  - Blocking
- Change the Nature



Nominal Complexity or the big O

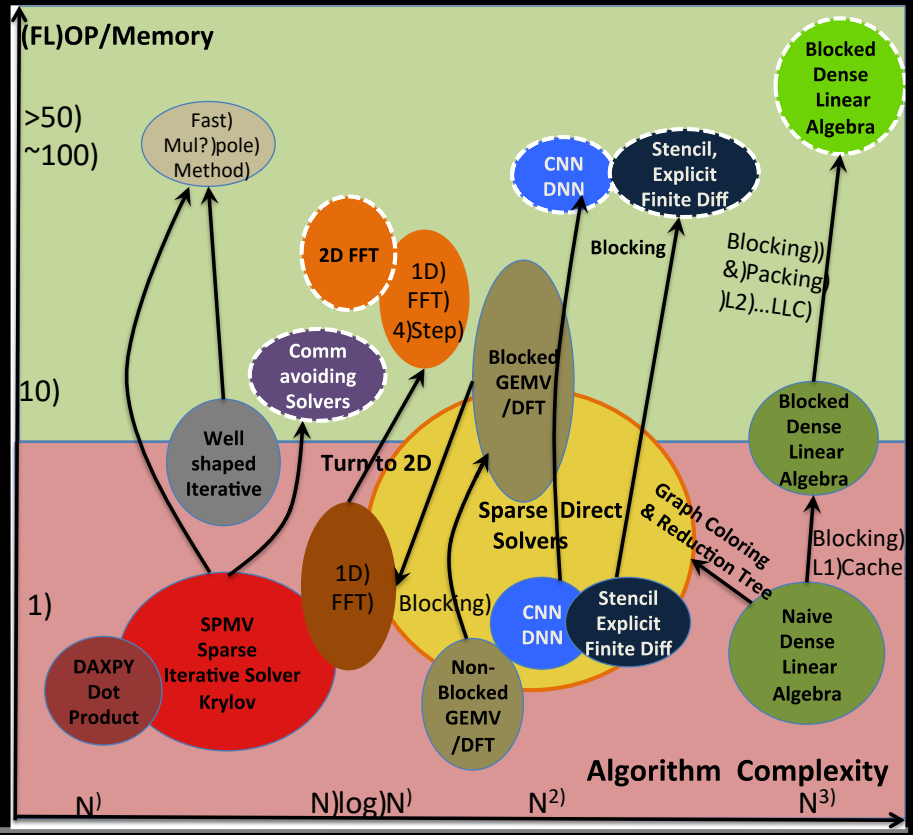
Scientific Computing Applications

- Locality
  - Blocking
- Change the Nature
- Parallelism
  - Communication
  - Memory Partitioning



Scientific Computing Applications

- Parallelism & locality
- Algorithm complexity vs. Memory behavior
- FLOP/Memory
- Algorithm changes memory behavior



Algorithm/Architecture Codesign